Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application

Anna Katerina Jordanous

Submitted for the degree of Doctor of Philosophy

Department of Informatics, University of Sussex

December, 2012

Declaration

I hereby declare that this thesis has not been submitted, either in the same or different form, to this or any other another University for the award of a degree.

Signature:

Acknowledgements

My two supervisors, Nick Collins and Chris Thornton, who helped me in different ways during their supervision: thanks to Nick for his continual support, versatility and intelligent input over the years (and a specific thanks for help towards devising the 'SPECS' acronym), and thanks to Chris for his assistance and advice (all was noted, even if advice was not always followed);

My two examiners, Alison Pease and Steve Torrance, for their thoughtful and considered input;

The Department of Informatics at the University of Sussex for partially funding this doctoral study and the Sir Richard Stapley Educational Trust, AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour), Creativity & Cognition/NSF (National Science Foundation) and CIM (Conferences on Interdisciplinary Musicology) for further financial support and opportunities;

My officemates over the years: Jens Holger-Streck, Thor Magnusson, Raphael Commins, Gareth White, Layda Gongora, Chad McKinney and especially Chris Kiefer (thanks for everything, Chris);

Alan Smaill and Bill Keller for their academic advice and collaborative contributions;

Phil Husbands, Geraldine Fitzpatrick, Meurig Beynon, Mike Joy and Jane Sinclair for their help with academic experience, feedback and funding;

Trisha Agarwal, Linda Robson and Rose Kelly for their belief and encouragement to start all this;

Till's bookshop in Edinburgh where I found Maggie's book, triggering my interest in creativity;

Sandra Deshors for being an inspirational example of determination, drive and resourcefulness;

Jamie Matthews for intelligent conversations, lunches, coffees and `J4mie Computing Supplies`;

Friends of the 'pink door', for contributing to such a supportive community during our time in Brighton, especially Seda Ilter, Francesca Conti, Daniela DeBono, Jane Stanford and Katha Koppert;

Mark Hedges, Tobias Blanke, Charlotte Tupman and Gareth Knight at the Centre for e-Research and the Department of Digital Humanities, King's College London, for advice and support this past year;

All who helped me in various ways, including (in addition to those listed above) Suzie Wilkins, Leon Baruah, David Mountford, Chris Starkey, Sam Lambourne, Patrick Billingham and Joe Dorrell;

The 'teatimers' in Informatics for the weekly sanity checks;

Andy Roberts' LaTeX tutorials at `http://www.andy-roberts.net/writing/latex` - a regular reference;

Almost last but definitely not least, my participants, who gave their time and effort to contribute to my research - thanks to all, especially the Case Study judges and Al Biles, Bob Keller and George E. Lewis

Mum & Paul, Sam.

For Trisha (1979-2011)

Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application

Anna Katerina Jordanous

Summary

This thesis proposes *SPECS*: a *Standardised Procedure for Evaluating Creative Systems*.

No methodology has been accepted as standard for evaluating the creativity of a system in the field of computational creativity and the multi-faceted and subjective nature of creativity generates substantial definitional issues. Evaluative practice has developed a general lack of rigour and systematicity, hindering research progress.

SPECS is a standardised and systematic methodology for evaluating computational creativity. It is flexible enough to be applied to a variety of different types of creative system and adaptable to specific demands in different types of creativity. In the three-stage process of evaluation, researchers are required to be specific about what creativity entails in the domain they work in and what standards they test a system's creativity by. To assist researchers, definitional issues are investigated and a set of components representing aspects of creativity is presented, which was empirically derived using computational linguistics analysis. These components are recommended for use within SPECS, being offered as a general definition of creativity that can be customised to account for any specific priorities for creativity in a given domain.

SPECS is applied in a case study for detailed comparisons of the creativity of three musical improvisation systems, identifying which systems are more creative than others and why. In a second case study, SPECS is used to capture initial impressions on the creativity of systems presented at a 2011 computational creativity research event. Five systems performing different creative tasks are compared and contrasted.

These case studies exemplify the valuable information that can be obtained on a system's strengths and weaknesses. SPECS gives researchers vital feedback for improving their systems' creativity, informing further progress in computational creativity research.

Submitted for the degree of Doctor of Philosophy

Department of Informatics, University of Sussex

December, 2012

Contents

Chapter 1

Introduction

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012).

Figure 1.1: *Wordle* word cloud of this Chapter's content.

N.B. Each Chapter is preceded by a word cloud summarising the words in that Chapter, produced using online software at `http://www.wordle.net`. The word clouds illustrate a key point underlying the work in Chapter 4: the meaning we attribute to something is reflected in the words we commonly use to describe it. The word clouds reflect what can be obtained through empirical inspection of language usage, which Chapter 4 will explore in greater detail.

Overview

This initial Chapter introduces the focus of this thesis work in Section 1.1, presenting the research question that is being addressed: *How should we evaluate the creativity of a computational creativity system?* Motivations for this work are outlined in Section 1.2. The work in this thesis centres around evaluation of computational creativity. Evaluation is presented as a key part of research (Section 1.3. Section 1.5 introduces a central assumption made during this work, that computational creativity is, in principle, an achievable possibility. Section 1.5 presents computational creativity as a research field and comments on its development and growth, previewing Section 1.5.2 which looks at how evaluation is viewed within computational creativity research. Section 1.5.2 focuses on the perception of evaluation as one of the 'Grand Challenges' of computational creativity.

Section 1.6 outlines what is hoped to be achieved by the work in this thesis. As Section 1.7 points out, although the immediate intended target audience of the thesis is computational creativity researchers who wish to evaluate the creativity of their systems, some of the contributions of this thesis can be useful to a wider academic research audience. To help readers navigate through the thesis to the parts most useful for them, a readers' guide to the thesis is offered in Section 1.7.

1.1 The focus of this thesis: Evaluating computational creativity

This thesis addresses the research question: *How should we evaluate the creativity of a computational creativity system?* Computational creativity is a youthful research discipline (Colton, 2008a; Cardoso, Veale, & Wiggins, 2009) and practical suggestions for creativity evaluation methodologies have only been presented within the last ten years. The aim of this thesis is to provide researchers with a practical, standardisable evaluation approach, to help them assess the creativity of their systems.

1.2 Motivations for this work

> 'A very important issue is the assessment of the creativity of programs.' (Colton, 2008a, p. 6)

> 'unless the motivations and aims of the research are stated and appropriate methodologies and assessment procedures adopted, it is hard for other researchers to appreciate the practical or theoretical significance of the work. This, in turn, hinders ... the comparison of different theories and practical applications ... [and] has encouraged the stagnation of the fields of research involved.' (Pearce, Meredith, & Wiggins, 2002, p. 119)

Pearce et al. (2002) highlighted a 'methodological malaise' faced by those working with computational music composition systems, following similar previous criticisms of AI research in general Bundy (1990). A lack of methodological standards for system development and evaluation, Pearce et al. argued, was leading to the 'stagnation' of this research area (Pearce et al., 2002, p. 119). Computational creativity research is in danger of succumbing to this same malaise.

Surveying the literature on computational creativity systems, one quickly finds evidence that systematic evaluation of creativity has been neglected.[1] Evaluation of computational creativity systems is often ad-hoc, not reported in a way that can be criticised or replicated, or omitted entirely. Researchers have not adopted one method as standard and creativity evaluation is often not conducted in an academically rigorous way. Sometimes it is entirely absent.

Currently many implementors of creative systems follow a creative-practitioner-type approach, producing a system then presenting it to others, whose critical reaction determines its worth as a creative entity. A creative practitioner's primary aim, however, is to produce creative work to be judged on its own merits and quality, rather than to critically investigate creativity; in general this latter investigative aim is important in computational creativity research. Creativity entails more than just the quality of the output, for example, what about novelty, or variety? Yet computational creativity systems are often described as creative without justification for this claim.

This cannot be the case for academic evaluation; the criteria by which you evaluate should be clearly stated, for rigour of approach and to enable comparison and criticism of evaluation results. Systematic evaluation is important for computational creativity research, allowing us to compare and contrast progress. Ignoring this evaluation stage deprives us of valuable analytical information about what our creative systems achieve, especially in comparison to other systems.

1.3 The role of evaluation and why it is needed

Evaluation helps identify where progress is being made and how the evaluated item can be improved upon. Without evaluation, how can research progress be demonstrated and tracked? And how can we understand and learn from our research without considering what has (and has not) been achieved?

Two different types of evaluation are summative evaluation and formative evaluation. In the context of this thesis, summative evaluation aims to provide a summary of a system's creativity, perhaps in quantitative form, for judgement of the amount of creativity demonstrated by that system. Formative evaluation, on the other hand, provides feedback on the system's strengths and weaknesses, to assist improvements and developments during ongoing development or in reflections on how the system worked well and how it could be improved. The distinction between summative and formative evaluation is analogous to two types of educational feedback: giving a mark representing the work's quality, or giving constructive feedback on how to improve the work, respectively.

Both types of evaluation are examined in previous computational creativity evaluation literature. Colton, Pease, and Ritchie (2001) advocate an entirely formative approach aimed at using evaluation feedback 'in the design of creative programs rather than in the assessment of established programs.' (Colton et al., 2001, p. 1). Ritchie (2007) focuses instead on summative evaluation to measure the

[1] Details of this survey shall be reported in Chapter 2 Section 2.3.

creativity of an existing program, proposing criteria 'which give some indicators of the extent to which that program has been creative on that occasion.' (Ritchie, 2007, p. 74).

This work aims more towards the generation of formative feedback to help improve existing systems and design new systems. It does however recognise the value in some summative feedback, to reward particularly creative achievements with positive evaluations, identifying what progress systems are making and what contributions are being made to knowledge. Evaluation in this work is treated as reviewing systems, comparing them to others to an appropriate degree to see which are more creative than others and in what ways, and learning from this evaluation and comparison.

> 'It is important to be explicit ... about the criteria that are being applied in making judgements of creativity.' (Ritchie, 2001, p. 3)

Ritchie's above point stresses that it is not sufficient just to say that a system has been evaluated; how the evaluation was performed should be easily transparent, so the evaluation process can be repeated on other systems for consistency and for comparison of evaluation results. A transparent evaluation process also becomes available to others for critique and perhaps to learn from, helping us sharing knowledge and learning from the work that we are doing. Sloman (1978) stresses that 'we can reasonably, though only tentatively, infer that something is possible if we have an explanation of its possibility' (Sloman, 1978, p. 44). Such an explanation should not be contradicted by other existing, established theories or laws.[2] Lack of these explanations would, Sloman argues, affect the development of shared knowledge and understanding:

> 'Unfortunately, the role of such explanations in our thought is obscured by the fact that not everyone who requires, seeks or finds such an explanation, or who learns one from other people, asks this sort of question explicitly, or fully articulates the explanation when he has understood.' (Sloman, 1978, p. 45)

1.4 A central assumption of this thesis: Computers *can* be creative

A central assumption of this thesis should be stated at the outset. Whilst it is acknowledged that there is some resistance to the idea of computers being creative, this thesis makes a working assumption that computational creativity is possible, taking a pragmatic outlook on computational creativity: *Given the assumption that computational systems can be creative, how does the computational creativity researcher measure how creative a computational system is?*

Whether computational creativity research makes progress towards simulating human creativity, or whether computers would actually be able to express true creativity (whatever 'true creativity' might be), is a debate which for practical reasons is best left for another time and place.[3] Although

[2] Excepting where existing theories are invalid and are disproved by the new theory; see Chapter 2 Section 2.2.

[3] Boden (2004) and Colton (2008b) address this debate to some extent and it is also approached from a more general standpoint by other authors (e.g. Dreyfus, 1979; Searle, 1980; Minsky, 1982)).

this thesis makes the working assumption that computational creativity is possible, it is necessary to acknowledge and consider this resistance to the idea and consider potential reasons for it.

> 'Robots are being created that can think, act, and relate to humans. Are we ready?' (Carroll, 2011, p. 67)

Some resistance to the possibility of a computerised system being creative does exist, whether the computer is in robot form or in a more traditional computer setup. Brian Reffin-Smith's 43 Dodgy Statements on Computer Art (Reffin-Smith, 2010) eloquently summarise the conflict of views on computational creativity, through reflections on both positive and negative aspects of computer art:

> '19. Using a computer merely to access the web is like using a Bugatti Veyron to deliver the papers. ...
>
> 37. "*i*", the imaginary square root of minus 1, is to the real numbers as the computer is - or should be - to art. ...
>
> 41. Of course computers and other devices will never fully understand flowing, allusive conversation. But they won't care.' (Reffin-Smith, 2010)

Negative preconceptions about computational creativity primarily exist because creativity is seen as something which is essentially human, a notable quality which should not be reducible to the form of a computer program. Computers are perceived only to be able to blindly follow instructions given to them by their programmers, with no intention or motivation guiding these actions (Searle, 1980; Dreyfus, 1979); how can the ability to be creative be expressed in such instructions? Identifying such instructions can itself be seen as a creative act, but the creativity is on the part of the programmer, not the computer receiving the instructions (Ritchie, 2001; Jennings, 2010a).[4]

A long-propagated view is that there are certain qualities, such as being creative, which a computer simply cannot be capable of (Nake, 2009; Dreyfus, 1979; Weizenbaum, 1976). As an example of this view, Colton (2008b) quotes from the June 1934 edition of *Meccano* magazine:

> ' "Truly creative thinking of course will always remain beyond the power of any machine" ' (Meccano magazine, June 1934, quoted in Colton, 2008b)

Resistance to computational creativity is often accompanied by the conception that computers cannot be creative beyond what the programmer provides the system (as discussed by, for example, Ritchie, 2001; Jennings, 2010a). There is a reluctance to accept that human skills and abilities that people prize could be replicated computationally. Additionally, the common preconception of engineering techniques as non-creative automata reinforces the belief that computers cannot do the kind of tasks where algorithms or instructions have not been clearly specified beforehand. This is perceived to require human level intelligence and demonstrates our superiority over computers.

> 'Creative artificial intelligence must always fight the impression that it is simply a fancy tool for expressing the programmer's creativity.' (Jennings, 2010a, p. 500)

[4] This was often found in the feedback from surveys carried out in Chapter 8, as reported in Section 8.1.1 of that Chapter.

This negativity influences people's judgements of computational creativity when they hear that a computer was responsible for producing an artefact (Ritchie, 2001; Colton, 2008a, 2008b).[5] Such negativity has propagated through to have some legal significance, as discussed in Chapter 3 Section 3.5. It has also had an adverse effect on computational creativity research progress (Colton, 2008a).

Notwithstanding the ease with which we sometimes assign computers human-like characteristics (Reeves & Nass, 2002; McCarthy, 1979), difficulties with identifying with computational creativity may contribute to negative impressions of computational creativity. Hennessey and Amabile refer to creativity as a 'distinctively human capacity' [p. 570], 'one of the key factors that drive civilization forward.' [p. 570]. As Boden (2004) describes:

> 'A work of art is an expression of human experience and/or a communication between human beings, so machines simply don't count.' (Boden, 2004, p. 7)

In conversations with musicians,[6] a common reaction emerges, that computer improvisers may progress to the level where they could replace humans, making the human improviser unnecessary. Colton (2008a) considers and refutes this objection:

> 'Why on earth would Lucien Freud stop painting just because a computer can paint as well? Moreover, people will always appreciate the blood, sweat, and tears expended by creative people in producing their works. (Colton, 2008a, p. 7)

In a recent interview (Blitstein, 2010), David Cope mentions an oft-encountered objection to his work with generative music creativity systems (Cope, 2005): 'aren't you taking away the last little thing we have left that we can call unique to human beings - creativity?' Blitstein (2010) reports Cope's controversial response to this worry about computational creativity somehow devaluing human creativity. Cope calls into question how creative humans can actually claim to be. As computers work without biases and prejudices, unlike humans, Cope argues that:

> 'humans are more robotic than machines. "The question," Cope says, "isn't whether computers have a soul, but whether humans have a soul." ' (Blitstein, 2010)

Computational creativity research showcases several ways of producing unpredictable behaviour that goes beyond explicit program instructions, such as connectionist computation, evolutionary computing, emergence or the use of randomness (examples can be found in Chapter 3 Section 3.4.1). Negative attitudes on computational creativity have however encouraged an anti-mechanism definition of creativity; creativity becomes what computers cannot do or that which we do not understand (Dreyfus, 1979; Weizenbaum, 1976). This suggests a shifting definition of creativity, for those resistant to the idea of computers being creative, essentially centred around progress in computational creativity. For computational creativity researchers, the most productive strategy may be to ignore such definitions but recognise the effect computational creativity is having on the perception on creativity for some.

[5] Again much feedback generated from the Chapter 8 surveys echoed this view.

[6] And particularly in the debrief for the survey on musical improvisation carried out for Chapter 6.

Boden poses what she terms as four Lovelace questions (Boden, 2004, pp. 16-17), inspired by Ada Lovelace's observations about the contemporary computer of her time (Babbage's Analytical Engine) only being capable of doing what humans program it to do (Lovelace, 1953):

1. Can computational models help us to understand creativity?
2. Can a computer *appear to* be creative? [this is followed up in (Colton, 2008b)]
3. Can a computer appear to recognise creativity?
4. Can a computer *really* be creative?

Boden (2004) argues that the first three questions can be answered positively. She de-emphasises the fourth question, deliberately not addressing it until the final chapter of (Boden, 2004), because she sees question 4 is philosophical, unlike the scientific questions 1-3, so requires a different treatment. Instead Boden concentrates her focus on the scientific questions 1-3.

It has been shown that 'people do have a tendency to discount and even dislike computer creativity' (Moffat & Kelly, 2006, p. 6). In a study with musicians and non-musicians, Moffat and Kelly found that on a subconscious level, musicians were often negatively influenced against a piece of music composed by a computer system, using subconscious aesthetic tendencies to like or dislike the music.[7] This effect was found even though the musicians were also found to be poor at using their expert knowledge to consciously detect that a piece of music is computer-composed.

As seen above, negative preconceptions against computational creativity are non-trivial. For the purposes of the current work, though, assuming that computers can potentially be creative allows us to make methodological and practical progress in evaluating computational creativity without being too restricted by these objections.

1.5 Computational creativity as a research field

As Chapter 3 Section 3.4.1 will report, computational creativity is generally seen as the demonstration by computers of activity, behaviour or output production that would be perceived as creative if demonstrated by a human. This definition has been developed in discussions over the past decade (Schmid, 1996; Wiggins, 2006a; Cardoso & Wiggins, 2007; Colton, 2008b; Cardoso et al., 2009).[8] [9]

Computational creativity research follows both theoretical and practical directions[10] and crosses several disciplinary boundaries across the arts, sciences and engineering. This has led to the emer-

[7] Somewhat surprisingly, this finding was not replicated to a significant degree in the non-musicians' behaviour.

[8] This definition is also advocated at http://www.computationalcreativity.net, a site hosting relevant information for computational creativity research, maintained by the steering committee for the computational creativity research community.

[9] According to reports by Simon Colton and Geraint Wiggins at the most recent conference on computational creativity (ICCC'12), the latest revision of this definition is likely to remove the specific reference to human creativity, instead referring to creativity in general; however the basic structuring of the definition remains the same, i.e. the demonstration of something that can be described as 'creative'. Chapter 3 will investigate in more detail what this actually entails.

[10] The work reported in this thesis uses creativity theory to provide practical support to that subsection of computational creativity research where creative systems are designed and implemented to model creativity or demonstrate creativeness.

gence of a community with varying and occasionally disparate aims and motivations, ranging from artistic goals to scientific exploration of creativity or the pursuit of software and/or hardware engineering achievements. Research within the field is influenced by artificial intelligence, computer science, psychology and specific creative domains that have received attention from computational creativity researchers to date, such as art, music, reasoning and narrative/story telling (Colton, 2008b; Widmer, Flossmann, & Grachten, 2009; León & Gervás, 2010; Pérez y Pérez, 1999, provide examples).

1.5.1 Development of the Computational Creativity research community

In examining current evaluative practice in the research field of Computational Creativity, it is important to understand how the field has developed. This places the practical discussions in context and highlights researchers' key aims and objectives within the field.

Though the field does not yet have a dedicated journal for research publication, the growth and active development of computational creativity is demonstrated by a healthy and sustained recent increase in workshop and conference activity, as well as a number of journal special issues on computational creativity research (featuring selected papers from prior research events).[11] Computational creativity research events have been taking place regularly since 1999, developing from satellite workshops at artificial intelligence conferences[12] leading to autonomous workshops (2007-2008)[13] and then to an annual conference series, the International Conference for Computational Creativity, or ICCC (2010-present), taking place in Portugal (2010), Mexico (2011) and Ireland (2012). This development has been accompanied by a substantial and increasing growth in the number of papers presented to such research events and in program committee sizes.[14] With the recent inclusion of computational creativity as a fundable research area for the European Union's Framework Programme 7,[15] a significant European source of funding, it can be said that computational creativity research is 'coming of age' (Colton, de Mataras, & Stock, 2009).

The question of how best to evaluate the creativity of computational systems has been described as a 'big question' in this research area, one which we may not even be able to start tackling properly in the context of current research.[16] This view has not been echoed in the calls for papers for research

[11] Knowledge-Based Systems 2006: 19(7), New Generation Computing 2006: 24(3), AI Magazine 2009: 30(3), Minds and Machines 2010: 20(4).

[12] Computational Creativity workshops have been held in conjunction with several AI conferences (AISB'99, AISB'00, AISB'01, ECAI'02, AISB'02, IJCAI'03, AISB'03, IJCAI'05, ECAI'06) case-based reasoning conferences (ICCBR'01, ECCBR'04) and linguistics conferences (LREC'04, NAACL'09).

[13] Autonomous workshops grew out of the International Joint Workshop on Computational Creativity series (2004-2008), which started through the coming together of communities from AI and from Cognitive Science, to hold joint research events on computational creativity. Separate symposiums have also been held, in Stanford, California (twice).

[14] Pre-2004 workshops typically contained 10-15 papers, with program committees of between 5-15 people. This has now grown to averages of 33 accepted papers and 42 program committee members over the 2010-2012 conferences.

[15] Objective ICT-2013.8.1, 'Technologies and scientific foundations in the field of creativity', in the draft for the EU FP7 programme for 2013. See http://computationalcreativity.net/index.php?news&nid=9 (last accessed November 2012).

[16] This is discussed in more detail in Section 1.5.2.

events. Creativity evaluation metrics and strategies have frequently appeared on the list of topics of interest for workshops and symposiums in the form of phrases such as 'Evaluation of Creativity' (workshops on creative systems, 2002-04), 'the assessment of creativity in AI programs' (AISB workshop, 2003), 'how we assess creativity in computers' (IJWCC 2007), 'Metrics, frameworks and formalizations for the evaluation of novelty and originality'[17] (computational creativity workshop, 2005) and the rephrasing 'Metrics, frameworks and formalizations for the evaluation of creativity in computational systems' (computational creativity workshops, 2006 and 2008). This last wording has appeared in the call for papers for all ICCC conferences to date (2010-2012).[18]

As will be discussed later in this thesis, in Chapter 3, in the past there has been only limited evidence of computational creativity researchers evaluating their systems' creativity and demonstrating to what extent their computational systems can actually be considered to be 'creative systems'. This may be partly due to the conflicting messages described above about whether creativity evaluation is a plausible thing to attempt. Another reason may be that reviewers' judgements and decisions may be influenced by personal criteria that may differ from reviewer to reviewer (Pérez y Pérez, 2012, personal communications), hence the emphasis on a need for evaluation may also differ across reviewers. The variety of disciplinary backgrounds that reviewers may come to computational creativity from (as mentioned above in this Section) makes Pérez y Pérez's point particularly pertinent. Additionally, this culture may have developed as a side-effect of the inclusive efforts to build up a community of computational creativity researchers; decisions on whether a paper should be accepted to a conference could be based around whether the paper would trigger interesting debate, rather than how academically rigorous its presentation was (Pease, 2012, personal communications). This was enhanced by the practice, since 2001, of accepting position/short papers (reports of work in progress or comments on research directions) alongside technical/long papers (detailed technical reports of creative systems or foundational theory). Whilst more thorough academic reporting is required for technical papers, a requirement which has been particularly imposed in the last few years (Pease, 2012, personal communications), position papers often report work in progress rather than completed work, so have less requirements imposed for academic rigour in reporting. Position papers encourage the reporting of current work and new methods even if the work is not yet fully completed. Unfortunately, as those proceedings often do not clearly distinguish technical papers from position papers,[19]

[17] The combination of novelty and originality is often used as a reductionist definition of creativity (Pease, Winterstein, & Colton, 2001; Peinado & Gervas, 2006; Pereira & Cardoso, 2006; Ritchie, 2007; Alvarado Lopez & Pérez y Pérez, 2008; Brown, 2009a; Chordia & Rae, 2010). Definitional issues shall be returned to later in this thesis, in Chapter 3.

[18] For ICCC'11, this phrasing appeared with a qualifier: 'quasi-formal approaches that, for example, argue for recognition without definition or that define the absence of creativity may have interesting implications for computational creativity'. This was probably in response to just such a evaluation framework offered by Colton (Colton, 2008b), which quickly became adopted more often than more formally stated predecessors such as (Ritchie, 2007; Pease et al., 2001), as shall be shown later in this thesis, in the survey presented in Chapter 3.

[19] See for example the proceedings for ICCC'11, ICCC'10 or IJWCC'07 (ICCC'11, 2011; ICCC'10, 2010; IJWCC'07, 2007) where a position paper is distinguishable from a full technical paper only by its number of pages.

this differentiation in quality can easily be missed, making a lack of evaluative (and other academic) rigour seemingly more acceptable in this community.

The issue of good practice in creativity evaluation has been highlighted by some researchers for many years now, for example Ritchie argued in 2001 that '[i]t is important to be explicit ... about the criteria that are being applied in making judgements of creativity.' (Ritchie, 2001, p. 3). As the field moves from its formative years of community development, into a position where it attracts enough research to support an annual international conference audience and journal special issues, evaluation has become more important for full research reports (Pease, 2012, personal communications). In conference, lack of evaluation in a paper has become a valid reason for rejecting a long paper or changing its status to that of a position paper (representing that work on the system is not yet complete). This is illustrated by the following (anonymised) reviewers' comments:

> 'This is the fourth paper I am reviewing for ICCC 2011 but the first one that takes evaluation seriously. In fact, two of the three papers I have reviewed so far don't even mention the issue of evaluation, and the third mentions it only in passing, as something someone might do one day. ... if this trend continues, then we as a community will make it harder to make progress.'
>
> 'This is a very strong paper. I just can't bring myself to give a "strong accept" to work which has such a dismissive attitude to evaluation.'

What is needed at this stage of the field's development is an established and standardised approach to evaluation for tracking and evaluating progress in computational creativity. As shall be seen and explored throughout this thesis, though,[20] the issue of how best to evaluate computational creativity systems has generated many more questions than solutions.

1.5.2 'Big questions' and 'Grand Challenges' in computational creativity

In colloquial discussions at the 2011 International Conference in Computational Creativity (ICCC'11), the question of how to evaluate computational creativity was referred to as one of the 'big questions' of this research area. Although some authors have proposed evaluation methodologies for creativity,[21] to some at ICCC'11, it seemed pointless to tackle such questions while they have not yet been dealt with sufficiently in human creativity research, despite decades more investigation.[22]

In aiming to evaluate a system's creativity, a number of complications exist. Human preconceptions about computational creativity, the lack of a creativity baseline to evaluate creative systems against and the general difficulty in defining creativity add to methodological issues such as who should evaluate the system, whether evaluation should be quantitative or qualitative, and how (or if) the creativity of two systems can be compared. In a recent research seminar presentation (Wiggins, 2008), Wiggins stated that he believes it is currently too awkward to assess creativity and therefore his evaluative choice was instead to assess the quality of his systems.

[20] Particularly Chapters 2 and 3.

[21] Existing evaluation methodologies for computational creativity are examined in detail in Chapter 2.

[22] Chapter 3 will look at research on how human creativity could be evaluated.

Cardoso et al. (2009) refer to evaluation in computational creativity as one of the *Grand Challenges* facing computational creativity researchers. This particular Grand Challenge, argue Cardoso et al., 'probably needs to be deferred until we are substantially more capable in general automated reasoning and knowledge representation' (Cardoso et al., 2009, p. 19).

It is important not to ignore a key issue in research merely because it is tricky to tackle. Instead we should acknowledge the issue and surrounding debate and look for a 'working' answer if necessary, simplifying assumptions to make the issue more tractable. This working answer may not be *the* answer and in fact *the* answer may not even exist or may change over time.[23] Taking a practical and proactive approach does however reduce perceived barriers to research, making 'hard' issues such as evaluation of creativity more manageable and less stifling.

This thesis argues that performing evaluation, based on a working understanding of creativity if necessary, provides more informative, more useful and more deeply grounded contributions than to perform no evaluation at all.

> 'if we are to make progress, the first thing we must do is face up to the things that make the problem so difficult. Then we can move forward toward a theory, without blinkers and with a good idea of the task at hand.' (Chalmers, 1996, p. xii, on the study of consciousness theory)

1.6 Research aims of this thesis

The research aims of this doctoral work are to make the following contributions:

- A practically applicable, standardised, flexible methodology for evaluating how creative computational systems are and generating constructive feedback.
- A clearer understanding of creativity research that crosses interdisciplinary divides and collects together findings from different academic perspectives.
- A working definition of creativity to use for evaluation.
- Evaluative feedback for a number of systems in a variety of domains, as a result of practically applying the methodological evaluation tool in case studies.
- Criteria for meta-evaluation of creativity evaluation methodologies and the application of these criteria to critically compare key methodologies.

1.7 Readers' guide

The apostrophe placing in the title of this Section is significant. This thesis is potentially useful for people from various academic backgrounds who are interested in creativity, computational or human:

- The main target readership of this thesis are computational creativity researchers who want to know how best to evaluate the creativity of their computational creativity systems.

[23] Changing perceptions of creativity over time will be considered more in this thesis, particularly in Chapters 3 and 10.

  - Some will mainly want to learn about the evaluation methodology and how to use it.
  - Other readers will want a more informed view of how the methodology was derived to deal with issues in creativity and evaluation.
  - Some readers may want to evaluate and compare creativity evaluation methodologies.
- Other readers may be working in areas similar to the systems evaluated in the case studies (or may have a system evaluated in one of these case studies) and will therefore be looking for information on how to make these systems more creative.
- In particular, Case Study 1 looks in detail at creativity in musical improvisation, reporting much information on how this type of creativity is manifested and evaluated. This is useful both for researchers modelling this type of creativity computationally and those interested in musical improvisation creativity from a music research perspective.[24]
- This thesis brings together research from several different disciplines, from computational creativity to areas such as psychology, education and law. Content especially in the earlier half of this thesis should be useful to those seeking a clearer understanding of creativity unhindered by academic 'blinkers' and discipline-specific research priorities.
- For people researching human creativity, it is possible that the methodology suggested in this thesis could also be applied to evaluate the creativity of people as well as computers.[25] This application falls outside the scope of this thesis work but would be very interesting to see.

With these different types of readership in mind, below is a summary of each Chapter's content:

Ch.1 Introduction to the thesis, its direction and research question, giving foundational content about computational creativity and evaluation.

Ch.2 Critical review of previous work in computational creativity evaluation, scientific method and a survey of current evaluative practice in this research.

Ch.3 In-depth investigation into how creativity has been defined in terms of computational and human creativity, for a more informed and comprehensive understanding.

Ch.4 Empirical identification of 14 key components of creativity through natural language processing tools and statistical analysis methods. The identified components collectively build up a multi-perspective definition of creativity, incorporating common themes across creativity research.

Ch.5 Presentation of the SPECS methodology: the *Standardised Procedure for Evaluating Creative Systems*, derived from the investigations into current practice in computational creativity evaluation, existing methodologies and the definitional findings.

Ch.6 Case Study 1: applying the SPECS methodology for a detailed and informed evaluation of the creativity of three musical improvisation systems.

Ch.7 Case Study 2: using the SPECS methodology to simulate how we initially judge creativity, with

[24] As partly illustrated by the acceptance of Jordanous (2010b), Jordanous and Keller (2011) to musicology conferences.
[25] This is discussed more in Chapter 10.

limited time and information, in an evaluation of five systems presented at ICCC'11.[26]

Ch.8 Application of other methods of creativity evaluation to the systems in the two case studies and critical comparison of the SPECS methodology with other creativity evaluation methods.

Ch.9 Reflections on how the SPECS methodology has been applied in the two case studies, including discussion of practical and theoretical issues that arose. The reactions of the computational creativity research community to earlier versions of this work are also discussed.

Ch.10 Summary of the SPECS methodology and conclusions about the contributions to knowledge made in this thesis, with suggestions for future methodology development and for adoption of the tool by the computational creativity research community.

1.8 Summary

This thesis focuses on evaluation of computational creativity, addressing the research question: how should we evaluate the creativity of a computational creativity system? As Sections 1.2 and 1.3reported, there is a clear need for evaluation in the field of computational creativity research, to track progress and identify strengths and weaknesses in our research.

A key assumption is made in this work, that it is possible for computers to be creative (Section 1.4). This is by no means a universal assumption; resistance to the idea of computational creativity abounds. Section 1.4 acknowledged this reaction against computational creativity, but concluded that for useful progress in computational creativity the assumption in Section 1.4 is necessary.

Computational creativity is a relatively new research area, growing significantly in recent years, particularly the past decade (Section 1.5). During this development, evaluation has been a recurring topic for discussion. The question of computational creativity evaluation itself has proven non-trivial and has been considered one of the 'Grand Challenges' of computational creativity research (Cardoso et al., 2009). This current work is intended as a significant, practical methodological contribution towards addressing this Grand Challenge. The methodological approach taken is influenced by key findings in a wide range of creativity research, both computational and human, and by research on existing creativity evaluation methodologies and standard scientific methodology.[27]

The work reported in this thesis, or parts of thereof, is potentially of interest to a wider readership than its immediate target audience (computational creativity researchers wishing to evaluate the creativity of their systems). To this end, a *Readers' guide* to the thesis was provided in Section 1.7, to guide readers to the parts of the thesis most relevant to them.

[26] The 2011 International Conference for Computational Creativity.

[27] See Chapters 2 and 3.

Chapter 2

Existing methodologies and issues in computational creativity evaluation

---

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012).

Figure 2.1: *Wordle* word cloud of this chapter's content

Overview

The issue of how best to evaluate computational creativity systems has generated many more questions than answers. A chronological overview of the suggestion and development of evaluation metrics and methodologies is presented in Section 2.1 and summarised in Figure 2.3. This is followed by a more detailed examination of the main methodologies to date (Colton, 2008b; Ritchie, 2007, 2001; Pease et al., 2001), considering their strengths and weaknesses. Section 2.1 also looks at smaller-scale creativity evaluation metrics and strategies that have been employed by researchers for evaluation.

While this thesis advocates that computational creativity is not solely a scientific endeavour (as was expanded upon in Chapter 1 Section 1.5), this thesis does pursue methodological recommendations. It is useful, therefore, to consider how developments of the scientific method could be relevant in developing good methodological practice for computational creativity evaluation (Section 2.2).

To identify trends in current evaluative practice for computational creativity and see how (and if) the contributions outlined in Sections 2.1 and 2.2 are applied in practice, Section 2.3 surveys 75 recent papers on creative systems, examining how each system is evaluated. We see that no evaluation methodology has become the standard for evaluating computational creativity systems. This is in no small part due to the number of practical and theoretical issues surrounding such an evaluation, which largely remain unresolved despite the research community's awareness and informed consideration of such issues (Section 2.3.5). Existing evaluation methodologies have their critics and there is no consensus within the computational creativity community about which methodology to adopt.

2.1 Existing evaluation methods

2.1.1 Overview of key evaluation methods in computational creativity

Overall, the major contributions to how to evaluate computational creativity systems have been in the last decade (Ritchie, 2001, 2007; Pease et al., 2001; Colton, 2008b). Prior to 2001, little attention was paid to devising practical methodologies for computational creativity evaluation, despite repeated calls for such work (Boden, 1990, 1994a; Bundy, 1994; Boden, 1998, 1999; Colton & Steel, 1999; Colton, Bundy, & Walsh, 2000; Wiggins, 2000).[1]

2001 was a fruitful year for computational creativity evaluation, with the publication of three relevant papers:[2] Ritchie's criteria approach (Ritchie, 2001),[3] the combination of tests in Pease et al. (2001) and investigations on how input to a creative system determines its output (Colton et al., 2001). Sections 2.1.2, 2.1.3 and 2.1.6 respectively report more details of each of these discussions. Only Ritchie (2001) has had lasting practical impact, being used a number of times for evaluation since 2001 (e.g. Gervás, 2002; Pereira, Mendes, Gervás, & Cardoso, 2005; Haenen & Rauchas, 2006).

[1] Such calls are easier to make than to fulfil.

[2] Somewhat reminiscent of the saying about London buses: 'you wait for a long time then lots turn up at once'.

[3] The criteria approach in Ritchie (2001) was later revised in Ritchie (2007).

The *creative tripod* evaluation framework (Colton, 2008b) was recently proposed, highlighting three qualities a creative system must display in its behaviour if it is to be deemed creative: *skill*, *imagination* and *appreciation*. This framework is summarised in Section 2.1.4. Ventura (2008) investigates whether the combination of the more informal creative tripod framework and Ritchie's formal criteria is sufficient to evaluate creative systems. Section 2.1.6 reports how Ventura reaches negative conclusions on this matter. Colton's approach, detailed in Section 2.1.4, has been adopted by several authors in the few years it has been available so far (e.g. Chan & Ventura, 2008; Colton, 2008c; Norton, Heath, & Ventura, 2010; Monteith, Martinez, & Ventura, 2010; Young & Bown, 2010). Generally this approach has been used by authors as a descriptive justification for why a particular system should be deemed creative but not for comparison of systems.

As well as providing a way to evaluate the creativity of a computational system, a key function of a creativity evaluation methodology is how it can be used to compare systems against one another on the basis of the level of creativity demonstrated. In practice, Ritchie's approach (outlined in Section 2.1.2) tends to have been the most frequently adopted quantitative evaluation method by the computational creativity community (e.g. Gervás, 2002; Pereira et al., 2005; Haenen & Rauchas, 2006). In comparison, although often cited, the tests offered by Pease et al. (2001) have not been applied in practice. As argued in Section 2.1.3, though, this paper makes a significant contribution to the literature on computational creativity evaluation.

Most recently, the FACE and IDEA models have been proposed (Pease & Colton, 2011b; Colton, Charnley, & Pease, 2011; Pease & Colton, 2011a; Charnley, Pease, & Colton, 2012). The FACE model is designed to represent creative acts and the IDEA model represents the impact of those acts. As reported in Section 2.1.5, the models collectively aim to distinguish whether a system is acting creatively or not, and whether an artefact is valuable or not.

The above-mentioned contributions and related work shall now be examined in closer detail.

2.1.2 Ritchie's empirical criteria for computational creativity

> 'The most extensive contribution to the discussion of how we could assess creativity in software has come from Ritchie' (Colton, 2008b, p. 2)

Graeme Ritchie has proposed a set of formal empirical criteria for creativity. The criteria were originally introduced in Ritchie (2001) then were revised and updated in Ritchie (2007). The criteria are situated in an overall framework describing the design and implementation of a creative computational system in set-theoretic form. Ritchie advocates post-hoc analysis of artefacts generated by the system, disregarding the process by which they were created. For systems that produce abstract rather than concrete results (Ritchie gives the example of analogies), Ritchie's approach is not applicable.

The criteria collectively describe aspects of the *typicality* and *quality* of the output of the creative system (and indirectly, the novelty of the system output). Two key mappings are used in the criteria:

typ - a rating of how typical the output is in the intended domain

> 'To what extent is the produced item an example of the artefact class in question?' (Ritchie, 2007, p. 73)

val - a rating of how valuable the output is

> 'To what extent is the produced item a high quality example of its genre?' (Ritchie, 2007, p. 73)

Ritchie emphasises the importance of assessing computer-generated artefacts both in terms of how typical an example they are of items in the target domain and in terms of atypicality. Further to this, an artefact may be typical of the domain but not be a good example, so the value rating is introduced to assess the quality of that artefact.

Originally a set of 14 criteria in Ritchie (2001), four new criteria were added and two existing criteria revised in Ritchie (2007).[4] Ritchie (2007) states that the criteria are not intended as a model of creativity or of the creative process, or as necessary evidence for creativity. Instead the criteria are formally defined observable factors that relate to creativity; exactly how these criteria should be implemented is left open to debate. The criteria can be combined in various ways, weighted or left out entirely as appropriate for the given creative domain. As well as weighting individual criteria, each criterion is parameterised, allowing further customisation of the criteria to individual definitions of creativity for different domains or systems.

The Criteria

The formal definitions of the 18 criteria can be found in Ritchie (2007) and in Appendix F. Here, the criteria are deliberately presented informally, with descriptors such as 'suitable' and 'high' substituted for the parameters left unspecified by Ritchie. It is hoped that any subsequent loss in formal semantics is balanced by a more immediate understanding of each criterion.

1. On average, the system should produce suitably typical output.
2. A decent proportion of the output should be suitably typical.
3. On average, the system should produce highly valued output.
4. A decent proportion of the output should be highly valued.
5. A decent proportion of the output should be both suitably typical and highly valued.
6. A decent proportion of the output is suitably atypical and highly valued.
7. A decent proportion of the atypical output is highly valued.
8. A decent proportion of the valuable output is suitably atypical.
9. The system can replicate many of the example artefacts that guided construction of the system (the *inspiring set*).
10. Much of the output of the system is not in the inspiring set, so is novel to the system.
11. Novel output of the system (i.e. not in the inspiring set) should be suitably typical.

[4] In Ritchie (2007), criteria 15-18 were added to the Ritchie (2001) set and criteria 8 and 10 were revised slightly.

12. Novel output of the system (i.e. not in the inspiring set) should be highly valued.
13. A decent proportion of the output should be suitably typical items that are novel.
14. A decent proportion of the output should be highly valued items that are novel.
15. A decent proportion of the novel output of the system should be suitably typical.
16. A decent proportion of the novel output of the system should be highly valued.
17. A decent proportion of the novel output of the system should be suitably typical and highly valued.
18. A decent proportion of the novel output of the system should be suitably atypical and highly valued.

The set of criteria[5] includes contradictions and inconsistencies between individual criteria, for example criteria 5 and 6, or criteria 17 and 18. This is probably to be expected for a concept like creativity which contains much ambiguity and is resistant to capture in a formal definition.[6]

Ritchie has no plans to carry out any further work on the criteria (Ritchie, 2009, personal communications), despite acknowledging the potential need for further development:[7]

> 'We do not claim that all these criteria are essential to an assessment of creativity, nor that they are the only such criteria that could be considered; rather, they are a first draft of a general catalogue of relevant factors' (Ritchie, 2007, p. 7)

Applications of Ritchie's criteria

Ritchie (2007) analyses how the 2001 version of the criteria has been interpreted and applied (Gervás, 2002; Pereira et al., 2005; Haenen & Rauchas, 2006). In practice, researchers have either chosen arbitrary levels for each of the parameters or re-used previously chosen parameter values for comparative purposes, with each criterion weighted equally in the overall evaluation.

Ritchie criticised these implementation decisions as compromising the evaluative power of the approach, though Ritchie's original recommendations on these matters were vague. It would have been useful for Ritchie to demonstrate how his criteria could be implemented in practice. There is data available for such a demonstration; (Binsted, Pain, & Ritchie, 1997) reports ratings of typicality and value of the products of JAPE, a joke generation system that Ritchie is involved with, but a criteria-based evaluation does not appear to have been performed to date.

The closest that Ritchie comes to using the criteria in practice is in Ritchie, Manurung, Pain, Waller, Black, and O'Mara (2007), on the STANDUP system for joke generation; however rather than applying the criteria, the paper only mentions them briefly to say that the criteria are not appropriate for evaluation in this case:

> 'these [Ritchie's] criteria are insufficiently subtle for making fine comparisons of creativity' (Ritchie et al., 2007, p. 97)

[5] Ritchie does not discuss how these criteria were selected for inclusion from a set of potential candidate criteria.

[6] See Chapter 3 for further discussion of this point.

[7] The criteria have been described as 'well-developed' (Colton, 2008b, p. 3) and have not been revised again since 2007.

Comments on Ritchie's approach

In his approach, Ritchie prioritises tests for artefact *novelty* and *quality*, disregarding as minor and unimportant other 'subsidiary tests of creativity' (Ritchie, 2001, p. 4) such as the autonomy of the creator's actions. The basis for this decision is not justified, except by statements such as:

> 'If a person produces a painting which is radically different from previous work ... and which is definitely a good painting, then that will usually be deemed creative. What is rarely brought into the assessment is *how* the person came up with the idea/artefact' (Ritchie, 2001, p. 4)

No supporting evidence for statements like this is offered. Many authors have questioned whether assessing creativity should involve more than solely judgements of the end product (e.g. Cohen, 1999; Pease et al., 2001; Boden, 2004; Colton, 2008b; Kaufman, 2009; Odena & Welch, 2009).[8]

Ritchie often uses non-committal language, for example [my emphases added]:

> 'we list a number of criteria which *might* contribute to a decision about creativity' (Ritchie, 2007, p. 67)
>
> 'define .. the attributes one *might* look for in order to decide whether or not a particular computer program had behaved creatively' (Ritchie, 2001, p. 4)
>
> '[we] state some criteria which *could* be applied in deciding how creative a program is, or has been' (Ritchie, 2007, p. 7)

Whilst this cautious approach guards Ritchie from full commitment to these views, it dilutes the strength of conviction of these recommendations.

Ritchie makes no recommendations on exactly how the criteria should be implemented in assessment and no example implementations are given. This leaves open questions of how best to choose appropriate values for parameters, criterion weights and even what criteria to include in assessment. To facilitate comparisons between systems, some standard settings can be adopted (Gervás, 2002; Pereira et al., 2005; Haenen & Rauchas, 2006).[9] To deal with this issue rigorously, Ritchie suggests generating values based on extrapolating from human assessment (essentially turning this into a supervised learning problem). Now, though, the researcher seeking an evaluation method has a new multi-dimensional, possibly non-linear problem to solve in order to extrapolate these values.

The distinction between being a typical member of a set and being a valued member of a set is one that is often highlighted (e.g. Wiggins, 2006a; Boden, 2004; Forth, McLean, & Wiggins, 2008). These two ratings are not necessarily distinct in Ritchie's definitions listed above and are difficult to separate completely in implementation, particularly if the two ratings are being produced by user assessment. For example, in Jordanous (2010c),[10] the criteria were implemented as an interactive fitness function for creativity in a genetic algorithm system but the user struggled to provide independent ratings for typicality and value. This issue arises again in Ventura (2008) where the two

[8] This debate is explored further in Chapter 5 Section 5.1.2 and also in Chapter 8.

[9] As Chapter 3 Section 3.6.4 will discuss, there is variety as to what is important for creativity in different domains, so questions arise as to the suitability of a standard set of parameter settings for all creative systems in all creative domains.

[10] The system in Jordanous (2010c) is included in Case Study 1 and will be described in Chapter 6 Section 6.2.

rating schemes are amalgamated into one rating, of image 'recognisability' (Section 2.1.6). Ritchie proposes that ratings schemes be formally outlined to avoid interpretation problems; another solution is to be careful in how to phrase the evaluation rating questions to users, as is done to good effect in Binsted et al. (1997) but less well in Ritchie (2007).[11]

Criteria 9 to 14 rely on an *inspiring set*: the knowledge or examples guiding program construction. Ritchie acknowledges that sometimes the information in this set may not be available, such as when no examples are explicitly used by the system to generate new examples, but does not make practical recommendations for how the criteria should be adapted for this situation.

In general, Ritchie's proposals acknowledge a number of theoretical issues, but are relatively impractical to apply for evaluation. Several implementation decisions are left to the choice of the evaluator, which has stimulated important discussion on how to make such implementation decisions accurately and has encouraged researchers to justify their decisions (Gervás, 2002; Pereira et al., 2005). This flexibility could however also possibly be taken advantage of for a more favourable evaluation of a particular system (for example one could weight highly the criteria where the system performs well, or tweak parameters to best fit the system's interpretation of creativity). The slight opaqueness of the formally stated criteria makes this harder to detect on inspection. Some researchers, including Ritchie himself, have also reported issues with implementing Ritchie's criteria (Ritchie et al., 2007; Jordanous, 2010c; Tearse, Mawhorter, Mateas, & Wardrip-Fonin, 2011). Applications of the criteria have dropped in recent years, in favour of other options or no evaluation of creativity.[12]

2.1.3 Pease et al.'s tests on the input, output and process of a system

Pease et al. (2001) discuss a wide range of aspects involved in evaluating the creativity of artificial systems. A combination of evaluative tests are proposed, based on:

- The input provided to the system.
    - ∗ *Input Measure:* items produced by the system are 'potentially creative' (Pease et al., 2001, p. 2) if they are not part of the inspiring set.
- The output produced by the system.
    - Novelty Measures.
        - ∗ *Transformational Measure:* categorising the novelty of items according to how they were generated from procedures within the program: *fundamental* novelty if a produced item is only possible once generation procedures have been developed in some way by meta level procedures; *mere* novelty if an item is produced using the procedures already in the program; and *none* otherwise.

[11] This will be demonstrated in Chapter 8 Section 8.2.1.

[12] As shall be shown in Section 2.3.

    - ∗ *Complexity Measure:* assessing the complexity of items by how complex and novel the generating process was, taking into account the size of the domain.
    - ∗ *Archetypal Measure:* how close in similarity a system product is to the nearest (previously-identified) archetypal example.
    - ∗ *Surprise Measure:* measuring the surprisingness of an event based on the probability of similar events occurring in that context.
    - ∗ *Perceived Novelty Measure:* human judges rate how novel they find items generated by the system, compared to systems generated by humans; the responses are used to categorise items as novel or not, accordingly.
  - Quality Measures.
    - ∗ *Emotional Response Measure:* human judges evaluate to what degree an item has affected them positively or negatively; the responses are used to categorise items according to the intensity of the response.
    - ∗ *Pragmatic Measure:* using unspecified (domain-specific) 'marking criteria' (Pease et al., 2001, p. 6) to judge to what extent an item meets an aim.
- The process(es) employed by the computational system.
  - Generation.
    - ∗ *Randomness Measure:* using probabilities of items being generated given specific input, with measures of difference between two items, for a measure of randomness that can be used to categorise items by randomness (high, low or none).
  - Evaluation.
    - ∗ *Evaluation of Item Measure:* statistically measuring how closely self-evaluation of an output item by the system correlates with external evaluation of that item.
    - ∗ *Evaluation of Process Measure:* comparing the quality measures from above on two comparable sets of output items. One set is produced by methods which can be transformed internally during program run-time and one by methods which cannot. The quality of the first set should exceed the quality of the second.

As in Section 2.1.2, these tests are summarised in informal language above.[13] All the tests generate quantified or categorised representations of the relationships between observable parts of the system and responses to the system.[14]

[13] For more formal descriptions of the tests, the interested reader is referred to their full presentation in Pease et al. (2001).

[14] The presentation of these formal statements may have discouraged some from applying the tests. The layout is dense with a lack of white space (probably due to space restrictions), making it more difficult to find each test within the paper.

Pease et al. state a number of assumptions that they make about the nature of creativity prior to determining the measurement methods proposed:

- (At least some part of) creativity is domain-independent so generalised measurements are preferable where possible. Even if this assumption is incorrect, it will still be valuable in revealing parts of creativity which are linked to specific domains but which we may have presumed were domain-independent.
- We should use human creativity as a guide for computational theory and as a yardstick to judge progress against.
- There is no distinction between genius and everyday creativity:

  > 'We assume that creativity in children, adults and geniuses is essentially the same phenomenon.' (Pease et al., 2001, p. 129)

- Creativity is (as mentioned above) a cyclic process of generating ideas and evaluating them.
- The creative process results in the production of item(s) (not necessarily physical objects).

Pease et al. admit that their choices of assessment methods are 'somewhat arbitrary' (p. 137). Like Ritchie (2007), they see their tests as initial suggestions, hoping to prompt further discussions along similar lines. Of the authors of Pease et al. (2001), some subsequent recommendations are made for creativity evaluation (Colton, 2008b; Pease & Colton, 2011b), but these are unrelated to those in Pease et al. (2001); in fact Pease et al. (2001) is not cited in these later publications.

Pease et al. (2001) discuss *meta-evaluation*, in other words how to evaluate a proposed evaluation. To assess their proposed evaluation tests, Pease et al. offer the following two criteria to consider:

1. How closely the proposed test maps to how humans evaluate creative behaviour.
2. How possible and practical it is to apply the test to systems.

Pease et al. claim that the first criterion can be evaluated empirically, presumably through experiments with human participants, and that the second can be evaluated by implementing the test methods in artificial intelligence programs. These claims, however, are by no means rigorously explored in Pease et al. (2001) or elsewhere. This discussion, although seemingly an obvious inclusion for authors writing about the importance of evaluation, is missing from other notable proposals on creativity evaluation (Ritchie, 2001, 2007; Colton, 2008b) and deserves more attention generally than the paragraph it receives in Pease et al. (2001).[15]

Communications between Pease et al. and Ritchie

The tests in Pease et al. (2001) are described by Colton et al. (2001, p. 6) as a set of 'general guidelines' to complement other 'concrete measures' such as Ritchie's criteria (Ritchie, 2001) or Colton et al.'s fine-tuning measure. Ritchie (2007) introduced Pease et al.'s tests as an 'elaboration' of the empirical criteria (Ritchie, 2007, p. 81). Pease et al. acknowledge Ritchie's work, describing

[15] In this thesis, Chapters 8, 9 and Appendix H return to the issue of meta-evaluation in greater depth.

their paper as an extension of Ritchie's framework (Pease et al., 2001, p. 129) and suggesting a number of refinements to the framework:

- The notion of the *inspiring set* is defined more formally (pre-empting criticisms of the vague specifications of the inspiring set (e.g. Ventura, 2008)).
- A distinction is made between *strong* (I_S) and *weak* (I_W) versions of the inspiring set:[16]
  - I_S: all items known by the programmer.
  - I_W: all items that the programmer deems to have influenced program construction.
- The typicality rating scheme, heavily relied upon but left unspecified in Ritchie (2001), is superseded by a formally stated 'Archetypal Measure', identifying typical (archetype) items and measuring how close in similarity a system product is to its nearest archetype.

Graeme Ritchie had some influence on the writing of Pease et al.'s paper, being thanked in the Acknowledgements section for his 'helpful comments on earlier drafts' (Pease et al., 2001, p. 137). In a later publication (Ritchie, 2007) on his evaluation criteria, though, Ritchie distances himself from Pease et al. (2001), criticising their approach in several ways:

- The tests proposed by Pease et al. are too unconnected and need to be more coherently linked.
- Pease et al. do not adequately justify their decisions on what aspects of creativity to test.
- Ritchie cannot see how Pease et al.'s tests could be incorporated in his formal framework:

> 'Although PWC [Pease et al. (2001)]'s proposals at first glance appear to be an elaboration of our basic framework, on closer inspection it is not clear where they would fit into, or alongside, our formalisation.' (Ritchie, 2007, p. 83)

2.1.4 Colton's creative tripod framework

The creative tripod framework (Colton, 2008b) is the culmination of discussion in several of Colton's previous publications (Colton & Steel, 1999; Colton et al., 2000, 2001; Pease et al., 2001; Hull & Colton, 2007). Colton emphasises the importance of considering the creative process when evaluating the creativity of a computer system. This follows from Pease et al. (2001) and represents a pointed departure from the product-focused standpoint taken in Ritchie (2001).[17]

The creative tripod represents three qualities that a creative system must demonstrate to some degree, in order to be considered creative: *skill*, *imagination* and *appreciation*. If a creative system can demonstrate each of these three behaviours, then Colton argues that this is sufficient for the system to be perceived as creative. Here Colton makes an important distinction; rather than positing the creative tripod qualities as necessary and sufficient conditions for the system to *have* if it is to be deemed creative, he argues that the system should only be *perceived* to possess these qualities. In

[16] This distinction is reminiscent of the division of artificial intelligence into *strong* and *weak* versions.

[17] Chapter 5 Section 5.1.2 examines the product/process debate in more detail.

other words, the challenge is to make a system appear to be creative to its audience, rather than to develop some level of creativity which exists in the system independently of an audience's perception.

Table 2.1 shows similarities between these three tripod qualities and the abilities outlined by Rabinow (as reported in Csikszentmihalyi (1996)) as necessary for a creative thinker: knowledge, motivation to explore, and evaluation. Parallels can also be drawn between the creative tripod qualities and Cohen's requirements for a 'behaviour X':[18] knowledge, emergence, awareness of what is emerging, and the motivation to act on what has emerged (Cohen, 1999, pp. 29-30).

Table 2.1: Parallels between Colton's creative qualities (Colton, 2008), Rabinow's requirements for creative thinkers (Csikszentmihalyi, 1996) and Cohen's 'behaviour X' (Cohen, 1999).

Rabinow	***Colton***	**Cohen**
Knowledge	*Skill*	Knowledge
Motivation to explore	*Imagination*	Emergence
Evaluation	*Appreciation*	Awareness, Motivation to act

The creative tripod provides three standard descriptors for creative behaviour, both for post-hoc assessment and during system development. Unlike Ritchie, Colton demonstrates how he envisages his framework being used for creativity evaluation, using his own systems as examples:[19]

- HR: a mathematical conjecture/concept discovery program (Colton, 2002).
  - **skill**: HR can generate new concepts from existing known concepts or carry out proofs/disproofs from empirical evidence.
  - **imagination**: HR can search the space of possible concepts and conjectures and make discoveries. For example HR discovered a conjecture about 'refactorable' numbers (Colton, 2008b) that was unknown to the system programmer (Colton).[20]
  - **appreciation**: HR demonstrates appreciation of the conjectures it generates through its measures of 'interestingness' (Colton et al., 2000).
- The Painting Fool: an artistic image generator (Colton, 2008b, 2008c).
  - **skill**: The Painting Fool can generate images in various different styles.
  - **imagination**: The Painting Fool takes an evolutionary approach to scene generation and can generate scenes it has not seen before.
  - **appreciation**: The Painting Fool can recognise emotion portrayed in portraits (via a machine vision module to assist this behaviour and make the system more autonomous).

[18] Cohen (1999) repeatedly uses the term 'behaviour X' as a veiled reference for creative behaviour.

[19] The issue of who decides if a system demonstrates the tripod qualities is not raised. In his examples, Colton evaluates his own systems, rather than using external judges, without discussing whether any bias is introduced as a result.

[20] This conjecture was considered publishable by the Journal of Integer Sequences (Colton, 1999).

Figure 2.2: The Creative Tripod, outlined by Simon Colton in 2008.

Continuing the tripod analogy, each leg of the tripod's legs is described as having three sections (pictured in Figure 2.2), each representing one perspective on the system: the *programmer*, the *computer system* and the *consumer* of the system. Each of these three parties can contribute skill, imagination and appreciation. For example, if some form of skilful behaviour is demonstrated from each of the three perspectives, relative to the system, then the 'skill' leg would be fully extended.

To determine whether a system is creative or not, however, Colton is interested solely in the *system*'s behaviour. Any contributions from the consumer or the programmer are irrelevant in determining *if* a system can be perceived as creative; they only contribute to *how* creative a system is perceived to be, once the system has demonstrated skill, imagination and appreciation.[21] Colton argues that if a creative system does not demonstrate these three behaviours, then that the system should not be perceived as creative.[22]

> 'Our position is that, if we perceive that the software has been skillful, appreciative and imaginative, then, regardless of the behaviour of the consumer or programmer, the software should be considered creative. Without all three behaviours, it should not be considered creative, but the more aspects which extend each leg of the tripod, the more creativity we should project onto the software.' (Colton, 2008b, p. 18)

It is unclear why Colton chose to discuss the programmer and consumer in relation to the tripod, if their contribution is not deemed crucial for determining the perception of creativity; the opinions of the audience are already included in how the creativity of the system is perceived by others. Also, Colton does not extend the tripod analogy further to consider how balanced the tripod is. If one of the three qualities is extremely well-represented in the system compared to the others, for example a

[21] Chapter 3 Section 3.6.4 considers the appropriateness of a single definition of creativity to cover all types of creativity.

[22] Here Colton makes an important distinction; rather than positing the creative tripod qualities as necessary components of a creative system, he argues that the system merely needs to be *perceived* to have these qualities. In other words, the challenge is to engineer a system that appears to be creative to its audience, rather than engineering a system that possesses a level of creativity existing independently of an audience's perception.

system that is highly skilful but weak in imaginative and appreciative capabilities, the tripod would be unbalanced and unstable, disguising weaknesses in some areas through its strengths in others.

2.1.5 Computational Creativity Theory: the FACE/IDEA models

The FACE and IDEA models form part of a wider research project to formally develop Computational Creativity Theory (Pease & Colton, 2011b; Colton et al., 2011; Pease & Colton, 2011a; Charnley et al., 2012). The FACE model is designed to represent creative acts and the IDEA model represents the impact of those acts. Collectively the models aim to distinguish between whether an artefact is valuable or not, and whether a system is acting creatively or not, with focus on the latter.

Pease and Colton (Pease & Colton, 2011b) partly motivate their models in response to a consideration of how versions of the Turing test have been applied in discrimination tests (Pearce & Wiggins, 2001) and as a direct test of the prevalent definition of computational creativity as tasks which if performed by humans would be considered creative. The Turing test (as adopted above) is criticised for various reasons: different styles of creativity are not equally recognised, or different manifestations of creativity across domains; contextual 'framing' information is ignored and evaluations are performed independently of context; there are opportunities to perform well on the test by 'window dressing' or producing shallow imitations ('pastiche') at the expense of genuine creativity; and evaluation benchmarks are high for computational systems if systems are to be judged at human standards of creativity (especially as the Turing test has not yet been passed by intelligent systems).

Pease and Colton suggest as an alternative the FACE and IDEA model. The FACE model *(Frame, Aesthetic, Concept, Expression of concept)* represents measures and measurement methods on context, aesthetics, concept(s) of interest and how they are expressed, respectively. For each of the *F, A, C* and *E* items, tuples represent a method for generating information on that item and a representation or measure of the item itself. The IDEA acronym represents the *Iterative Development Execution Appreciation* cycle, which assumes an ideal audience *i* and measures the effects that one creative act *A* has on *i*, such as change in well-being (*wb*) or the cognitive effort for appreciation (*ce*).

Several quantitative measures are proposed within this model, measuring disgust, divisiveness, indifference, popularity, provocation, acquired taste, instant appeal, opinion splitting, opinion forming, shock and subversion. Some assumptions are made (but not yet fully justified) by the authors as to why certain outcomes are preferable, such as the amount of cognitive effort or the extent to which a creative system arouses divisiveness in its audience.

At this early stage of development (the FACE and IDEA models were first published in mid-2011), FACE and IDEA are proposed not as the end solution for evaluation but as a 'beginning in our efforts to avoid some of the pitfalls of the TT' (Pease & Colton, 2011b, p. 7). There are plans to develop sub-models of aspects of creativity, with several suggestions listed, including: affect, analogy, appreciation, audience, autonomy, blending, community, context, and curiosity. The end goal for this

work is a comprehensive, detailed formalisation of computational creativity:

> 'Using the foundational terminology for creative acts and impact described above, we plan to expand each term into a formalism containing conceptual definitions and concrete calculations using those definitions which can be used for the assessment of creativity in software. In doing so, we hope to contribute a *Computational Creativity Theory* which will provide a strong foundation for objectively measured progress in our field.'

Some brief examples exist of the FACE model being applied to describe systems (Colton et al., 2011) or guide their development (Colton, Goodwin, & Veale, 2012). Currently though, full worked examples of the FACE/IDEA models being used for evaluation of computational creativity have not yet been published by the team behind Computational Creativity Theory or by other researchers. This is probably due to the stage of development it is in; this is a current project, with the models being developing on an ongoing basis, even occasionally to the level of the precise terminology being used.[23] With these ambitious aims for this current project, however, it will be interesting to see how this work develops on its promising potential.

2.1.6 Combining creative evaluation methodologies together

Colton et al.'s addition to Ritchie's criteria

Colton et al. (2001) is a collaborative contribution from Graeme Ritchie and two authors of Pease et al. (2001), Simon Colton and Alison Pease. Working within the formal framework described in Ritchie (2001), Colton et al. investigate how input affects program output, specifically whether the program overfits to its input ('fine-tuning') rather than producing new items. They conclude that creative programs should be as general as possible and that the value of a program encompasses more than the ability to generate specifically-targeted artefacts.

Colton et al. (2001) extends Ritchie's criteria to include measures of the amount of 'fine-tuning' (replicating the input data rather than generating novel and valuable output) and estimates of the generality of the program. As with Ritchie's criteria, these measures require details of the inspiring set (items that influence the way the program is implemented, either directly or implicitly). If details of the inspiring set are not known, applying the measures becomes problematic.

Colton et al. (2001) is intended to complement Ritchie's criteria. When Ritchie's criteria have been applied for evaluation, though, Colton et al.'s proposals have either not been used (Haenen & Rauchas, 2006; Jordanous, 2010c) or have been considered separately to the criteria analysis, almost as an afterthought (Gervás, 2002; Pereira et al., 2005). Rather than take the opportunity to incorporate the contribution of Colton et al. in his updated version of the criteria, Ritchie (2007) merely reports the details of Colton et al. (2001) as a 'related proposal' (Ritchie, 2007, pp. 81-82).

[23] At the International Conference on Computational Creativity 2012, the 'E' of FACE was interchangeably referred to by the model authors as 'Example', rather than 'Expression'.

The other contributions of Colton et al. (2001) are to cast Ritchie's proposals as a guide for program construction rather than post-hoc assessment and to stress there is no 'right answer' when assessing creativity. The latter leads to the conclusion that there is little worth in comparing systems:

> 'Under certain circumstances, it may be possible to use such measures [measures to estimate the creativity of a program] to determine whether one program is more creative than another. However, the circumstances would have to be very special, taking the design, input and output of both programs into consideration, and it is still likely that such a comparison would be deemed unfair. (Colton et al., 2001, p. 1)

Colton et al. (2001) conclude from this that formative feedback is more useful than summative evaluation, a conclusion that this thesis agrees with. To disregard comparative evaluation completely, though, is contentious. People can make comparisons such as 'x is more creative than y', even though there is rarely any ground truth or correct answer to compare such statements to. Also, how do we show progress in our research if we cannot compare our systems against previous work?[24]

Since the publication of this paper, each author of Colton et al. (2001) has retracted the view that comparison between systems is unfair and that there is little worth in using measures of creativity to assess computational systems after implementation. Ritchie (2007, also 2009 personal communications) continues to present his criteria as a tool for post-hoc analysis of creative systems. Pease and Colton (2011b) investigated the application of the Turing Test (as inspired by Turing, 1950) for computational creativity evaluation (finding it lacking). Colton (2008b) offered a new evaluation framework for creative systems (as described in Section 2.1.4). In a recent description of the progress of computational creativity research, Colton (2008a) advocated the comparison of creative systems:

> 'We've recently reached the stage where there is a sufficiently large number of such programs for us to be able to compare and contrast them in a meaningful way.' (Colton, 2008a, p. 6)

Incorporating Ritchie's criteria and Colton's creative tripod in a combinatory framework

The creative tripod is presented by Colton (2008b) as an informal alternative to Ritchie's formal criteria approach (Ritchie, 2007). Ventura (2008) tests the combination of the 2007 version of Ritchie's criteria and Colton's creative tripod, in a thought experiment using an imaginary artistic computer system called RASTER. RASTER stands for 'Ridiculous Artist Supporting Thought Experiment Realization'; as this acronym indicates, RASTER is intentionally designed to be uncreative. The system generates images by comparing random binary patterns to existing images retrieved online, returning any random patterns that are close enough to replicate the existing image being checked.

Ventura's motivation is to test the sufficiency (as opposed to necessity) of this combination: is there an element of creativity (or a decidedly non-creative element) that the Ritchie-Colton combination

[24] As later Chapters in this thesis will show (Chapters 6, 7, 8), although Case Study 2 found only limited value in comparison, some interesting observations came out of this comparison from learning from other systems strengths. Case Study 1 demonstrates how direct comparison between related systems supplied much valuable feedback for system improvements.

framework is not able to detect? By exploring exactly what this combination does or does not provide, Ventura aims to identify how the framework can be made more robust.

During the thought experiment, a key decision is made:

> 'In our *gedanken* experiment, since we are interested simply in (artefact) class membership (how recognizable an image is), the two rating schemes of the 3-tuple become redundant (that is, **typicality and quality are equivalent for this task** [this emphasis added]), and we can think of the 2-tuple definition as capturing this by compressing both into the single rating scheme *r*.' (Ventura, 2008, p. 12)

If typicality and value ratings are treated as the same rating *r*, Ventura shows that several of the criteria reduce to duplicates of other criteria. Only 3 distinct criteria remain for RASTER. This is partly because Ventura claims RASTER has an empty inspiring set, i.e. he is stating that there are no items that inspire or guide the implementation of the program. The online database of images is not an inspiring set, argues Ventura, because they do not guide the search program within the set. Whilst Ventura concedes that one could question his arguments for choosing not to use the online database as an inspiring set, he is deliberately highlighting and exploiting the weaknesses in Ritchie's definition of the inspiring set, a definition which is surprisingly informal given the thorough formal treatment applied to the majority of the content of Ritchie (2007).

7 distinct criteria remain for iRASTER, which is the same as RASTER except that it applies genetic operators to an inspiring set of images to generate potential output patterns instead of randomly generating patterns to be tested. This leads to the suggestion of a 'meta-heuristical application' (p. 15) of Ritchie's criteria:

> 'if most of the criteria *can not* be applied, then the system in question may not be a likely candidate for attribution of creativity.' (Ventura, 2008, p. 15)

Whilst this is a sensible observation, there is nothing in Ritchie's papers prohibiting the use of only three criteria to assess systems; this is allowable in Ritchie's proposals. As highlighted in Section 2.1.2, Ritchie repeatedly says that the criteria are not intended as a definitive set but instead as a set of possible tools for analysis from which the researcher can pick. No criteria are advocated as essential and no recommendations are made for what criteria to include or exclude, how many criteria are needed and how each criterion should be weighted in a summary measure of results. This presentation of the criteria affords little protection against attacks such as Ventura's.

A slightly different conclusion can also be drawn from Ventura's arguments. As Ventura is rating artefacts of RASTER and iRASTER for recognisability rather than typicality and value, he is arguably no longer assessing RASTER and iRASTER for creativity, but instead is assessing for value (in terms of recognisability). This slightly undermines Ventura's contribution, as the criteria become less fit for the purpose that Ventura is applying them for. Ventura almost reaches the same conclusion after seeing that most criteria reduce to duplicates of each other when the typicality and value rating schemes are combined, but criticises the applicability of the criteria rather than the application.

Ventura (2008) argues that it is possible to engineer Ritchie's criteria and Colton's qualities to find uncreative systems to be creative. To counter this, Ventura recommends that creativity assessment should also incorporate measures of efficiency in terms of output quantity during the system runtime and the ratio of items produced to items discarded during the creative process. These extra measures would help discredit RASTER from being judged creative, although iRASTER may still be found to be creative. This invites new counter-examples to then be proposed to highlight other inadequacies, such as a system that keeps all items produced, or one that does not produce output at all. One wonders if this would lead to a positive iterative process of improvement, or a more negative exchange haggling over insignificant details. Ventura also questions if the ability of a system to self-promote should be necessary for it to be deemed creative, though this is only posed as a closing question.

Rather than combining the two frameworks, Ventura mainly applies both frameworks separately to the same systems to see if they individually generate positive assessments of creativity. Most of the paper is devoted to looking at Ritchie's criteria (approximately 5 pages) as opposed to the combination of Ritchie and Colton's work (approximately 1 page in total), or Colton's creative tripod on its own (less than 1 page in total).[25] Although it is doubtful that alternative combination mechanisms would lead to new conclusions being reached, different combinations of the two frameworks could be beneficial, for example [my emphasis added]:

> 'we should be implementing Ritchie's criteria into artefact generation software, so that it could better *appreciate* its own creativity.'[26] (Colton, 2008b, p. 6)

Ventura notes that it would not be appropriate to formalise the three qualities that form Colton's tripod, because Colton (2008b) does not advocate that a system should *have* these three qualities, but the system should be *perceived* to have them. Ventura questions the informativeness of any rating system involving subjective judgements by humans. He seems to conclude that any evaluation by humans could be seen as uninformative so should therefore be disregarded:

> 'This ... leads us to ask whether any rating scheme that is subjective (human evaluated) is not *ipso facto* subject to a No-Free-Lunch type argument that demonstrates that it will be uninformative on average.' (Ventura, 2008, p. 18)

2.1.7 Other creativity metrics and evaluation strategies

Despite the recent attention paid to creativity evaluation, no one methodology is emerging as a standard tool. In the absence of an accepted standard methodology, several authors have proposed smaller-scale creativity metrics or evaluation strategies, as outlined below. To date, these proposed metrics and strategies have been used in isolation rather than in combination with other metrics. These metrics

[25] The rest of the 9-page paper gives details of the thought experiment systems or other details such as references.

[26] This has been attempted in Jordanous (2010c), where Ritchie's criteria were used as a fitness function for a genetic algorithm. This system is reported in more detail in Chapter 6 Section 6.2.

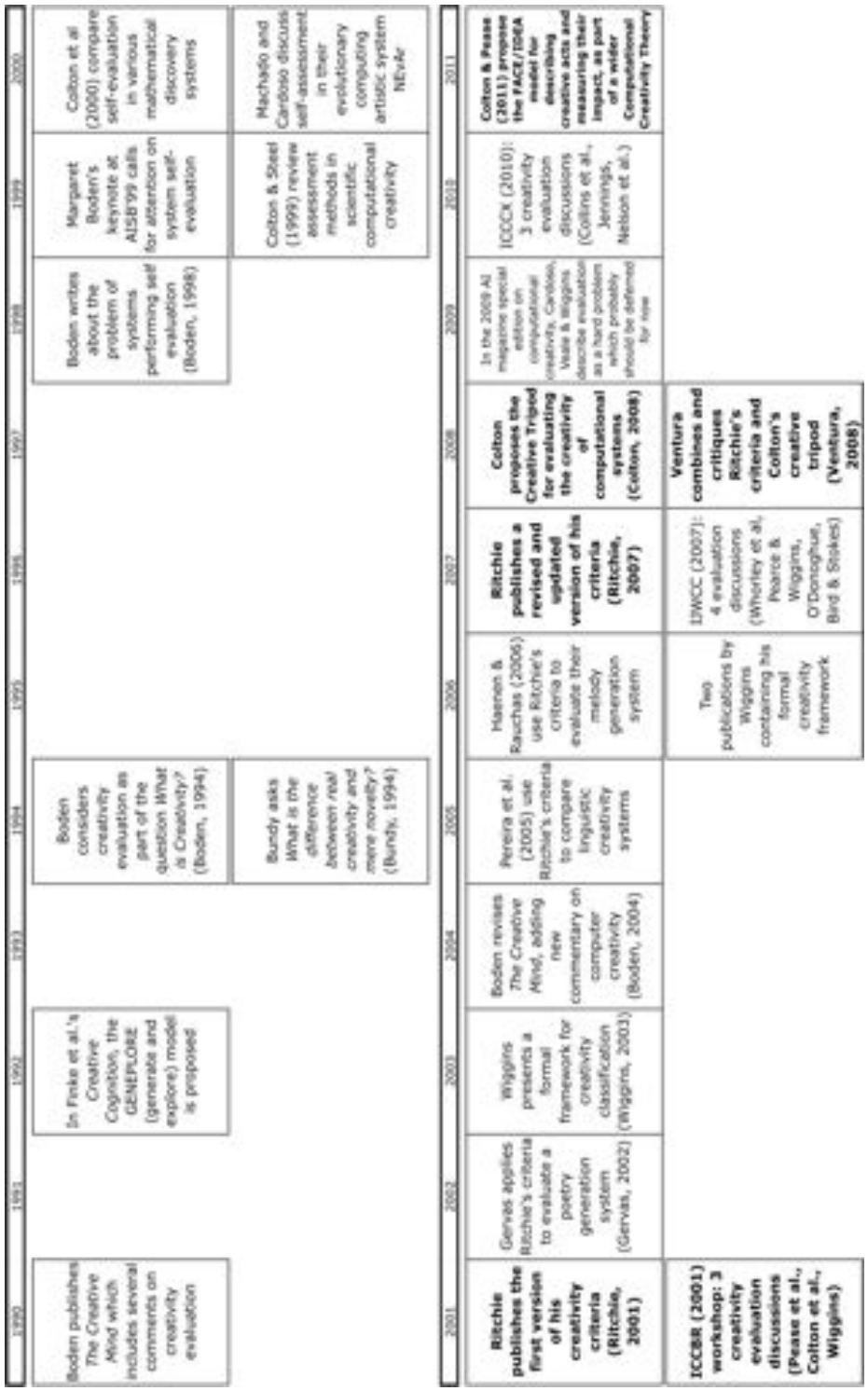


Figure 2.3: Timeline of important contributions to computational creativity evaluation from 1990 to present (2011). Major contributions are highlighted in **bold font**.

are not being re-applied by other researchers,[27] meaning that as yet the methods have acquired little practical significance on a research community scale. If treated collectively, though, these evaluation strategies and metrics make a growing contribution to the evaluation toolkit available to creativity researchers. This Section explores the options available from existing research.

Wiggins has proposed a framework for categorising creative systems (Wiggins, 2006a, 2003, 2001) inspired by Boden's proposals on creativity (Boden, 2004). The framework describes system details formally according to seven formal rule sets and functions relating to the system's conceptual space (i.e. the set of all possible items that could conceivably be output by the system). One of these rulesets, $\mathcal{E}$, is the set of rules used to evaluate items in the conceptual space. Though the framework is intended to be used to 'analyse, evaluate and compare creative systems' (Wiggins, 2003, p. 1), Wiggins carefully states that he does not contribute to the debate on creative evaluation:

> 'I am making no attempt here to discuss or assess the value of any concepts discovered: while this issue is clearly fundamentally important [citing Boden (1998), Ritchie (2001), Pearce and Wiggins (2001)], it can safely be left for another time.' (Wiggins, 2006a, p. 453)

Wiggins's work has tended to focus on quality evaluation rather than creativity evaluation (Whorley, Wiggins, & Pearce, 2007; Pearce & Wiggins, 2007; Whorley, Wiggins, Rhodes, & Pearce, 2010). In particular Pearce and Wiggins (2007)'s melody generation system was evaluated using a variation of Amabile's Consensual Assessment Technique (CAT) (Amabile, 1996), intended by Amabile for evaluating the creativity demonstrated by humans in a quantitative way by expert judges. CAT was adapted slightly in Pearce and Wiggins (2007) to assess the quality of output ('stylistic success') from their system rather than the creativity of the system itself. This decision was perhaps influenced by the authors' substantial background in musical quality evaluation (e.g. Wiggins, Miranda, Smaill, & Harris, 1993; Pearce & Wiggins, 2001; Wiggins, 2001, 2003; Pearce, 2005; Wiggins, 2006a).

A statistical approach to creativity evaluation is taken in O'Donoghue (2007), demonstrated on a system that generates analogies. Initial tests for novelty were first carried out, comparing generated analogies against the system's knowledge base. Following this, ratings supplied by human evaluators were used to assess the produced analogies on 'qualities associated with creativity' (O'Donoghue, 2007, p. 34). These qualities include novelty and quality (attributed to Ritchie (2001) by O'Donoghue), plus criteria relating to Colton et al.'s 'interestingness': empirical plausibility, novelty (again), surprisingness, applicability, utility, comprehensibility and complexity of concepts and conjectures (Colton et al., 2000; Colton & Steel, 1999). The evaluators also provided ratings on the validity of the analogies. Statistical tests were performed on this evaluation data (first McNemar's test, then the more powerful Mann-Whitney (Wilcoxon) test) to test the hypothesis that the mean and median creativity ratings for 'valid' analogies were greater than the equivalent ratings for 'invalid' analogies. Essentially, although critical of the 'artefact-centric' approach in Ritchie (2001) (O'Donoghue, 2007, p.

[27] As the Section 2.3 survey of evaluative practice will show.

32), O'Donoghue takes a similar approach, obtaining various ratings of the produced artefacts based on their novelty, typicality and value, before manipulating these ratings (using statistical tests, not formal criteria) for an overall creativity evaluation of the system.

Another perspective on evaluation as part of the creative process is provided by Bird and Stokes (2007). Working in a domain of pattern recognition and appreciation by robots, Bird and Stokes's approach to evaluation described properties of an artefact and then assigned that property a value of merit or 'de-merit'. Bird and Stokes (2007) do not give details of how this property value is assigned, keeping this part of the discussion in Bird and Stokes (2007) brief, but they do stress that evaluation is a necessary part of the robots' creative process.

From a case-based reasoning perspective, Byrne, Schnier, and Hendley (2008) offer a theoretical presentation of a creativity metric according to the creative potential stored in a system's repository of information. Case-based reasoning involves extrapolating from example 'cases' to generate solutions to a related problem. Byrne et al. (2008) suggest measuring the similarity between the problem solution proposed and the case base used to generate that solution. The greater the dissimilarity, the greater the 'misuse of knowledge' (Byrne et al., 2008, p. 38), which Byrne et al. (2008) suggest is strongly linked to creativity. Byrne et al. propose that the amount of processing required to transform the example case to the solution could be measured, perhaps using graph distance metrics, although they do not apply their suggestions for any practical evaluations.

Collins, Laney, Willis, and Garthwaite (2010) employ the Wilcoxon's two-sample statistical test on their music harmonisation generator. Again the metric examines similarity between generated output and the system's knowledge base, however their purpose is the reverse of of Byrne et al. (2008). Despite Collins et al. describing their test as a creativity metric, they actually measure how closely the system can replicate the test set (similar to the approach of Whorley et al. (2007)). In other words, Collins et al. (2010) actually propose a correctness metric rather than a creativity metric, a distinction which they briefly acknowledge:

> 'This paper has presented a metric for evaluating the creativity of a music-generating system. Until further evaluation has been conducted (by human listeners rather than just by the creativity metric), we are cautious about labelling our overall system as creative.' (Collins et al., 2010, p. 9)

The blurring of evaluative aims, between assessing quality and creativity, is a theme that recurs not only in Collins et al. (2010), but also in several computational creativity system evaluations.[28]

2.2 Scientific method and its relevance to computational creativity evaluation

It is useful to take a step back from specific computational creativity methodology and consider more broadly what makes good methodological practice for evaluation of research. There is a large body

[28] This shall be shown in Section 2.3 and Chapter 5 Section 5.1.8.

of work on *scientific method* that is relevant for such considerations. As shall be described here, this body of research has not arrived at a single universally-agreed scientific method, suitable for all such research, despite much concerted effort from scholars over centuries. Despite this, however, it is useful to review the contributions that have been made to the development of scientific method. Such a review allows us to learn from these contributions and the associated debates and to see what general methodological principles have arisen that could be applied for computational creativity evaluation.[29]

Scientific method can be thought of as a standardised practice for scientists to follow in their pursuit of accurate, simple and comprehensive theories about the world. The scientific method as we know it today can (arguably) be said to date from the work of Bacon (1878, originally 1620) (Thagard, 1988). As a scientist who worked in a practical, applied manner, based around conducting experiments, Bacon argued that scientific method should involve starting from the data you have through observations and experimental evidence and trying to explain this data.

2.2.1 Verification of hypotheses

In scientific method, a hypothesis or theory is a proposition or assertion (or set of related propositions/assertions) that acts as the context for scientific investigation. It is debatable whether the scientist starts with a hypothesis and uses it to make predictions to be tested, or if the hypothesis is formulated post-discovery of the facts (Hacking, 1981a). As Hacking (1981a) concluded in his examination of Lakatos' work,[30] timing and ordering becomes less important over time, once a hypothesis has been adopted as part of our body of scientific knowledge. The end result of scientific method is aimed at making contributions to knowledge but there is a distinction between discovery of such knowledge and its justification,[31] the latter of which scientific method is concerned with.[32] In the case of this thesis' aims, scientific method concerns justifying how creative a computational system is, given some hypothesis or theory of what it means for that system to be creative.

In one model of scientific method using hypotheses, if one forms a hypothesis based on prior ideas and speculations (and perhaps some initial evidence), and can then deduce from that hypothesis a particular conclusion(s), then the hypothesis is confirmed by empirical evidence confirming that conclusion(s). This is referred to as the *hypothetico-deductive* model, where:

> 'Start with hypothesis H. Use logic to deduce predicted observation O. If O is observed, then H is confirmed... but if not-O is observed H is falsified' (Thagard, 1988, p. 192)

Induction is a method based on similar principles as hypothetico-deductivism, but proceeding in an opposite way. With induction, if one can collect enough examples of evidence indicating a

[29] To clarify, this thesis makes no claims that computational creativity is solely a *scientific* endeavour (see Section 1.5).

[30] The contributions of Lakatos to scientific method will be reviewed later in this Section.

[31] A similar distinction between discovery and justification/verification has been made when examining different stages of the creative process, for example by Poincaré (1929); see Chapter 3 Section 3.4.2.

[32] As we shall see later in this Section, Feyerabend (1993) disagrees with this viewpoint on the focus of scientific method.

particular pattern, then this pattern can be formulated into a hypothesis supported by the evidence of those examples. For computational creativity evaluation, induction would be the process of building a theory about how a system is (or systems are) creative, given several existing examples of that type of system being creative.[33] If, on the other hand, a hypothesis about the creativity of a system(s) is largely formed from previous speculations and then tested successfully with confirmatory evidence from examples of creative systems, this would fit the hypothetico-deductive model.

A central assumption underlies both induction and deduction: a hypothesis can be *verified* by the existence of confirming examples. This assumption is affected by two counter-arguments: the paradox by which a hypothesis can be confirmed by seemingly irrelevant evidence, and the problems encountered when falsifying evidence is discovered during the search for empirical evidence. These counter-arguments were offered by Hempel and Popper (following Hume) respectively (Bird, 1998).

The first counter-argument is centred around the *paradox of confirmation* (Bird, 1998, p. 92, originally proposed by Hempel)[34] where a hypothesis h is verified not only by the existence of evidence e in the scenario covered by h but by the existence of $not(e)$ in the complementary of that scenario. Hempel's example considered a birdwatcher who has a theory that all ravens are black. If you saw only black ravens, then this is evidence for the theory. (A white or green raven would falsify the theory). Hempel saw that logically, the 'all ravens are black' statement is equivalent to 'all non-black things are non-ravens' (as described by Edmonds & Eidinow, 2001, p. 129). Hence, to see a red robin - or a yellow sun - is to provide confirmatory evidence for the 'all ravens are black' theory; herein lies the paradox. This is troublesome as it seems to contradict common sense; why should a yellow sun be relevant evidence towards verification of a statement about black ravens?[35] To apply this to computational creativity evaluation, a chair, deemed as uncreative, could be offered as evidence for a theory that 'all computer systems of type X are creative'.

The second counter-argument, proposed by Popper, led to the proposal of a new method to adopt in scientific practice: falsification, which shall be examined next in this review of scientific method. Popper argued that the process of verification associated with induction (and hypothetico-deductivism) was not a sound scientific method to work with, as a string of positive results showing a theory to be seemingly valid can be contradicted by just one result showing the theory to be false:[36]

> 'The man who has fed the chicken every day throughout its life at last wrings its neck instead,

[33] Given the emphasis on novelty in creativity (Chapter 3 Section 3.4.1, the formation of a theory from previous examples would need to be carefully handled; direct copies of existing examples of systems deemed as creative would be unlikely to themselves be considered creative unless, perhaps, the likenesses were purely coincidental and unintentional.

[34] This is also known as the raven paradox or as Hempel's paradox, due to the philosopher Carl Hempel using the hypothesis 'all ravens are black' to illustrate this paradox originally (Bird, 1998).

[35] In Bird (1998), the example is of the hypothesis 'all bats are blind'. According to the paradox of confirmation, this hypothesis can be verified both by the existence of a blind bat and of a non-blind non-bat (in Bird's example, a 'sharp-eyed red-setter' (Bird, 1998, p. 92)), although the latter is seemingly irrelevant to the original hypothesis.

[36] Popper was preceded in his objections by Hume, who questioned whether inductive reasoning was a reasonable scientific method and questioned the assumption that past evidence is an adequate basis for future predictions (Bird, 1998).

> showing that more refined views as to the uniformity of nature would have been useful to the chicken.' (Russell, 1912, p. 98)

2.2.2 Falsificationism

Popper (1972) pointed debates on scientific method into a different direction from induction, seeing observation and experiments as *theory-laden* rather than free of context.[37] With induction, it can be questioned how valid so-called universal statements are, where the construction and verification of those statements is based upon contextual experiences that are open to interpretation (observations, experimental results, and so on). Therefore Popper suggested falsification as an epistemologically more secure scientific method: constructing a hypothesis and seeks the *falsifying* example to disprove it. A scientific hypothesis therefore cannot be proven beyond doubt, but it can be disproven. Popper argued that scientists' focus should be on the search for falsifiers, rather than the search for supporting evidence.Positive results from testing (even rigorous, detailed and varied testing) can only be temporarily used as supporting evidence to *corroborate* a theory. On the other hand, a negative result 'may always overthrow it [the theory]' (Popper, 1972, p. 33).

Predictions are deduced from a theory in order to test it; such predictions are prioritised if they are more amenable to testing, if they are not deducible from current theoretical knowledge or if they actually lead to conclusions which contradict this current knowledge. The predictions are compared against experimental evidence gained from reproducing the predicted scenario and making observations.[38] Popper did not require that these statements must be tested in all possible ways but, for Popper, if a statement is untestable then it cannot be deemed scientific (Popper, 1972, p. 48). Statements in science must be testable. Scientific statements also need to be objective [p. 44-48], hence corroboratory empirically evidence must be reachable and reproducible by others.[39]

Popper felt that the need for scientific method was driven by scientific fallibility and the need to be able to work within such fallibility, to avoid committing mistakes to scientific knowledge where possible (Popper, 1972, pp. 49-50). In fact Popper advocated a more tentative and temporary adoption of scientific knowledge. Though we can hold a subjective opinion that we are sure of a statement being true, we cannot be objectively certain of this. 'The demand for scientific objectivity makes it

[37] Bird (1998) gave as example the statement ' "I see a lemon" '(Bird, 1998, p. 166). To label what is being observed as being a lemon does however imply various other facts, such as the fact that the observed item comes from a lemon tree.

[38] As Bird (1998, p. 132) has pointed out, observation involves the observer actively interpreting sensory information; it is possible for two observers to interpret the same sensory information in two different ways. Hence potential subjectivity is introduced into the gathering of empirical evidence. Bird (1998) sees a way to avoid this: 'The observer must be able reliably to recognize the facts being observed. If the observer is not reliable then neither will his reports be reliable.' (Bird, 1998, p. 132). Reliability of the instruments and processes involved would also be of concern.

[39] In Chapter 5 Section 5.5 it shall be discussed how this emphasis on *scientific* knowledge distances scientific method somewhat from computational creativity evaluation. For example, the dynamic nature of creativity and its existence as a continuous rather than discrete quality may jar somewhat with the traditional interpretation of a hypothesis or theory in scientific method that is (a) treated as static and (b) often exemplified in statements which are assigned a discrete truth value which would not be affected by changes in time and context.

inevitable that every scientific statement must remain *tentative for ever*' (Popper, 1972, p. 280).[40]

2.2.3 Structure and growth of scientific knowledge

How did Popper envision that science should progress, if nothing can be proven? (And, following on from this, how can we prove how creative our computational systems are?) Also, as Bird (1998) has questioned, how does Popper's falsification deal with a lack of appropriate empirical evidence, corroborative or falsifying? This question is pertinent to computational creativity researchers; elements of unpredictability in creativity mean that the evidence a researcher is searching for in evaluation may not be retrievable on demand; in fact the predictability of a particular piece of evidence arising (for example an expected output instance) may be detrimental to the system's overall perceived creativity, if it produces only expected evidence. Bird (1998, pp. 8-9) gives the example of Darwinian theory which is corroborated by finding fossils, but which is not falsified should fossils for a certain period not be found; these gaps in the fossil records can be argued to be due to fossils not being found yet, rather than those fossils not existing. Nothing can be proven beyond possible doubt, according to Popper, but if a hypothesis has been corroborated by thorough and diverse testing, it should not be succeeded by a new theory without 'good reason'(Popper, 1972, pp. 53-54), such as replacing it with another version of the theory which is more testable, or if the consequences of the hypothesis are falsified by later empirical evidence. This explanation does not sufficiently remove Bird's issue with lack of evidence, but has given some basis for progression in scientific work. The arising of this issue acts as a warning for computational creativity evaluation; where possible in our evaluations we should seek types of empirical evidence that accommodate the *expected unexpectedness* of creativity.

Looking at scientific progression more generally, Popper saw the trend for progression along the 'path of science' (Popper, 1972, p. 276-281) as the development of more universally applicable theories, by continually proposing (and testing) theories at a higher level of universality than current theories[41] (while not pursuing theories which are so general that they could easily be falsified by outliers or extreme scenarios, or become tautologies (and therefore untestable). The ultimate Popperian aim is to discover deeper and more general problems and corresponding theories and to conduct more detailed, rigorous, diverse testing of hypotheses. This is an attractive proposition for the development of theories on what creativity entails in the computational systems being evaluated. Overly-specific theories on what makes a system more or less creative may end up overfitting to specific instances

[40] Popper no doubt intended this 'tentative' adoption of knowledge to be related to the positive and/or negative evidence that had been discovered at a particular point in time, rather than to changing definitions. Nevertheless, as Chapter 3 will show, attempts to pin down the dynamic nature of creativity into hypotheses fit well with the concept of tentative, temporary adoption of knowledge (scientific or otherwise), rather than the adoption of absolute and permanent truths.

[41] Bird (1998) has pointed out (Bird, 1998, pp. 179-180) that Popper has in fact allowed induction to be part of his scientific method: if enough positive evidence is rigorously gathered for a theory, it should be given more credence than a theory with less thorough corroboratory evidence. Popper (1972, p. 276) acknowledged that the 'path of science', as he sees it, could be described as 'quasi-inductive'.

of creativity in a particular context and time, rather than accommodating the dynamic[42] nature of creativity. They are also less likely to be useful to the research field as a whole, or to subsections of the research field who are working in the same or similar domains.

Popper did acknowledge that in practice, one may try to discount falsifying evidence due to experimental error or unreliability, or treating discrepancies between evidence and theory as anomalies that will be resolved as we advance our knowledge of the theory further (Popper, 1972). Bird (1998) described a potential situation for when this last clause may apply; in the 19th century, Lord Kelvin calculated that that the solar system could not have existed for a long enough time for life to have evolved as evolutionary theory predicts, based on cooling rates of the Sun and Earth and assumptions (based on scientific knowledge at the time) that there are no heat sources on the earth and that the Sun's energy is generated solely from combustion. Creationists see this as evidence for falsification of evolutionary theory, especially as creationists believe that the world is only a few millennia old. In his calculations, however, Kelvin did not (and could not) account for sources of energy which had not yet been discovered, such as radioactive decay in the Earth or alternative energy generation processes in the Sun such as fusion. Taking these discoveries into account, Bird (1998) pointed out how Kelvin's calculations no longer provide falsifying evidence for the theory of evolution.

A question is raised here on how falsification would deal with scenarios where falsifying evidence becomes invalid after later scientific discoveries, or where a single and possibly inaccurate falsifying exemplar stands alone amongst a large body of corroboratory evidence.[43] This concern is very relevant for computational creativity evaluation, where a universally valid hypothesis on the nature of computational creativity may be somewhat unrealistic.[44] Presumably, continuous testing (as advocated by Popper) would detect where evidence was no longer adequate for falsification. It appears from Popper's accounts in Popper (1972), however, that scientists should concentrate their time and resources on seeking falsifying evidence for those theories which have not yet been falsified, perhaps at the expense of continuing to rigorously and routinely test those theories which have been falsified.

The suggestions in Lakatos (1978) deal better with scenarios such as Kelvin's supposed falsification of the earth's lifespan. Lakatos took Popper's scientific method and adapted it towards a 'methodology of scientific research programmes' (Lakatos, 1978, title and throughout text). For Lakatos, occasional negative evidence for a theory - 'anomalies'(Lakatos, 1978, pp. 4-6) - should not necessarily lead to that theory being falsified and therefore rejected. Along these lines, it should also be acceptable to introduce occasional 'ad hoc' explanatory hypotheses to support the theory (Lakatos,

[42]The dynamism of creativity as a general concept is addressed in Chapter 3.

[43]The latter scenario is particularly pertinent in statistical hypothesis testing, widely used in psychology research and other disciplines, where a hypothesis is tested and success is measured against p values. A hypothesis may be true at least 95% of the time (p=0.05) or 99% (p=0.01) and so on. As Bird (1998) has pointed out, a hypothesis which is satisfied by 99 trials out of 100 would presumably be regarded as falsified by Popper, but on the other hand would be considered to be statistically very significant, if a scientist is concerned with p values.

[44]See Chapter 3.

1978, p. 33). This approach could better accommodate theories which did not universally cover all scenarios and contexts, as may be expected in computational creativity.

Lakatos (1978) considered how to avoid situations where theories are accepted even if they generate an unacceptable amount of anomalies and rely upon too many ad hoc hypotheses of unreliable nature. Lakatos differentiated between scientific research programmes which can handle anomalies, as opposed to those that have no strategies or heuristics for dealing with anomalies as they arise. In a scientific research programme, theory consists of a *hard core* which is unchangeable and which defines the programme, plus an *auxiliary belt* which can be changed and adapted as necessary (Lakatos, 1978, pp. 48-49). A scientific research programme may deal with anomalies either using negative heuristics (a strategy of not abandoning the hard core of the theory when anomalies occur), or positive heuristics (advising what to do when anomalies occur).

Research programmes should be seen as 'progressive' (Lakatos, 1978, pp. 33-34) if later theories are more informative than earlier ones, later theories can explain earlier theories and later theories have more corroboratory evidence than earlier ones. Otherwise the research programmes would be described as 'degenerating' (Lakatos, 1978, pp. 33-34). Computational creativity evaluation would be interested in pursuing progressive rather than degenerating research programmes, where theories could develop and become more informative, generally useful and more accommodating of corroboratory evidence over time.

Like Lakatos, Kuhn (1962) considered scientific progress more widely than for individual theories. For Kuhn, science is not the linear accumulation of theories that move closer and closer to true theories about the world. Instead science goes through cycles of *normal science*, *crisis* and *scientific revolutions*. Scientists work within a particular *paradigm* ('a conceptual scheme representing a group's shared commitments and providing them with a way of looking at phenomena' (Thagard, 1988, p. 36)). Different paradigms are favoured at different points in time, given the state of knowledge at that time and the cyclical point that scientific research related to that paradigm has reached.

Normal science represents the stage of scientific research that had traditionally been represented in scientific method till that point, defined by Kuhn as:

> 'research firmly based upon one or more past scientific achievements, achievements that some particular scientific community acknowledges *for a time* as supplying the foundation for its further practice.' (Kuhn, 1962, p. 10, my emphasis added)

The emphasis in the quote above highlights a fundamental part of Kuhn's contribution, which precedes Lakatos in considering how scientific research communities move from one dominant set of theories to explain certain phenomena (or one dominant paradigm, in Kuhnian terminology) to another.[45] Over time, Kuhn observed, the current *status quo* of scientific knowledge in a particular domain would encounter anomalies and problems, for which theories would be extended and/or re-

[45] Kuhn (1962) refers to this as a *paradigm shift*.

shaped in order to accommodate these issues within the current theoretical framework. What Kuhn referred to as 'crisis' (throughout Kuhn, 1962) is to reach a stage where the existing paradigm (theoretical framework) is being stretched and adapted in numerous ways, such that (to use Lakatos' terminology) the 'hard core' of the theory is being challenged by anomalies beyond that point which the 'auxiliary belt' can accommodate those anomalies. This would usually not be a discrete point in time, but would be a process of recognition over time by increasing numbers of scientists (but perhaps not the whole of that scientific community). Some of these scientists would identify an alternative paradigm (either prior to the crisis point or in response to the recognition of crisis) and conduct revolutionary science, i.e. working within this alternative paradigm, rather than working within the existing paradigm (normal science). Gradually a scientific revolution would occur, where the alternative paradigm replaces the previous paradigm, as more and more of the community adapt their work to fit in with the new paradigmatic structure and theories.

For example, Einstein's relativity theory disrupted the cumulative extension of theory from Newton. Newtonian mechanics could no longer be seen as the full and complete scientific truth, as Einstein's theory disproved various aspects of Newtonian mechanics. From a Kuhnian viewpoint, this was an example where the normal science period of Newtonian mechanics was disrupted during the period where Einstein conducted revolutionary science, which established a new scientific paradigm.

The Kuhnian view of scientific development can also be illustrated through the paradigm of computational creativity, an area where the gathering and structuring of knowledge is in relatively early stages given its comparative youth as an area of research (Chapter 1 Section 1.5). One can see the scenario where one model on what computational creativity entails is favoured at first, with many researchers in the community evaluating their systems according to that model, until a growing body of evidence encourages a paradigm shift towards using an alternative model of creativity.[46]

New paradigms are not necessarily better than the preceding paradigm, but would replace the older paradigm. Scientific method would be used in the 'normal science' phase, to assist progress within a paradigm, but would not contribute to the larger-scale progress Kuhn identifies, from one paradigm to another. What Kuhn's contribution allowed, as did many of those outlined above, is that it became acceptable for science to be seen as fallible. The pertinent issue therefore became how to work within scientific fallibility and reconcile this fallibility against the perception of scientific knowledge being the truth. Scientific method can therefore be seen as the tool to help scientists work within the theoretical framework that they have chosen to adopt at that time.

[46]For example, it could be argued that in research on human creativity, a paradigm shift has occurred, moving from treating creativity as divergent thinking, to using a more holistic multi-perspective or multi-stage model of creativity (Chapter 3 Section 3.4.2). It may also be tentatively suggested from the evidence from the survey of current practice later in this Chapter (Section 2.3) that a paradigm shift occurred from the Ritchie (2007) criteria model of creativity to the creative tripod model proposed by Colton (2008b), although the quantity of evidence is far too small to make more than a suggestion of this at this point. A further paradigm shift towards wide-scale adoption of Computational Creativity Theory (Colton et al., 2011) could possibly occur, but it is too early at this point in time to make such a prediction with any concrete evidence.

2.2.4 The existence (or not) of a standard scientific method

One question remains throughout the above discussion: is there such a thing as '*the* scientific method'? (Bird, 1998, p. 237) Bird has argued that there is 'there is no such thing' (Bird, 1998, p. 273); instead there is a 'spectrum of methods and principles' (Bird, 1998, p. 258) (and rules of thumb, heuristics, and so on (Bird, 1998, p. 259)). In fact, the idea of working to a prescribed scientific method has been rejected by Feyerabend (1993), who argued that the adoption of any prescribed method, principle, rule or theory can lead to situations being encountered where exceptions need to be made. On this reckoning, all methods and hypotheses should be considered as potentially useful in scientific practice, with none ruled out *per se*. It may be that the greatest knowledge gains may be obtained by proceeding '*counterinductively*' (Feyerabend, 1993, p. 29), contradicting established theories.

Feyerabend advocated what he describes as 'an anarchistic methodology and a corresponding anarchistic science' (Feyerabend, 1993, p. 21). This 'anarchism' in choosing and applying methods allows the researcher to actively avoid situations where progress in research is inhibited by (perhaps inappropriate) methodological rules and principles. In this way, Feyerabend argued, scientific progress would not be limited by older but well-established theories, which may persist despite being poorer than newer but less-established contradictory theories.

There is one exception to this general approach - the single principle 'that can be defended under *all* circumstances and in *all* stages of human development.' (Feyerabend, 1993, p. 28):[47]

> '*The only principle that does not inhibit progress is:* anything goes' (Feyerabend, 1993, p. 23)

Adopting this approach 'against method' (Feyerabend, 1993, title), Feyerabend rejected the adoption of a single scientific method.[48] For our purposes, although some practical guidance in evaluation methodology would be useful, to inform computational creativity researchers of available methods and develop practice across the research community, Feyerabend warns us of the negative aspects of being overly prescriptive in such methodological recommendations.

2.2.5 Scientific method review: general conclusions

From the above considerations, we can clearly see that a single, standardised and universally-agreed-upon scientific method does not yet exist. Several debates exist over various relevant details, such as how and when a hypothesis is formed, or the choice of which of various individual methods should be cast as 'the' scientific method, for optimum scientific progress and advancement of knowledge, or whether there actually is a single appropriate scientific method for all scientific disciplines. These ongoing debates have been introduced above; for more detailed coverage of scientific method, the

[47] It seems appropriate for there to be exceptions to an approach that actively encourages exploiting exceptional cases.

[48] Although critical of methods such as induction and falsification, Feyerabend is sympathetic to the ideas in Lakatos (1978), as he sees them as 'anarchism in disguise' (Feyerabend, 1993, p. 181). In the Author's preface, Feyerabend describes the work in (Feyerabend, 1993) as a response to Imre Lakatos, to whom the work is dedicated.

interested reader is referred to: Bird (1998), particularly Chapters 8 and the discussion of Popper's work in Chapter 5; the collected papers in Hacking (1981b); and the texts referenced during the above discussions in this Section (especially Popper, 1972; Lakatos, 1978; Feyerabend, 1993).In the scope of this thesis, what can be learned from the discussions in this Section is the general approach of scientific method that has been established over centuries of debate. Good practice here should involve clearly identifying a relevant hypothesis[49] and empirical consequences of that hypothesis. The hypothesis should be tested by gathering appropriate evidence to support (or falsify, if taking a Popperian stance) empirical consequences of that hypothesis. Such an approach should be borne in mind (alongside the comments above that relate scientific method to the aims of this thesis) when considering appropriate methodological steps to follow for computational creativity evaluation. [50]

2.3 Survey of current practice in computational creativity evaluation

Having identified that there are a number of contributions offered for creativity evaluation, it is useful to see what methods are being used in practice, by examining evaluations reported in recent relevant publications. The purpose of this examination is to survey to what extent evaluation (and specifically evaluation of creativity) is undertaken in computational creativity research. The survey looks at whether evaluation is carried out systematically, rigorously and as an expected and normal part of the research process in this field. The survey will look at how systems are evaluated, by whom, and to what extent. Evidence of methods that have been adopted as standard across the community will also be collected. Considering the progress in the research field at a level transcending individual systems and individual researchers, it is also useful to see how systems are compared against the current state of the art in that research strand, by looking at creative systems within the context of comparison to similar related systems. This allows a system's researchers to clarify the novel achievements and contributions to knowledge of an individual system, as well as identifying relative weaknesses of the system which may be addressed by learning from other research.

2.3.1 Survey methodology

A literature search was carried out to find all journal papers that present details of a computational creativity system. Using the *Web of Knowledge* and *Scopus* databases, various combinations of words and phrases such as 'computational creativity', 'creative system', 'creative computation', 'system' and 'creativity' were used as search terms. The search explicitly focused on finding all reports of computational creativity systems where the system was intended to be creative.

[49] Or several related hypotheses.

[50] Section 2.3.3 considers how current practice in computational creativity evaluation reflects scientific method principles. Chapter 5 Section 5.5 considers how the SPECS methodology (derived in this thesis) relates to scientific method and how it contextualises the approach of scientific method in the frame of computational creativity research evaluation.

Table 2.2: Sources for papers included in survey on evaluation of computational creativity systems.

Publication source	Year	Total no. of papers surveyed	No. of papers presenting creative systems
Minds and Machines 20(4) Special Issue on Computational Creativity	2010	8	3
AI Magazine 30(3) Special Issue on Computational Creativity	2009	6	2
New Generation Computing 24(3) Special Issue on Computational Creativity	2006	6	4
Knowledge-Based Systems 19(7) Special Issue: *Creative Systems*	2006	4	3
Other journal papers retrieved in a literature search	1996, 2002, 2010	4	2
1st International Computational Creativity Conference (ICCCX)	2010	33	25
Dagstuhl Seminar on Computational Creativity	2009	36	10
International Joint Workshop for Computational Creativity (IJWCC)	2008	18	12
International Joint Workshop for Computational Creativity (IJWCC)	2007	17	14
Totals		132	75

The resulting collection of papers was supplemented with papers from journal special issues on computational creativity (if these papers had not already been retrieved in the literature searches). Reflecting the current balance of conference/workshop publications to journal publications in computational creativity, papers from recent Computational Creativity research events were also added to the survey.[51] Details of these sources are listed in Table 2.2 and in full in Appendix A.

For each paper, the following information was recorded (as illustrated in Figure 2.4):

1. Paper details (Paper title, Authors, Publication year, Publication venue, No. of pages).
2. Creative system details (Whether or not a computational creativity system is presented in the paper, Creative domain of the system (e.g. music, visual, reasoning etc.), Name of the system being presented, if any).
3. Evaluation details:
    - Is system evaluation mentioned at all in the paper?
    - Has a system evaluation been performed and described in the paper?
    - Brief description of evaluation done.
    - Does the paper contain a section on evaluation?[52]
    - Do the authors state the aims of their evaluation and/or their evaluative criteria?
    - Is the main aim of evaluation to assess creativity (including quality of output/system) or (just) quality of output/system?
    - Is a standard creativity evaluation method used? (e.g. Colton, 2008b; Ritchie, 2007)
    - Is the system compared to other systems by different authors?

[51] Proceedings from events in 2007 onwards were included in the survey. Proceedings from creativity research events prior to 2007 are not readily available in an online format, making them difficult to locate for this survey and also less likely to have influence on researchers today unless they were one of the relatively few people who attended that workshop (in comparison with attendances of such events in more recent years).

[52] which may or may not have 'Evaluation' in the title of the section.

Table 2.3: Creative domains addressed by creative systems in the paper.

Domain	No. of papers	Domain	No. of papers
linguistic creativity	23	gameplay	2
music	19	music and linguistic creativity	2
artistic creativity/visual images	17	reasoning	2
creativity support tools	3	furniture arrangement	1
design	2	movies	1
Ecosystem simulation	2	scientific creativity	1

- Have independent evaluators been used? (i.e. not the researchers themselves.)
- If independent evaluators are used, are experts or novices (or both) used as evaluators?
- What factors/criteria are used to evaluate the system?

4. Discussion of the system's creativity.
   - Degree of discussion of the system's creativity in the paper.

     [none, very little, some, a lot, this is the main focus of the paper]
   - Number of pages in the paper that discuss the system's creativity.
   - Percentage of pages in the paper that discuss the system's creativity.

2.3.2 Survey findings

Of the 132 retrieved papers, 75 papers presented details of computational creativity systems, with the remaining papers discussing theoretical aspects of creativity or other related issues.

The domains which are most tackled by creative systems in the survey are language (23), music (19) and art/images (17). Table 2.3 lists the creative domains represented in the survey.

Survey results are summarised in Table 2.4. Looking in more detail at the 75 surveyed systems:

- 58 out of 75 papers mentioned system evaluation in some way; 17 did not. Of those 58 papers:
  - 41 papers reported details of evaluation that had been done. 17 papers discussed potential evaluation strategies but had not actually performed any evaluation.
  - 38 papers contained a section(s) devoted to evaluation. 20 did not. Of these 38:
    - 18 papers had a section with the word *Evaluation* in the section title.
    - 20 papers had a section on evaluation which did not have *Evaluation* in its title.
  - Nearly all of the papers that mentioned evaluation (52 out of 58) stated or at least mentioned what assessment standards the system was evaluated by. 6 papers left this unspecified, merely stating that the system was successful without justifying why.

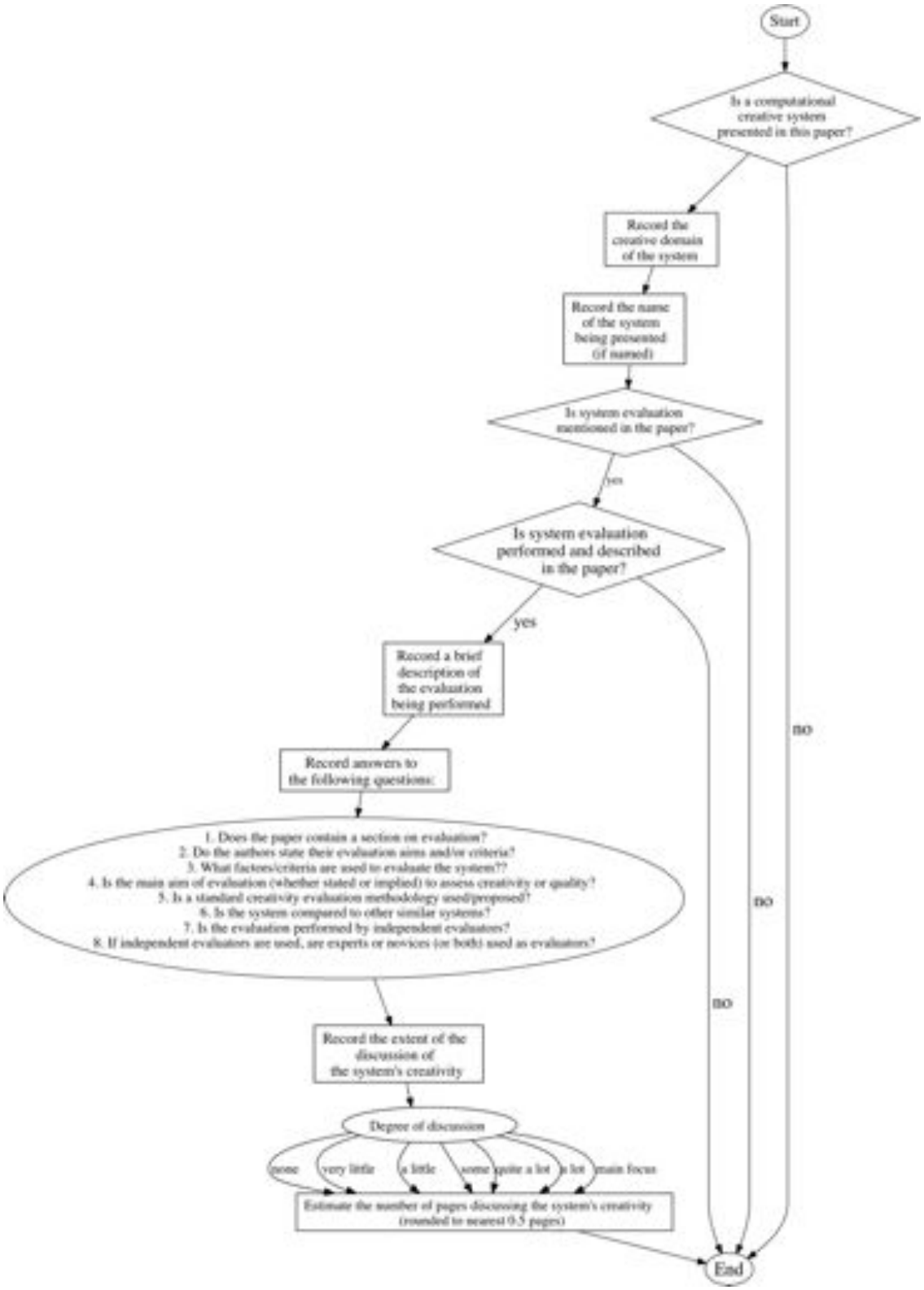


Figure 2.4: Evaluation survey process represented diagrammatically.

- Of the 75 programs presented as creative systems, only a third (26 systems) were critically discussed in terms of how creative they were.
  - 32 systems were evaluated based solely on the quality or accuracy of system performance compared to a human performing that task. This set of 32 systems includes 3 systems which were described as being assessed on how creative the systems were, but which were actually assessed by the quality of the system's performance.
  - 1 paper evaluated its system in terms of knowledge gained for future research.
  - The remaining 17 papers had no critical discussion or evaluation of the creative system.
- 20 papers referred to existing creativity evaluation methodologies to evaluate their system:
  - 5 papers used Colton (2008b).
    - 1 further paper referenced Colton (2008b) but did not use this methodology.
  - 4 papers used Ritchie (2007, 2001)
    - 1 further paper referenced Ritchie (2001) but did not use this methodology.
  - 1 paper used the Consensual Assessment Technique (Amabile, 1996).
  - 3 papers used creativity theories (Boden, 1990; Wiggins, 2003) to classify their system.[53]
  - 5 papers proposed new metrics (Bird & Stokes, 2007; O'Donoghue, 2007; Whorley et al., 2007; Collins et al., 2010; Nelson, Brummel, Grove, Jorgenson, Sen, & Gamble, 2010).
  - Of the 18 papers that applied recognised creativity evaluation methodologies:
    - 10 papers used the methodologies to measure how creative their systems were.
    - 6 papers adapted the chosen methodology to measure the quality of the systems.
    - 2 papers used 'creativity' methodologies that actually measured quality.
  - In half of the papers reporting details of a performed evaluation (38 out of 75), instead of using a creativity metric, either evaluation was performed using non-standard, system-specific methods not specifically designed for creativity evaluation or the evaluation methodology was left unstated.
- 11 systems were directly contrasted to existing similar systems or to human performance:
  - 8 systems were compared to systems by authors other than the paper authors.
  - 2 systems were compared to systems by that author(s) but not those of other authors.
  - 1 system was evaluated using human performance on the same task as a control in the evaluation experiment.
- A further 4 papers discussed related systems without performing evaluative comparisons.

[53] Creativity theories were not used specifically to evaluate how creative the system was, but to justify the system as being creative by positioning it in the context of creativity theory.

- Only a third of systems (25 / 75) were evaluated by people other than the paper authors:
  - 8 systems were evaluated by domain experts or the target users.
  - 4 systems were evaluated by novices in that domain.
  - 4 systems were evaluated by a range of people with differing expertise in that domain.
  - The expertise of evaluators in the remaining 9 systems was left unstated.

Journal papers generally tend to undergo a more stringent review process and are written and edited over a longer time frame than conference papers. One could hypothesise that if the survey was focused only to journal papers, ignoring conference papers, then a more rigorous and scientific approach to evaluation would be seen. To a certain extent this hypothesis is validated in this survey, although significant problems still remain with the evaluations performed. Taking the journal papers that present a computational creativity system (14 out of 75):

- Out of 14 journal papers that presented details of a computational creativity system, system evaluation was mentioned in almost all papers (12 out of 14) and evaluation was performed and reported for all but one of these 12 papers.
- The evaluation process was more transparent in the set of journal papers. 11 papers clearly stated the evaluation criteria being used for evaluation and 10 papers devoted a section of the paper to reporting methodological details and results of the system's evaluation.
- Again, though, although each system was presented as being creative, not all papers critically discussed their system's creativity. Only 7 out of 14 systems were evaluated for creativity, though at 50% this is an increase of 15% compared to the full set of 75 papers included in the survey. The other systems were evaluated only for accuracy and quality of performance.
- A standard creativity evaluation methodology was mentioned in 2 of the 14 papers. Of these 2 papers, 1 paper informally used a methodology (Ritchie, 2001) as a prompt for discussion of their system's creativity. 1 paper mentioned Ritchie's approach but did not apply it.
- Only 5 systems were compared against the performance of other similar systems. Of these 5 systems, 1 paper compared its system to existing systems in the same domain but used different criteria for evaluation. 2 of these 5 papers compared the system to its research competition in terms of quality of performance but not in terms of creativity exhibited by the system.

Figure 2.5 presents a quantitative representation of the papers' creativity discussions, using a count of the number of pages in each paper that contain some mention of the creativity of their system (measured in units of half a page and expressed as a percentage of the number of pages in the paper).[54] The 75 papers surveyed were on average 9.9 pages long, with an average of 1.4 pages per paper that contained some discussion of computational creativity. Over half the papers (40 / 75) contain less than

[54]Occasionally the percentages do not give the correct totals, e.g. the percentages for the 5^{th} and 6^{th} lines should sum to 77% (% of papers mentioning evaluation) but actually sum to 78% due to rounding of decimal places.

Table 2.4: Summary of the evaluation approaches used in the 75 surveyed papers that presented a computational creativity system.

Paper makes at least a mention of evaluation	77%
Paper states evaluation criteria	69%
Paper performs evaluation of their system	55%
Paper contains section(s) on Evaluation	51%
Main aim of evaluation: Creativity	35%
Main aim of evaluation: Quality/Accuracy/Other	43%
Mention of creativity evaluation methodology	27%
Use of creativity evaluation methodology to evaluate system's creativity	24%
System compared to other systems	15%
System compared to other systems produced by other researchers	11%
Systems evaluated by people other than system implementers	33 %

10% of such discussion, with an average of less than 1 page per paper. This includes 26 papers that do not discuss the creativity of their system at all; this was somewhat surprising for papers that are specifically aimed at a computational creativity research audience, presenting a computational system as a creative system. Overall the papers vary a great deal as to how much they discuss computational creativity. Based on the subjective rankings allocated during the survey:

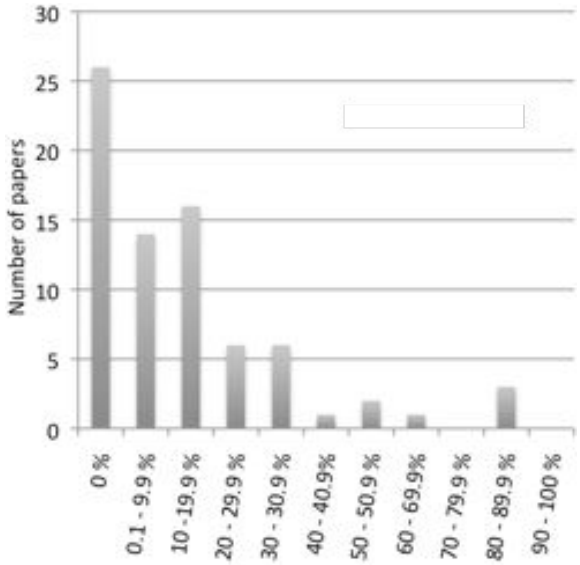


Figure 2.5: Percentage of pages per paper that contain some discussion of the system's creativity.

- 6 papers' *main focus* was on the creativity in computational systems.

- 3 papers discussed the creativity in computational systems *a lot*.
- 3 papers discussed the creativity in computational systems *quite a lot*.
- 19 papers discussed the creativity in computational systems in *some* depth.
- 2 papers discussed the creativity in computational systems in *a little* depth.
- 16 papers discussed the creativity in computational systems in *very little* depth.
- 26 papers included *no* discussion of the creativity in computational systems.

To summarise the key findings from this survey, evaluation of computational creativity is not being performed in a scientifically rigorous manner. From 75 computational systems presented as being creative systems, the creativity of a third of these systems were not even critically discussed when presented to an academic audience in paper format. Half the papers surveyed did not contain a section on evaluation. Only a third of systems presented as creative were actually evaluated on how creative they are and less than a quarter of systems made any practical use of existing creativity evaluation methodologies. Occurrences of evaluation being done by people outside the system implementation team were rare, as were any examples of direct comparison between systems, to see if the presented system represents some research progress in that field.

2.3.3 Considering the survey results in the context of scientific method

It is apparent that across the time period covered by the surveyed papers, scientific method is not being adopted as standard within the community for computational creativity evaluation. In the majority of papers, there was no reference to an overriding theory (or competing theories) about what computational creativity is.[55] No papers formally stated hypotheses for testing the systems in terms of in what circumstances the system could be creative, though the 24% of papers that used a creativity evaluation methodology did indirectly state their hypotheses for testing, by their adoption of their chosen methodology. Of the 58 papers that performed some sort of evaluation (of any aspect of the system, not just its creativeness), 52 papers stated evaluation criteria but rarely in the context of an overriding theory or hypothesis from which these evaluation criteria were drawn[56] or an explanation of why those criteria had been adopted.

Given that only 35% of papers did perform evaluation of the creativity of their systems, and also given the diversity and sparse adoption of methods being used for evaluation, the evidence as a whole suggests that the community as a whole is not working within a particular paradigm or theoretical framework for creativity, nor that there are competing paradigms which are being adopted by smaller subgroups within that community. Instead, creativity evaluation (and to some extent evaluation in general) is shown in this survey to be progressing along a more ad-hoc, individualistic basis.

[55] Exceptions here include those papers that adopted a version of the Ritchie (2007) formalisation of creativity, or Colton's characterisation of creativity in the creative tripod model (Colton, 2008b).

[56] Again, those systems that used existing models of creativity were notable exceptions.

As discussed in Chapter 1 Section 1.5, it would be inaccurate to identify the computational creativity community as a solely scientific community. Many researchers within the community would not see themselves as scientists. The desire to improve computational creativity systems and build our knowledge of computational creativity, though, is demonstrated in the recent strengthening of the calls for better evaluative practice (as discussed in Chapter 1 Section 1.5.

From the review of scientific method in Section 2.2 of this Chapter, what emerges is the usefulness of having a hypothesis or theory and clearly stating how it is being tested. In the case of scientific method, this would refer to the empirical conclusions arising from the adoption of a hypothesis, which can be tested. In the context of computational creativity evaluation, what hypotheses are being/should be tested in this scenario? The focus of this thesis is on how computational creativity researchers should best evaluate the creativity of their systems, therefore the question becomes: what does it mean for this system to be creative? A hypothesis for testing would therefore centre around what answer a computational creativity researcher gives to this question, and the hypothesis would be tested by how they employ this answer for evaluating the creativity of their system. No attempt shall be made to define that hypothesis or its conclusions here;[57] however the need for clarity over these aspects has emerged from Section 2.2 as being important for progress, but has been demonstrated to be lacking in current research practice in computational creativity, according to the survey results.

2.3.4 The types of evaluation being used in the surveyed papers

This Section describes and critiques the types of creativity evaluation encountered in the 75 surveyed papers, highlighting any notably useful or inadequate examples in approaches taken either in one paper or across a larger cross-section of the surveyed literature.

In general, no creativity evaluation methodology has emerged from the Section 2.3 survey as the standardised choice for computational creativity evaluation, leading to a lack of a communal approach and a scarcity of examples of any particular methodology to guide researchers for good practice within the field of computational creativity research. Where references to existing creativity methodologies appear in papers, they are mentioned in discussion or only loosely applied for evaluation, rather than being applied as intended by the authors. For example, Norton et al. (2010) mention Colton's creative tripod during the introduction and in occasional further references but do not explicitly apply the creative tripod to discuss the creativity of DARCI.[58] The DARCI system (Norton et al., 2010) is evaluated for other aspects beyond its creativity, namely how well it learns adjectives and how well DARCI can label images with associated words, by surveying users. Interesting findings are uncovered such as how disagreement between humans on interpretations of images implies that DARCI does not have to agree with the common consensual opinions of humans all the time. These findings

[57] The issue of whether one universal answer exists for the question "what makes this computational system creative?" shall be examined in Chapters 3 and 5.

[58] Chapter 7 onwards will return to the evaluation of the creativity of this system as part of Case Study 2 in this thesis.

in turn lead the authors to conclude that 'Clearly, other metrics are needed to truly evaluate DARCI' (Norton et al., 2010, p. 32). [59] Similarly, after some discussion of the creativity of the STANDUP system in Ritchie et al. (2007), particularly in relation to JAPE (Binsted et al., 1997), a similar system to STANDUP and one that some of the authors have worked with, Ritchie and colleagues decide that 'these criteria [Ritchie (2007)] are insufficiently subtle for making fine comparisons of creativity' (Ritchie et al., 2007, p. 97). The authors then move evaluative focus to more quality-oriented tests measuring improvement in communication and soliciting target users' feedback on the system's usability and effectiveness. Another citation of Ritchie's criteria does not fully acknowledge the scope of Ritchie's proposals: Riedl and Young (2006) justify their use of novelty and value in relevant theory (Ritchie, 2001), without acknowledging the importance placed on typicality of products, the inspiring sets that assist the system, or other content in Ritchie (2001) on modelling creativity.

A view of creativity that is often adopted in the surveyed papers[60] is to treat creativity as the combination of novelty and value of the system, particularly of its output (Gero, 2010; Pérez y Pérez, Aguilar, & Negrete, 2010; McCormack, 2009; Aguilar, Hernandez, Pérez y Pérez, Rojas, & Zambrano, 2008; Alvarado Lopez & Pérez y Pérez, 2008; O'Donoghue, 2007; Pereira & Cardoso, 2006; Gomes, Seco, Pereira, Paiva, Carreiro, Ferreira, & Bento, 2006; Riedl & Young, 2006).[61] To illustrate, in his discussion of future evaluation Gero (2010) plans to examine the quality of analogies produced by his analogy generator system and compare the output to previous generated analogies (i.e. novelty). Often in the survey, a system's creativity would be evaluated via inspection of a combination of novelty, value and some (variable) third factor that is usually more relevant to the system's needs as a whole (Machado & Nunes, 2010; Saunders, Gemeinboeck, Lombard, Bourke, & Kocaballi, 2010; Saunders, 2009; Riedl, 2008; Gervás, Perez y Perez, Sosa, & Lemaitre, 2007; O'Donoghue, Bohan, & Keane, 2006).[62] Saunders and colleagues, for example, examine the novelty, 'hedonism' (or pleasingness, i.e. value) and curiosity present in their system (Saunders, 2009; Saunders et al., 2010). For example, Machado and Nunes (2010) plan (in the future) discuss future plans for incorporating automated evaluation (although not evaluation of the entire system) based on the diversity (hence contextual novelty), value and complexity of the generated output. Dividing 'value' into two separate perspectives, Gervás et al. (2007) considers the novelty, interestingness and coherence of tales produced by MEXICA, the story generation system.

[59] DARCI has since been evaluated in more detail (Norton, Heath, & Ventura, 2012), using the creative tripod model.

[60] The view of creativity as the combination of novelty and value, or factors heavily related to these two such as originality and appropriateness (respectively), is explored further in Chapter 3 Section 3.4.1. A more detailed report on how various surveys evaluate novelty and value will be given in Chapter 5 Section 5.3.3, during examination of the coverage of existing evaluation tests for *Originality* and *Value*.

[61] Cohen (2009b) pointedly rejects measuring the creativity of system by its ability to generate novel and 'distinctive' output, instead evaluating his artistic system AARON through feedback from audiences at exhibitions and by filtering output for exhibition using his own taste and judgement.

[62] Chapter 3 Section 3.2.3 will make note of the views expressed by Weiley (2009), that creativity contains more than just novelty and value, but is instead definable as 'novelty, value and "x" ' (Weiley, 2009).

Authors rarely see a need to justify modelling creativity as {novelty and value}, probably due to its prevalence within the field. An exception to this is where Riedl and Young (2006) ground their use of novelty and value in relevant theory (Ritchie, 2001), though Riedl and Young uses Ritchie (2001) to justify the use of novelty and value without (as mentioned above) acknowledging the contribution made to modelling creativity by the entire 2001 paper by Ritchie. The diversity of tests employed[63] for value (or its strongly related counterparts) show what the above-mentioned example in Gervás et al. (2007) illustrates: that one interpretation of system value does not necessarily correspond to another. Also, evaluation of novelty (or originality, newness) is often examined in the papers cited above according to how dissimilar the system's artefacts are to previous output or other existing examples of creative output in that domain (e.g. Saunders et al., 2010; Alvarado Lopez & Pérez y Pérez, 2008; Gervás et al., 2007). On the other hand, appropriateness is often evaluated according to how similar the system's output artefacts are to known examples (e.g. Pérez y Pérez et al., 2010; McGreggor, Kunda, & Goel, 2010; Alvarez Cos, Perez y Perez, & Aliseda, 2007; Gomes et al., 2006). Hence across the field as a whole, there is a stark inconsistency as to whether to prioritise the generation of artefacts which are dissimilar from existing artefacts, or whether to pursue the generation of artefacts which are similar to existing artefacts, arising directly from the adoption of 'novelty + value' as the underlying model of creativity. Such a contradiction is clearly not helping the identification of coherent and consistent strategies to adopt across the field.

Another relatively frequent evaluation scenario in the surveyed papers is that authors conduct the evaluation of their system by how its output is received by its intended audience (Machado & Cardoso, 2002; Ritchie et al., 2007; Widmer et al., 2009; Edmonds, 2009b; Cohen, 2009a), though this is usually in the form of exhibitions or informal solicitation of feedback rather than more formally controlled experiments and evaluative feedback is not generally collected in a systematic way, especially on the system's creativity (as opposed to the quality or aesthetics of its generated products).[64] Some authors (Widmer et al., 2009; Bushinsky, 2009; Edmonds, 2009b; Cohen, 2009a; Machado & Cardoso, 2002) conducted tests seeing how their creative system fared against human opposition, occasionally in a 'blind' manner reminiscent in some ways of a Turing test (Turing, 1950) where the judge does not know if the creative entity is a human or a machine.[65] Others use a competition type of evaluation by seeing how one system performs in competition with others (e.g. Bushinsky, 2009; Widmer et al., 2009; Pérez y Pérez et al., 2010) or directly present the output of their systems

[63] See Chapter 5 Section 5.3.3 for details.

[64] A partial counter-example to this is the work by Pearce and Wiggins (2007), who use a derivative of the Consensual Assessment Test (CAT) (Amabile, 1996) (see Chapter 3 Section 3.4.2) to conduct controlled tests to collect feedback, though they employ their version of CAT to evaluate the quality of their system, rather than its creativity.

[65] The direct use of a Turing-style test has been criticised by Pease and Colton (2011b) for encouraging pastiche and homogeneity rather than diversity, rewarding creativity that stays within constraints of a given stylistic boundary without testing those boundaries (c.f. Boden (2004)'s exploratory rather than transformational/combinatorial creativity), while shifting the development emphasis on superficial matters (relative to creativity) such as the front-end of the program.

to audiences to see how they are received (e.g. Machado & Cardoso, 2002; Edmonds, 2009b; Cohen, 2009a). These are interesting and useful forms of overall system evaluation but have not been used to specifically address *creativity* evaluation;[66] for example, if Bushinsky (2009)'s Deep Junior chess player wins matches against human and machine opponents, it is clearly performing well as a chess player, but in the question of how creative Deep Junior is, this type of evaluation provides no answers. Also, the choice of using competitive scenarios for evaluation is tempered by how clearly the judging criteria are specified, how relevant they are and how closely they are adhered to. As Widmer et al. (2009) remarks, even though their YQX musical system is tested by a variety of contests, acknowledged flaws in YQX's musical knowledge are not noticed by the judges. Widmer et al. (2009) feel the need to add a critical discussion of their system using standards identified from related literature: intentionality, conscious awareness of form and structure, aesthetics, imagination, skill and self-evaluation. Stating these criteria is good practice (though as found in the findings in Section 2.3.2, not common practice) as the criteria can be critiqued, and forms part of this paper's contribution, as the criteria have been derived from study of relevant literature and can be learnt from by others who wish to compare similar systems or perform similar evaluation.

Widmer et al. (2009) is not the only example where specific evaluation criteria for creativity are stated, though it is one of only a few isolated examples where the criteria used are explicitly justified in terms of their applicability and relevance for creativity as demonstrated by that system. In general, papers that perform some form of evaluation do state the criteria by which they are evaluating their system. In papers where the opposite is true (Bushinsky, 2009; Cohen, 2009a; Edmonds, 2009b) the system is not being evaluated for creativity but for the quality of the system: the evaluation being performed is either left down to aesthetic judgements of the appeal of the system's products (Cohen, 2009a; Edmonds, 2009b) or seeing how the system performs in a competition, specifically, a chess match (Bushinsky, 2009). Failing to explicitly *justify* choices of evaluation criteria, particularly related to how creative the system is, is however a recurring theme in many of the papers reviewed (e.g. Riedl, 2010; Zhu & Ontañón, 2010; Gervás & León, 2010; Saunders et al., 2010; Saunders, 2009; Gervás et al., 2007; Hassan, Gervás, León, & Hervas, 2007; Oliveira, Cardoso, & Pereira, 2007; Veale, 2006b). In such cases, evaluation criteria are stated but the reasons why they have been chosen as appropriate evaluation criteria are not justified. Occasionally one cannot see the wider relevance of why they have been chosen to evaluate the system, particularly the system's creativity. For example, Zhu and Ontañón (2010) 'propose three dimensions to classify the landscape of analogy-based story generation: 1) the scope of analogy, 2) the specific technique of computational analogy, and 3) the story representation formalism'. It is not clear why these particular dimensions were selected. Saunders and colleagues (Saunders, 2009; Saunders et al., 2010) evaluate their agent-based systems using the Wundt curve as a measure of novelty against 'hedonism', or pleasingness (Saunders, 2009; Saun-

[66] As has been stressed throughout, this thesis work concentrates on the issue of *creativity* evaluation.

ders et al., 2010) and also develop the notion of curiosity as the driving force behind actions taken to reduce uncertainty through exploration of a creative field over time, without relating this notion back to what it means to be curious, and without developing the wider question of why this combination of factors is used for evaluation. In a discussion considering various aspects of the story-generation system VB-POCL, e.g. coherence of stories produced, limitations, Riedl (2010) asks: 'Is VB-POCL creative?' but then focusses on a 'practical creativity' that only addresses whether the program terminates successfully and the ability of the system to find stories that other computational systems cannot (without direct reference to some systems). Again, justification for selecting these criteria (and not others) is left implicit rather than being explicitly stated. The evaluation in Gervás and León (2010) is related back to the broader creative system framework proposed by Wiggins (2006a), with the use of an evaluation function E (required as part of this framework) to evaluate if the story is 'good' or not. This function E is based on evaluating a number of 'significant variables', story-specific variables that represent the story content from a number of perspectives, in detail, such as the interest of the story, danger in the story, tension, amount of action, and hypotheses made.[67] The question of whether the system is creative or not is not addressed (except via its partial placing within Wiggins' framework). Hassan et al. (2007) does consider how creative certain aspects of their story generation system are, looking at measurements of world construction and story construction parameters, content planning and sentence planning. These evaluation criteria are used to discuss creativity inherent in the system, but are extremely closely fitted to the system in Hassan et al. (2007) and how it operates. No reasons are given for choosing these specific criteria to discuss the creativity of the system and questions arise as to whether these criteria are the most relevant for evaluating creativity. They are also limited in their usefulness for evaluating other story generation systems, for comparison, if for example another system does not implement the construction of story worlds. There is no overall consideration of the whole system's creativity on a more general level.

This inadequacy in justifying evaluative choices occurs across several of the papers surveyed, but not uniformly. There are a relatively small group of papers which do justify the criteria chosen for evaluation, such as Widmer et al. (2009), mentioned above, which chooses evaluation criteria based on what has been seen to be important (independently of the system development, in relevant literature), or Monteith et al. (2010) and (Chan & Ventura, 2008),where the chosen evaluation criteria were based on the adoption of the creative tripod (Colton, 2008b) as an underlying model of creativity.

Saunders et al. (2010) and Oliveira et al. (2007) also offer examples of another common evaluation inadequacy in the reviewed papers: plans for future evaluation are outlined but actual evaluation has not been carried out yet. This situation occurs reasonably often (e.g. Machado & Nunes, 2010; Saunders et al., 2010; López, Oliveira, & Cardoso, 2010; Aguilar et al., 2008; Forth et al., 2008;

[67] Similar (but less detailed) evaluations are made in an earlier related paper (Gervás et al., 2007), where the system is evaluated on its ability to produce novel, coherent and interesting stories, but this is not based within Wiggins' framework.

Hull & Colton, 2007; Oliveira et al., 2007; Swartjes & Vromen, 2007). For example Machado and Nunes (2010), who offer an evolutionary image generation system, talk about automating the fitness function of their system (which currently uses user intervention) in future work, partially to allow the authors to evaluate quality, diversity and complexity of the images. As another example, in López et al. (2010) no actual evaluation has taken place, but listening tests are proposed to gauge how fluent the generated songs are. (Hull & Colton, 2007, p. 142) acknowledge 'that evaluation of systems and the artefacts they produce is an essential aspect of computational creativity which is missing from the work presented here and we aim to fill this gap.'

Evaluation of aspects of the system other than its creativity tended to occur more frequently and was often more rigorously conducted, as might be expected for scenarios where evaluation methodologies and approaches are already established, such as tests for precision and recall (e.g. Chuan & Chew, 2007; Veale, 2006b) or statistical evaluation based on testing hypotheses about the system (e.g. Colton, Gow, Torres, & Cairns, 2010; Mozetič, Lavrač, Podpečan, Novak, Motaln, Petek, Gruden, Toivonen, & Kulovesi, 2010), or cross-entropy measures to examine the distribution of produced artefacts (Whorley et al., 2010, 2007).[68] Though a broader evaluation is important in the research process, this survey is specifically concerned with systems that are being presented as *creative systems*. If evaluation of the *creativity* of these "creative systems" is not being performed, then claims are being made about the system's ability to be creative that are are not addressed or verified.

In some cases, evaluative tests are conducted on the system which purportedly evaluate the system's creativity but which actually only measure the system's quality (Collins et al., 2010; Riedl, 2010; O'Donoghue, 2007). As mentioned above in Section 2.1.7 Collins et al. (2010) propose creativity metrics but then acknowledge that their tests instead evaluate the quality of their system. O'Donoghue (2007) mentions several factors for evaluation such as novelty, plausibility, surprisingness, applicability and usefulness (citing Colton and Steel (1999)) and does conduct initial tests for novelty (against the existing knowledge base of the system), but then focuses all evaluative testing and discussion specifically on the quality of the analogies generated by his system. The slight change in focus is briefly acknowledged by O'Donoghue (2007) in the conclusion. Similarly, although Riedl (2010) devotes paper space to discussing what it would entail for the VB-POCL story generation system to be productive, the focus then moves towards a 'practical creativity' (Riedl, 2010, p. 609, p. 612,) that only tests for successful program termination and for the ability ability of the system to find stories that other computational systems cannot (without direct reference to some systems). Returning to the issue noted in the previous paragraph, the quality of the program for story telling is

[68] Whorley and colleagues Whorley et al. (2010, 2007) uses a cross-entropy measure (Whorley et al., 2007) to evaluate their system and present some preliminary results. Their system, however, uses cross-entropy measures in model construction and also to evaluate the models, hence the models are being evaluated by the same criteria that is used to shape the models' generation rather than using an independent evaluation measure. 'An information-theoretic measure, cross-entropy, is used to guide the construction of models, evaluate them, and compare generated harmonisations. The model assigning the lowest cross-entropy to a set of test data is likely to be the most accurate model of the data.' (Whorley et al., 2010, p24).

evaluated in place of addressing the factors that Riedl has previously highlighted in another surveyed paper as important for determining how creative the system is: 'the output of a creative system [such as this one] must be novel, surprising, and valuable' (Riedl, 2008, p. 42).

Sometimes evaluation (and in particular, creativity evaluation) does appear prominently in the paper, but solely as a component of the system's operation rather than an investigation of the system's overall creativity(Gervás et al., 2007; Bird & Stokes, 2007; Miranda & Biles, 2007; McCormack, 2007; Alvarado Lopez & Pérez y Pérez, 2008; Saunders, 2009; Saunders et al., 2010; Pease, Guhe, & Smaill, 2010; Pérez y Pérez et al., 2010; Jordanous, 2010c).[69] While it is important to demonstrate the system's ability to reflect and self-evaluate (Poincaré, 1929; Wallas, 1945; Finke, Ward, & Smith, 1992; McGraw & Hofstadter, 1993; Boden, 1998, 2004), this should not be at the expense of evaluation of the overall system, to evaluate claims of it being a creative entity. This is recognised by some authors; for example, in the surveyed work of Saunders and colleagues (Saunders, 2009; Saunders et al., 2010) on multi-agent systems, only the agents are evaluated, rather than the system as a whole. Future evaluation is planned, however, to quantify behavioural diversity between agents and in system as a whole and also to test "how humans interacting with an artificial creative system construe the agency of the robots."(Saunders et al., 2010, p. 107).

Where the creativity of the system is evaluated in some way, often papers do not perform evaluation but merely give summative statements that cursively describe the success of their results, with no analysis or justification. These authors are losing the opportunity to test and demonstrate the success of their systems in more methodical and academically recognised ways, weakening the contribution of their paper. For example, for overall evaluation:

> 'We are interested in problems of computational creative discovery where computer processes assist in enhancing human creativity or may autonomously exhibit creative behaviour independently.' [with no questioning of whether either of these two aims are achieved] (McCormack, 2009, p. 1)
>
> [on their harmonic progression system:] 'it is a creative system based within the "post-tonal" harmony found in certain 20th century musical styles.' (Eigenfeldt & Pasquier, 2010, p. 16)

To summarise, various types of creativity evaluation have been seen in the 75 surveyed papers (with no particular approach or method standing out as a universal option across the field). While positive examples of good creativity evaluation practice exist, for example the consideration of what makes the YQX system creative (Widmer et al., 2009) in relation to the surrounding research on musical expression, there are a number of inadequacies in the surveyed literature:

- If existing creativity methodologies are mentioned, they are often not applied as intended but referenced more briefly.
- A fairly popular underlying model of creativity that was adopted for evaluation was to treat

[69] It is acknowledged here that Jordanous (2010c) is one of the papers criticised on this issue.

creativity as the combination of novelty and value of the system,[70] especially in terms of its output. Occasionally a third factor is also used, which varies depending on the system. The manifestation of value in a system (or of appropriateness, or usefulness, or quality, etc) has been interpreted and tested for in different ways. Occasionally tests for appropriateness or quality in particular reward the opposite of what is normally evaluated for novelty, i.e. (dis)similarity between the system's products and existing examples of creative output in that domain. Adopting this model is therefore leading to some contradiction in strategies adopted across the field.

- A number of authors evaluated their system by how its output is received by its intended audience; however this is usually related to the system's aesthetic qualities and/or usefulness rather than relating to the system's creativity.[71]
- In some cases a system was evaluated by comparing its output to that of other systems (perhaps in a competitive way) or to examples of human creativity in the same domain. These were often blind tests, such that the evaluator did not know the creator responsible for the artefacts they were evaluating. While the comparisons generally provided useful feedback, again these tests were often used to evaluate system quality rather than creativity, or reward aspects of the system that are not necessarily contributory to their creativity (Pease & Colton, 2011b).
- Usually papers that performed evaluation stated their evaluation criteria fairly clearly. In several papers, though, the relevance of the evaluation criteria used was not justified, particularly if used for evaluation of the creativity demonstrated by that system. Occasionally the choice of evaluation criteria seemed inappropriate, overly specific or somewhat ad hoc.
- Papers that detailed evaluation plans for future work but that reported no evaluation had actually been carried out were noted in the survey disappointingly often.
- Often papers did not include evaluation of creativity at all, despite the systems in the papers being presented as creative systems.
- Looking at one specific example, Whorley et al. (2010) uses the same measure to evaluate generated musical models that is being used to generate the models, rather than using an evaluation measure that is independent of the generative process. This could lead to skewed results as the metric is essentially being applied to evaluate itself.
- In some cases, creativity evaluation tests are proposed and carried out which actually measure the system's quality rather than its quality.
- Sometimes evaluation (and in particular, creativity evaluation) only appears in the paper as a component of the system's operation rather than as an overall evaluation of the system's

[70] Chapter 3 Section 3.4.1 will look at how the combination of novelty and value has been used as a model of creativity in research into both computational and human creativity.

[71] As the Case Studies in this thesis shall show (see in particular the survey of human opinion in Chapter8), people can struggle to decide how creative they think something is and find it difficult to justify their opinions once formulated, instead relying on intuition or judgements of quality as a substitute for judgements of creativity.

creativity.

- It was fairly common to see papers labelling their system as creative with little methodical analysis or investigation, through only brief and cursory descriptions of how their system is creative.

Where specific examples of less adequate evaluation criteria/practice are given and commented on, this is done in terms of thinking about possible improvements or noting matters that need to be treated with more care or attention. A more detailed report of deployed evaluation methods is in Chapter 5 Section 5.3.3, which focuses on what aspects of creativity are being tested.[72]

2.3.5 Reasons behind the current state of creativity evaluation practice

The survey findings reported above show that for creativity evaluation, there is a lack of direction and standardisation of practice within the computational creativity research community, to the extent that creativity evaluation is often reported poorly or missed out entirely. It is important to try and understand why this situation has developed within the computational creativity research community. Notwithstanding the negative preconceptions about computational creativity that need to be overcome if a system is to be fairly evaluated by an audience, the idea of evaluating the creativity of computational systems is sometimes seen as being too complex to attempt.[73] Nonetheless, the research community is, in general, well informed and conversant with key issues and problems behind the evaluation of the creativity of 'creative systems'. Debates on such issues and problems have taken place often during these formative years of this research field, both in discussions between people in the field (Pease, 2012, personal communications) and in the context of relevant publications.[74]

The International Computational Creativity Conference (ICCC) showcases current trends and advances in computational creativity research. One discussion session at the ICCC'11 conference debated the current evaluative culture in computational creativity. This discussion arose after the conference session including my paper at ICCC'11, which presented the results of the evaluation survey as reported in Chapter 2 Section 2.3, highlighting weaknesses in evaluation of creativity in this research field and motivating the need for a standardised and comprehensive creativity evaluation methodology (Jordanous, 2011a). This paper briefly discussed issues around defining and evaluating creativity as well as reporting some content from later Chapters of this thesis.[75] This debate was particularly illuminating, with several points raised as to why researchers did not include (or did not report)

[72] Chapter 9 Section 9.2 will return to the question of what is not being tested for, or is being treated inadequately, in the context of how the suggestions of this thesis address such inadequacies and omissions.

[73] This issue arises at various points in this thesis, mainly in Chapters 1 and 3).

[74] As discussed earlier in this Chapter, in Section 2.1.

[75] In Jordanous (2011a), the 14 components of creativity derived in Chapter 4 were proposed as part of a solution to the definitional problem, while the Evaluation Guidelines (Chapter 5 Section 5.2) were proposed as a heuristics-based solution to some of the evaluation issues. The paper also reported results from a pilot study evaluating musical improvisation systems, a preliminary study to Case Study 1, in Chapter 6.

creativity evaluation in their papers. The discussions demonstrated how people had engaged with the issues surrounding creativity evaluation on a number of levels. It encapsulated the most recent community-wide discussion on these issues, with opinions being voiced by people from a variety of academic backgrounds and levels of experience. In their contributions to discussion, several people also concurred that a closer examination of creativity evaluation was necessary and a more systematic approach should be taken by the field as a whole.

Below is a review of the main points arising at this 2011 discussion with regards to the content in Jordanous (2011a) and my responses to them, both at the time (when relevant) and upon further reflection.[76] All quotes are transcribed from a recording of the discussion session and anonymised.

Researchers' reasons for not doing evaluation

> 'I'd like to ask Anna whether you asked the authors of these 75 papers ... why they've not done sufficient evaluation ... because I think there might be reasons, good reasons ... we aren't necessarily bad researchers for not doing evaluation, there's good reasons why people don't do it, and I was wondering if you have any answers to that?'

Although my paper (Jordanous, 2011a) was careful not to label research with little scientific evaluation as 'bad', the paper was critical of such research if the research was positioned as scientific research and/or made claims that the system was creative without justification. Citing warnings about developing a 'methodological malaise' in computer music composition systems (Pearce et al., 2002), my response emphasised that even though creativity evaluation is complex to address, more rigour in our approach to evaluation would help maintain the credibility and progression of our research.

Evaluation of creativity could include evaluations done in alternative environments to the traditional idea of formal evaluations.[77] Differing views arose during discussion as to whether this type of evaluation would be relevant content for a technical paper. This thesis adopts the view I gave in the discussion; if evaluative feedback is useful in verifying claims made in a paper that would otherwise be left unverified, then that feedback should be included. People mentioned the validity of using audience evaluation and how different information can be gained from using people with different personal contexts to evaluate, such as through levels of expertise, geographical context or academic inclinations and expectations.[79] There are scenarios where a rigorous, formal evaluation procedure

[76] It should be noted that the discussion session covered four other papers, so some points described here are influenced by other presentations as well as mine. Only those points relevant to my paper have been included. Perhaps because my presentation was last in this session, immediately preceding the discussion, or perhaps because the critical nature of my paper made my presentation quite controversial, my paper received a large percentage of attention during the discussion.

[77] As can be seen in the Evaluation Guidelines described in Jordanous (2011a),[78] a researcher might choose to define creativity in art (in whole or in part) as a positive audience reaction at an exhibition (Step 1). *If* this definition could be justified as an appropriate interpretation of *creativity* in art (as opposed to quality) then audience reaction would be a standard to test the system by (Step 2). A test for this standard (Step 3) would therefore be to exhibit the artwork produced by a system and gauge audience response. In fact, the component *Social Interaction and Communication* (see Chapter 4 and Jordanous (2011a)) could incorporate exactly this type of evaluation.

[79] A suggestion from members of the steering committee for future ICCC conferences was to include an exhibition of works from creative systems. Such exhibition events would potentially be a new avenue for the systems to use to gather feedback from audience reactions.

may not be appropriate, as illustrated above. This does not mean that the system is left unevaluated. The evaluation survey in Section 2.3 treated as evaluation any form of getting feedback on the system. The situation being highlighted by the evaluation survey was where systems were developed and presented without feedback *of any kind* having been elicited.

Difficulties in finding systems for comparison to your own

> '... the point about comparing to other systems ... in another field like machine learning for example, it's pretty easy to find ten other systems that are doing exactly what you claim to be doing and you sort of run them all and compare them straight across. But in this field it seems much trickier to find even one other system thats doing exactly what you're doing and therefore I'm not sure that you can do that, and secondly I'm not sure how you would do it.'

Section 2.3 criticised situations where systems are evaluated without contextual references to the context of research progress in that area as a whole. If a researcher is unaware of related work by others, then they cannot learn from that work and their own work is potentially less likely to make a highly relevant contribution to the advance of research in that particular field. As a research field, activity in computational creativity research has increased greatly over the last decade or so, as shown by the progress in research events and the recent journal special editions focused in this area (see Section 2.3). Many creative systems have now been developed.[80] If similar systems exist and a body of research builds up in a particular area, then it is useful to consider how research in that area is progressing collectively. This benefit has been demonstrated in research into narrative/story-generation systems (Gervas, 2009), a long-standing and thriving research area within computational creativity research (e.g. Meehan, 1981; Turner, 1994; Bringsjord, 2000; Pérez y Pérez & Sharples, 2004; Peinado & Gervas, 2006; Peinado, Francisco, Hervás, & Gervás, 2010; Tearse et al., 2011).

This question sparked a debate about whether the central aim of a creative system was to generate products that no other systems generate, or to generate behaviour distinct from all other systems. If a system is unique, then how can it be compared to other systems? My specific response to this point was to question how one would be able to find out if a system is producing unique results, if its output is not compared with other systems.[81]

There are situations where a creative system operates in a niche where no other system exists. For example, the ERI-Designer is believed to be the sole exemplar system of creativity in furniture arrangement (Pérez y Pérez et al., 2010; Aguilar et al., 2008). In these cases, direct comparisons between two equivalent systems cannot be made, however comparisons could be made between the system and humans performing the same task (as in Pérez y Pérez et al., 2010), or with a considered comparison of the appropriate crossovers with systems operating in a reasonably similar domain.[82]

[80] The Section 2.3 survey looks at 75 papers describing such systems.

[81] For example, *Originality* is one of the 14 components used to evaluate systems in Case Studies 1 and 2 (Chapters 6 and 7 respectively). It is unclear how performance on this component could be evaluated on one system in isolation. *Originality* does, however, form an important part of creativity, as seen in the Case Study Chapters and in Chapter 3 Section 3.4.1, and is important for determining the uniqueness of a system.

[82] ERI-designer, mentioned above, could perhaps be compared to architectural design systems or game design systems.

Another point made during discussion was that different versions of the same system could be compared, to see what improvements have been made and to measure progress. An example was described in a comment during discussion of a different paper:

> 'I tend to apply just engineering solutions and try and get new results and look at the results I've had before and see if they're better and so on and apply modifications'

Some of the papers reviewed in the Section 2.3 survey did extend and develop existing systems (e.g. Whorley et al., 2007, 2010), MEXICA and its variants (Gervás et al., 2007; Montfort & Pérez y Pérez, 2008; Pérez y Pérez, Negrete, Penãlosa, Aṽila, Castellanos, & Lemaitre, 2010), ERI-designer (Aguilar et al., 2008; Pérez y Pérez et al., 2010). Generally comparison between different versions of the system was limited but was present to some degree.

The original questioner clarified that their question was about whether it was appropriate to compare systems in different domains with different requirements. A point made during discussion was that a broad evaluation from a wide perspective can be performed on systems which are fundamentally different; we can learn both from the evaluation results and by understanding the ways that the systems are different.[83] Distinctions were drawn between evaluation of one system as a single research project and the evaluation of progress of a particular strand of research.

There are some types of systems that are so fundamentally different that there is no area of crossover to compare; however as Chapter 3 Section 3.6.4 will point out, some aspects of creativity are universal across different systems. Whilst the amount of crossover between system domains determines the extent to which the systems can be compared,[84] we do not need to be restricted to evaluating systems that are very similar before meaningful comparisons start to emerge.[85]

Difficulties of evaluating creativity

A perennial issue was raised during discussion, concerning the complex nature and difficulties of evaluating creativity. One comment focused on the somewhat contradictory nature of creativity:

> 'some things that seem like perfectly reasonable evaluative criteria, that on the surface everybody would say yeah, you should - yeah that system should score highly on that, in fact in the long run may be detrimental.'

An example was given of how human creativity can involve contradictions, non-systematic processes and even flaws, all of which we try to remove from our systems in the processes of refinement.

Another comment put into words the variety of opinions expressed during discussion:

[83] The SPECS methodology which is introduced in this thesis, in Chapter 5 and onwards, enables such comparison, by identifying similarities and differences between domains in terms of aspects of creativity and priorities for each domain. Such a comparison will be demonstrated in Case Study 2 (Chapter 7).

[84] For example, see Chapter 7.

[85] Another issue mentioned during this part of the discussion was that it was often difficult to locate materials to use for evaluation, such as the system's source code or products. This is a valid concern that was also encountered during Case Study 1 (Chapter 6) and is part of the justification for evaluating systems based on limited available materials and information (Case Study 2, Chapter 7). See Chapter 9 Section 9.1 for further discussion.

> 'One of the things actually that I'm a little confused about from the discussion currently, is - I'm curious what everyone right now was thinking about - is the issues and the scope of what we're talking about with creativity in that you [ICCC'11 attendees] seem to have a lot of different approaches ...'

This comment developed some examples of different notions of creativity:

- To what extent creativity is seen as domain-specific.[86]
- Whether creativity should be evaluated at a specific point in time or over a longer period (or if this choice should vary depending on the type of system or the specific creative domain).
- Who is evaluating the system.[87]

It was questioned to what degree human creativity should be used as a basis for computational creativity. This impacts upon whether assessment methods for human creativity are relevant for computational creativity, an especially pertinent point given that human creativity assessment is by no means a solved problem.[88] While the prevalent understanding of computational creativity by researchers is centred around the computational modelling or simulation of human creativity, there are a number of alternative interpretations of computational creativity.[89]

Multi-dimensionality of creativity

> 'Even people that I think believe creativity is something that you identify with a multidimensional space, I often hear them slip back and forth into language that suggests it's an either/or, a binary choice, it's creative or not. And so the two plus two as a creative act, I thought illustrated very well that creativity is a continuum. On the one hand we've got things which are creative acts but not creative, but it's really a multidimensional space, it's not about 'is creative' or not creative.'

This person's comment is perceptive and raises valid points to remember.[90] Although time limitations prevented much discussion on this matter in my talk, one emphasis made during the talk was that taking a confluence-style approach to creativity[91] allows us to obtain more detailed feedback about how to make our systems more creative.

Different types of evaluation

> 'For me there's five ways that a computational creative system can be evaluated: first you get a paper, and it can be accepted or rejected, and there's the experiments you can make to evaluate the system, and that's the evaluation you're talking about, then there's the evaluation of peers in the domain in which the system is creative, then the expert, like the critics and then finally the audience. And so my question is quite simple: did you have a look at how many of these systems actually met their audience?'[92]

---

[86] As will be discussed in Chapter 3 Section 3.6.4.

[87] This will be discussed further in Chapter 5 Sections 5.1.4 and 5.1.5.

[88] See Chapter 3 Section 3.4.2.

[89] This will be investigated in Chapter 3, particularly Sections 3.4.1, and was briefly mentioned in Jordanous (2011b).

[90] These points previews the sentiments of Chapter 3 Sections 3.2.2 and 3.4.2 in particular.

[91] Chapter 3 Section 3.4.2 introduces the confluence approach to creativity.

[92] See (Eigenfeldt, Burnett, & Pasquier, 2012) for a more formal presentation of these 'five ways that a computational creative system can be evaluated'.

This comment was useful in listing several ways in which a system can be evaluated. My response confirmed that the survey took an inclusive view of evaluation and included scenarios where the system was presented to an audience and audience feedback.[93]

This observation is included to acknowledge the five aspects of evaluation highlighted by this informative comment. A similar variety of points of views was acknowledged during discussions on evaluation at the 2009 computational creativity seminar at Dagstuhl, including the perspectives of 'viewer/experiencer', 'creator' and 'interactive participant (Brown, 2009b, p. 1). This thesis concentrates on post-implementation evaluation of a system, but incorporates within that the possibility of peer evaluation, expert evaluation and audience evaluation.[94] If one of these types of evaluation is prioritised over others, the system can still be treated in the survey as being evaluated in some way. The evaluation survey above highlights situations where systems were developed and presented without feedback *of any kind* having been elicited.

Producing a single score as evaluation of a creative system

> 'I think there's difficulties in trying to ascribe a single number to a creative system? Because I think that creativity is a function of what inputs you provide in a given instance and what you get out, so putting a single number on that is really hard.'

This view resonates with the stance taken in this work that formative feedback is a useful result of evaluation and also echoes views previously expressed in discussions on evaluation at the Dagstuhl seminar in computational creativity (Brown, 2009b) that '[e]valuation can feed back into the system to affect (hopefully improve) future performance' (Brown, 2009b, p. 1). [95]

Issues in providing a standard tool across creativity

> '[Something] reminded me of something I read ...that people were biased to move towards conjunctive definitions of things. And they will go to a level of abstraction necessary to define something conjunctively ... coming up with abstractions and formalisms that covered all the bases. ... And it's a guarantee that the evaluation strategies that you come up with ... are only going to cover a subset of what your current community covers, unless it's a useless abstraction.
>
> And so the question is how do people on the panel believe about that, as you move to improving evaluation, if you're not conscious of having different evaluation criteria for different parts of your community, then parts of your community are going to leave, that's a guarantee, or your community is going to accommodate this disjunction. If your formalism is invented to cover all the bases, it's not going to. How are you, the community, going to handle that fragmentation?

This question was aimed more towards another paper preceding that discussion section (Colton et al., 2011) than my own paper, however it neatly phrased how a tool for assisting researchers needs

[93] In the evaluation Case Studies, Chapters 6 and 7, the *Social Interaction and Communication* component incorporate details about a system's interaction with an audience.

[94] As described in Chapter 5 Section 5.1.6.

[95] There are pros and cons to the approach of measuring creativity via a single score. In the Case Studies in this thesis, the option to produce a single score to measure creativity was purposefully mentioned only briefly and was not followed up on, as it seemed somewhat contradictory to split creativity into various components and then sacrifice the more detailed qualitative feedback on each component for a single score. See Chapter 9 Section 9.1 for further consideration of this decision and its consequences and alternatives.

to be flexible enough to be adopted for several different types of creativity, without falling into a trap of being all-inclusive but being virtually useless for definitional purpose and any real information.

My response (as a discussion panel member) clarified the importance I placed in a standardisable methodology that could be parameterisable and customisable, allowing an appropriate level of domain-specific detail to be given, without being overly restrictive or abstract. As several of the papers in that particular session were directed towards a more clear and transparent approach to research in computational creativity, I also emphasised the importance of being clear about the criteria chosen for evaluation, so that these choices can be critically reflected upon and repeated where appropriate, for greater consistency across research.

Community-defined criteria for evaluation

> 'I have a proposal that I think is for a sort of middle ground that gets at finding some common consensus on what evaluation criteria are within this community, without requiring them as formalised as in the sense that we are discussing.
>
> So in the evolutionary computing community, there's an annual prize for human-competitive results. And there's a panel that decides on the prize. They have very explicit criteria, but they're decided by humans, not by a formalism. And they include things, some of which are actually relevant to this community, like if the system has produced an artefact which has been patented, that is a good thing, and that counts. ...
>
> there's a sort of forum for deciding - in our community here it would be what counts as computational creativity and have discussions around that, in a venue that people care about, if they're competing, and in which criteria are explicitly articulated and debated, and work is recognised.'

This comment is included not as a reflection on my work, as such, but as an acknowledgement for the need for engagement and contributions on a wider basis, that could potentially develop creativity evaluation practice from a community perspective rather than by individual researchers or small teams. It will be intriguing to see if this suggestion is taken further.

2.4 Summary

The past ten years have seen discussion of how best to evaluate the level of creativity demonstrated by computational creativity systems. Several creativity metrics and evaluation methodologies have been proposed, as detailed in Section 2.1. In practice the most significant of these have been Ritchie's empirical criteria approach (Ritchie, 2007, 2001) and Colton's 'creative tripod' framework (Colton, 2008b), with useful contributions to discussion from Pease et al. (2001), Colton et al. (2001), Ventura (2008), Pease and Colton (2011b). As of the time of writing, no one evaluation methodology has emerged as a standard evaluative tool for computational creativity researchers to use, to allow the research community to measure progress using a common methodology.

Often the aim of evaluation has been to see if the systems can replicate human performance as measured against a test set, rather than whether the systems can surpass prior creative achievements by humans and/or computers. These two aims can be confused, especially in the absence of a standard

evaluation methodology for creativity, though as later Chapters will show, these aims need not be mutually exclusive and can be incorporated within the same evaluation framework.

More general observations on best methodological practice can be drawn from reviewing the body of research on scientific method. While computational creativity is not solely a scientific pursuit, the process of verification of a hypothesis in scientific method is closely analogous to the methodological aims of this thesis, for computational creativity evaluation. Section 2.2 highlighted the importance of being clear about what hypothesis is being used in a piece of research, and what evidence is needed to test this hypothesis. Ongoing debates exist over the specific type of method to employ, whether a process of verification or falsification should be followed and if a single method is in fact appropriate for all scientific research. What does emerge during Section 2.2, however, is the importance of clarity in the general approach to evaluating research and confirming theories and hypotheses, and the linking of any evaluative or verification processes to testable parts of a theory.[96]

Section 2.3 has investigated how computational creativity researchers are currently evaluating their creative systems. The key conclusion of the survey was that evaluation of computational creativity is *not* being performed in a scientifically rigorous manner:

- The creativity of a third of the 75 'creative' systems was not critically discussed.
- Half the papers surveyed did not contain a section on evaluation.
- Only a third of systems presented as creative were actually evaluated on how creative they are.
- A third of papers did not clearly state or define criteria that their system should be evaluated by.
- Less than a quarter of systems applied existing creativity evaluation methodologies.
- Occurrences of evaluation by people outside the system implementation team were rare.
- Few systems were comparatively evaluated, to see if the presented system outperforms existing systems (a useful measurement of research progress).
- General principles of scientific method are not being followed by the community as a whole.

Several reasons and justifications for this situation can be seen, as discussed in Section 2.3.5. In particular, the considered debate (prompted by an earlier presentation of this thesis's proposals (Jordanous, 2011a)) at the 2011 conference on computational creativity (ICCC'11) highlights the community's general awareness and acknowledgement of issues that have hampered computational creativity evaluation to date, as well as some thoughts on how to address these issues. Clearly, though, these issues are far from resolved; more active investigation is needed into suitable evaluation methodologies, learning from what has been done so far.[97] In order to evaluate computational creativity, a clearer understanding of creativity itself would be beneficial in informing such investigations. Chapter 3 will more closely examine definitions of the word 'creativity'.

[96] Chapter 5 Section 5.5will look at this further, including the development of the suggestion of a more holistic approach by Bird (1998), and the expansion of scientific method principles to research encompassing more than scientific practice, as is the case with computational creativity research.

[97] This will form the focus of Chapter 5.

Chapter 3

Defining creativity

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012) and a peer-reviewed conference paper (Jordanous & Keller, 2012).

Figure 3.1: *Wordle* word cloud of this Chapter's content

Overview

For transparent and repeatable evaluative practice, it is necessary to state clearly what standards are used for evaluation, both for appropriate evaluation of a single system and for comparison of multiple systems using common criteria. Defining standards for creativity evaluation is by no means straightforward; as Section 3.1 discusses, there is a lack of consensus on the exact definition of the word *creativity* which hinders creativity research progress.

Section 3.2 examines how several issues and debates adversely affect the chances of deriving a suitable, universally accepted definition of creativity. These debates mostly discuss the general nature of creativity, though some specific concerns arise in computational creativity research.

Definitions of computational creativity generally refer to interpretations of human creativity in a computational manifestation (Section 3.4.1), showing the need for a working definition of creativity itself. Section 3.3 argues that dictionary-style definitions such as those in Figure 3.2 are inadequate and impractical for evaluation. Research into human creativity is explored for a suitable definition, with key contributions considered in Section 3.4.2. Some philosophical considerations of how we interpret the meaning of concepts like creativity are presented in Section 3.4.3. Section 3.5, investigates how creativity is defined legally, such as for copyright. Again, though, no universally accepted definition is available, despite several offerings. What does emerge from this research are varying and occasionally contradictory opinions on what is, and is not, creativity. Several different perspectives on creativity exist, which should be considered when conducting research to decide the standpoint taken. These perspectives and the decisions taken in this work are described in Section 3.6.

3.1 The need to define creativity

> '[there are] problems related to assessing artefacts in a domain where the assessment criteria themselves are in flux' (Colton, 2008b, pp. 7-8)

Evaluation standards are not easy to define. It is difficult to evaluate creativity and even more difficult to describe how we evaluate creativity, in human creativity as well as in computational creativity (Sternberg & Lubart, 1999; Veale, Gervás, & Pease, 2006; Kaufman, 2009). Even the very definition of creativity is problematic (Plucker, Beghetto, & Dow, 2004). Consequently there is a lack of benchmarks or ground truths with which to analyse research progress.

Hennessey and Amabile pose a significant question and offer two answers:

> 'Even if this mysterious phenomenon can be isolated, quantified, and dissected, why bother? Wouldn't it make more sense to revel in the mystery and wonder of it all?' (Hennessey & Amabile, 2010, p. 570)

1. To gain a better understanding of creativity, learning from previous work by people such as Picasso, da Vinci or Einstein (cited as examples of creative geniuses and 'amazing individuals' (Hennessey & Amabile, 2010, p. 570)).

2. To learn how to boost people's creativity. In a set of slightly grandiose statements, creativity is hailed as some kind of miracle cure for society's woes, 'driving civilization forwards' (Hennessey & Amabile, 2010, p. 570).

Kaufman (2009) argues that creativity can and should be studied and measured scientifically, but that the lack of a standard definition causes problems for measurement. Plucker et al. (2004) make recommendations for best creativity research practice, based on their literature survey:

> 'we argue that creativity researchers must
> (a) explicitly define what they mean by creativity,
> (b) avoid using scores of creativity measures as the sole definition of creativity (e.g., creativity is what creativity tests measure and creativity tests measure creativity, therefore we will use a score on a creativity test as our outcome variable),
> (c) discuss how the definition they are using is similar to or different from other definitions, and
> (d) address the question of creativity for whom and in what context.' (Plucker et al., 2004, p. 92)

In short, the standards used to judge creativity need to be specified and justified. A more objective and specified definition of creativity enables researchers to make a worthwhile contribution (Torrance, 1967; Plucker et al., 2004; Kaufman, 2009). This is as relevant for computational creativity as it is for human creativity. Computational creativity researchers must be clear about what is meant by a program being creative, when discussing how creative it is; otherwise how can research progress be tracked? How can the strengths and weaknesses of different research projects be compared and contrasted? (Ritchie, 2001; Pearce et al., 2002; Colton, 2008b; Pease & Colton, 2011b). Looking at human creativity research, Plucker, Beghetto and colleagues stress the need for definition:

> 'Without an agreed-on definition of the construct, creativity's potential contributions to psychology and education will remain limited.' (Plucker et al., 2004, p. 87)

> 'Explication of a definition of creativity is necessary for the study of creativity to continue to grow, thrive, and contribute meaningfully to our understanding of the processes and outcomes across and within various domains.' (Plucker & Beghetto, 2004, p. 155)

To stress this point further, it is argued that the lack of a standard definition for creativity research weakens the 'legitimacy' and validity of that research, impeding progress (Rhodes, 1961; Plucker et al., 2004). This concern applies to computational creativity and human creativity research alike.

> 'we have to have a clear notion of what we mean by a program "being creative". Without that, discussion of the pros and cons of different internal mechanisms are just speculation.' (Ritchie, 2001, p. 3)

3.2 Perils and pitfalls in defining creativity

Several problems and issues arise when attempting to define creativity, as shall be seen in this Chapter. Some are specific to computational creativity, others relate to creativity research in general.

3.2.1 A standard definition of creativity?

> 'However you think of creativity - *whether* you think of creativity - it is all likely different from how your neighbor thinks of it. There are many different ways in which someone can be creative, and there are almost as many different ways that people try to measure creativity.' (Kaufman, 2009, p. 9)

No consensus has been reached on a standard, comprehensive definition of creativity, due to a variety of issues involved in understanding and defining what creativity is. A multitude of definitions of creativity have arisen, in various depths of analysis and for different purposes. Whilst there is some agreement as to what creativity entails, definitions diverge according to what is prioritised as most pertinent in their viewpoint of creativity. Identifying a standard definition that fits all manifestations of creativity is non-trivial and is arguably a flawed approach to take (see Section 3.6.4).

3.2.2 High expectations of a weak term

We have an intuitive but tacit understanding of the concept of creativity that we can access introspectively. For comparative purposes and methodical, transparent evaluation, this introspective understanding is not sufficient.[1] The title of von Hentig's 1998 book on creativity neatly sums up the word 'creativity' (von Hentig, 1998): *Kreativität: Hohe Erwartungen an einen schwachen Begriff*, which translates[2] as *'Creativity: a high expectation of a weak term.'*. The point that von Hentig makes is that we expect the word *creativity* to convey a great deal, despite the word itself being ill-defined.

> '[c]reativity defies precise definition ... even if we had a precise conception of creativity, I am certain we would have difficulty putting it into words' (Torrance, 1988, p. 43)
>
> 'it may be possible to use measures of creativity to determine whether a program is being creative at all, but this is also problematic, because creativity is such an overloaded and highly subjective word.' (Colton et al., 2001, p. 1)

Many authors have expressed the difficulties in coining a definition of creativity in words (Rhodes, 1961; Torrance, 1988; Sternberg & Lubart, 1999; Boden, 2004; Ritchie, 2006). In particular, how can such a definition be tested once it is derived? (Boden, 1998; Heilman, 2005; Cardoso et al., 2009; Dietrich & Kanso, 2010). There are competing opinions on what is or is not creative, with no baselines or ground truths to test a definition of creativity against (Plucker et al., 2004; Colton, 2008b; Hennessey & Amabile, 2010; Norton et al., 2010). Often the creative process is not directly observable, even with current neuroscientific advances (Dietrich & Kanso, 2010; Heilman, 2005).

Time and context factors also affect how creativity is defined. Individual and societal definitions of creativity may shift over time as knowledge and practices develop, or cultural norms change (Boden, 1998; Colton, 2008b). The audience which critically receives the creative work plays a crucial role in determining the perceived creativity of that work. Social peer pressure may influence perception of

[1] This will be shown in an evaluation survey performed on systems in Case Study 1 (Chapter 8 Section 8.1.1), where several participants called for creativity to be defined to help them to evaluate the creativity of the Case Study 1 systems.

[2] Thanks to Jens Streck for making me aware of this book and for translating the title from German to English.

creativity, according to social groupings, reputation, and on a larger scale according to cultural standards. If creativity really 'is in the eye of the beholder' (Cardoso et al., 2009, p. 17)/(Widmer et al., 2009, p. 44), then if someone or something is creative and this is not recognised, then it is unclear whether that person or work is still creative, or if it only becomes creative when recognised. Boden (2004) and Colton (2008b) question whether perceived creativity is reflective of actual creativity, should such a distinction exist.

Creativity has a multi-faceted nature, being made up of many different factors and components (Sternberg, 1988; Torrance, 1988; Sternberg & Lubart, 1999; Boden, 2004; Plucker et al., 2004; Dietrich & Kanso, 2010). The multi-dimensionality of the concept of creativity, discussed in more detail in Section 3.4.2, causes problems; how does one know that all aspects have been included and that the definition is both inclusive and accurate (Plucker et al., 2004)? Also, which components are most important? There may be no answer to these questions. People see creativity from their own perspective (detailed further in Section 3.6), within a set of biases and assumptions which they may not even be aware of (Hennessey & Amabile, 2010; Moffat & Kelly, 2006):

> 'In essence, all of these researchers may be discussing completely different topics, or at least very different perspectives of creativity. This is not merely a case of comparing apples and oranges: We believe that this lack of focus is tantamount to comparing apples, oranges, onions, and asparagus and calling them all fruit. Even if you describe the onion very well, it is still not a fruit, and your description has little bearing on our efforts to describe the apple.' (Plucker et al., 2004, pp. 88-89)

3.2.3 Perceptions of mystery and wonder in creativity

> 'It's hard to claim that something is scientific or that it can be measured when the first people to talk about it were also talking about muses and demons.' (Kaufman, 2009, p. 2)

Creativity is often said to have an undefinable or mysterious side to it, that we cannot identify. This impression may originate from times when creativity was thought of as mystical, relying on the presence of a 'muse' or divine inspiration (Williams, 1976; Albert & Runco, 1999; Williams, 1967; Boden, 2004; Kaufman, 2009). The word 'creativity' as we use it today is etymologically derived from the same sources as a 'Creator' or 'Creation' (Williams, 1976), indicating a godlike quality that may be inaccessible to mere mortals (and computers). In the current print edition of the Oxford English Dictionary (Simpson & Weiner, 1989), the definition of the word 'create' begins with:

> 'Said of the divine agent: To bring into being, cause to exist' (Simpson & Weiner, 1989, p. 1134).

An objection to computational creativity, and more widely to the study of creativity in general, is whether it may have a detrimental effect on our sense of the wonder and mysticism of creativity:

> 'Forget computers, for the moment: the conviction is that *any* scientific account of creativity would lessen it irredeemably. ... [There is a] widespread feeling that science, in general, drives out wonder. Wonder is intimately connected with creativity. All creative ideas, by definition, are valued in some way. Many make us gasp with awe and delight. ... To stop us marvelling at the creativity of Bach, Newton, or Shakespeare would be almost as bad as denying it altogether. Many

> people, then, regard the scientific understanding of creativity more as a threat than a promise.' (Boden, 2004, pp. 277-278)

Rhodes argues that creativity is not something mysterious (Rhodes, 1961):

> [Recent creativity research has] 'rendered into baloney many former sacred cows. For instance ... the idea that people are born to be either creative or lacking in creative ability, the notion that creativity is more a way of feeling than a way of thinking, the idea that creativity is something mysterious, and the notion that the word creativity applies to a simple, uncomplicated mental process that operates in restraint.' (Rhodes, 1961, p. 306)

This undefinable part of creativity is reflected in Weiley's coining of creativity as 'novelty, value and "x" ' (Weiley, 2009). It is often inadvertently included in discussions about creativity, for example O'Donoghue (2007) qualifies a list of components of creativity with the final component of 'e.t.c.' (O'Donoghue, 2007, p. 34) rather than providing a complete standalone list of components.

Several researchers have however agreed with the sentiments expressed above in Section 3.1, that research into creativity helps us understand creativity better and develop our own creativity (Boden, 2004; Plucker et al., 2004; Kaufman, 2009; Hennessey & Amabile, 2010). Creativity is generally considered a positive quality to possess[3] and to know how to develop one's creativity is considered desirable and beneficial, as illustrated by the wide availability of materials for this purpose.[4]

3.2.4 Overriding definitional problems

Veale et al. (2006) prefaces the *New Generation Computing* 2006 special issue on computational creativity by outlining three different ways in which authors in this issue respond to the difficulties of defining creativity (Veale et al., 2006, pp. 204-205):

1. Work without a definition of creativity or with an approximate definition, in the hope that practical work will point towards a clearer definition.
2. Identify how metaphors are used in creativity and what existing computational heuristics and models best represent such metaphors.[5]
3. Produce systems that model human examples of creative work; replicating the results of creativity rather than taking a fuller perspective.[6]

[3] Notwithstanding the acknowledgement of a 'dark side' of creativity (Kaufman, 2009), where creativity can be manifested in a negative way. For example one could argue that the actions of Hitler or the 'September 11' bombers in New York were creative, although the atrocious consequences of their actions would provoke negative reactions to this conclusion vastly in excess of any distaste towards computational creativity.

[4] The results of a questionnaire on creativity in musical improvisation, which will be reported in Chapter 6 Section 6.3.2, will also show how creativity is viewed as positive, with all responses treating creativity as a desirable aspect of improvisation. Responses could be summed up by this comment from one participant: 'I wish I had more of it, definitely'.

[5] A little bias is perhaps introduced here by Tony Veale, one of the authors of Veale et al. (2006), highlighting the approach he favours above other approaches to modelling the creative process such as those outlined in Section 3.4.1.

[6] In *four P* terminology, described in Section 3.4.2, this focuses on Products at the expense of Person, Process and Press.

In contrast, Rhodes (1961) 'appeals' that 'we do not throw out the baby with the bath water just because the water is cloudy.' (Rhodes, 1961, p. 310). Though it is problematic to define creativity, this does not mean that we should avoid such a task. The next Sections will review several such attempts to define creativity, categorised as:

- Dictionary definitions of creativity (Section 3.3).
- Research definitions of creativity (Section 3.4):
  - Definitions of computational creativity (Section 3.4.1).
  - Definitions of human creativity (Section 3.4.2).
- Philosophical reflections on how to interpret the meaning of concepts like creativity (Section 3.4.3).
- Legal definitions of creativity (Section 3.5).

3.3 Dictionary definitions of creativity

Dictionary definitions give a short and succinct account of word meanings. Figure 3.2 shows several dictionary definitions of the word 'creativity' and the related words 'creative' and 'creat'.[7] The most frequent words used in these definitions (excluding common-use English words such as 'the', 'and', and so on) are represented in a word cloud format, in Figure 3.3. This diagram shows the prominence of a number of concepts in how creativity is defined, such as:[8]

- cause
- something
- produce
- make
- new
- original
- imaginative
- existence
- ability
- quality
- rank
- occasion

The definitions in Figure 3.2 tend to be self-referential, defined in terms of words derived from the same word stem 'creat-'.[9] These definitions often contain (or on occasion entirely consist of) phrases such as the 'ability to create', or 'creative power or faculty'. Essentially, creativity is defined in terms of itself. This does not help develop an in-depth practical understanding of creativity.

The brevity of the definitions in Figure 3.2 also affects the utility of dictionary-style definitions for evaluation. For use as evaluation standards, definitions should be of sufficient length and detail, or they provide only an overview rather than investigating more subtle points which may be significant. Dictionary definitions however tend to be short and heavily abbreviated, due to space restrictions.

Dictionary definitions assume words can be defined independently of context. Some of the definitions in Figure 3.2 show that context is important when considering the meaning of a word, making

[7] Definitions are taken from various print dictionaries, as indicated (Simpson & Weiner, 1989; Treffry, 1998, 2000; Robinson, 1999; Garmonsway & Simpson, 1969; Babcock Gove, 1969; Barnhart, 1963; Sinclair, 1992).

[8] The selection of words listed here ignores the words 'creativity', 'creative' and 'create'.

[9] The problem of self-reference in definitions of creativity has also been noted by Cope (2005).

Figure 3.2: Dictionary definitions. N.B. Definitions of 'creativity' are often included under definitions of 'creative' or 'create'. For readability, some definitions are edited slightly to standardise formats and remove etymological/grammatical annotations.

Figure 3.3: Word cloud representation of the words used in the dictionary definitions of Figure 3.2. Generated by *Wordle* online software (*http://www.wordle.net/*).

some interpretations more relevant than others. For example, in understanding what creativity is, one use of 'create' that is less relevant is its use to refer to an actor playing a new theatrical role. Instead of using words independently of context, we tend to use words to refer to things within a contextual frame of reference (Evans & Green, 2006; Haiman, 1980). Cognitive linguistics studies language as it is *used*, rather than relying on pre-determined meanings. Cognitive semanticists, who research the meaning of words and phrases using cognitive linguistics, follow a central principle:

> 'word meaning cannot be understood independently of the vast system of encyclopaedic knowledge to which it is linked' (Evans & Green, 2006, p. 243)

To construct a mental representation of the meaning of the word involves a dynamic process of *conceptualisation* (applying concepts to background knowledge). The point is that words cannot be used out of context, therefore they have no meaning or relevance that is independent of context. Semantics (what words mean) should not be separated from pragmatics (how words are used) (Haiman, 1980). Instead, word meaning is grounded in our experiences and knowledge. The words themselves act as reference points to relevant contextual knowledge, rather than 'neat pre-packaged bundles of information' (Evans & Green, 2006, p. 221).

3.4 Research definitions of creativity

3.4.1 Defining computational creativity

As the interest of this thesis is in computational creativity research, it is important to examine how computational creativity has been defined by researchers to date.

Computational creativity = human creativity performed by computers?

A prevalent definition of computational creativity centres around human creativity being demonstrated by computer systems: computational creativity is where computers exhibit behaviour which, if seen in a human, would be perceived as creative (Colton, 2008b; Wiggins, 2006a; Cardoso & Wiggins, 2007). Whilst such a definition is intuitively simple, it reveals little about what creativity actually is; nullifying what is an aim of much computational creativity research (Widmer et al., 2009).

In computational creativity research, this type of definition has often been adopted (Schmid, 1996; Wiggins, 2006a; Cardoso & Wiggins, 2007; Colton, 2008b; Cardoso et al., 2009). To some extent it simplifies matters to reproduce the only example we have of creativity, i.e. human creativity. If 'creativity really is in the eye of the beholder' (Cardoso et al., 2009, p. 17), though, then there are significant repercussions for the scientific validity of evaluating computational creativity systems. A researcher could report their system as the most creative and progressive system of its type, on the basis that they themselves view it that way, without further justification. Another researcher might counter this with a similar claim about their own system, for the same justification. This would quickly lead to stagnation in research (Pearce et al., 2002) rather than progress.

Alternative definitions of computational creativity

Alongside the prevailing current definition of computational creativity (to refer to human creativity in a computational setting without defining creativity itself), there exist complementary perspectives on computational creativity which incorporate creativity theory in some way.

Reduction of creativity to novelty + value Peinado and Gervas define computational creativity as 'how to create something new and useful at the same time.' (Peinado & Gervas, 2006, p. 290). The *novelty* (related concepts: originality, newness) and *value* (related concepts: usefulness, appropriateness, relevance) of creative products have often been identified as the two main aspects of creativity (Pease et al., 2001; Peinado & Gervas, 2006; Pereira & Cardoso, 2006; Ritchie, 2007; Alvarado Lopez & Pérez y Pérez, 2008; Brown, 2009a; Chordia & Rae, 2010).[10] Here computational creativity borrows from psychological research into creativity (e.g. Finke et al., 1992; Mayer, 1999; Sternberg, 1999; Boden, 2004; Plucker & Beghetto, 2004; Kaufman, 2009; Rowlands, 2011). Mayer (1999) refers to this combination as the 'basic definition of creativity' (Mayer, 1999, p. 450). Table 22.1 of Mayer (1999), reproduced here in Figure 3.1, summarises the 'Two Defining Features of Creativity' (Mayer, 1999, p. 450) as used in Sternberg (1999).

[10] As seen in Chapter 2 Section 2.3.4, though, inconsistencies have arisen in how this model is adopted.

Table 3.1: Mayer's summary of how novelty and value (or highly related concepts) are used to define creativity by different authors in various chapters of Robert J. Sternberg's influential *Handbook of Creativity* (1999) (Mayer, 1999, Table 22.1, p. 450).

Author (Chapter)	Feature 1: Originality	Feature 2: Usefulness
Gruber & Wallace (5)	novelty	value
Martindale (7)	original	appropriate
Lumsden (8)	new	significant
Feist (13)	novel	adaptive
Lubart (16)	novel	appropriate
Boden (17)	novel	valuable
Nickerson (19)	novelty	utility

In their survey of creativity definitions usage in research, Plucker et al. (2004) found that of the 34 out of 90 surveyed articles that included a definition:

> 'The most common characteristics of explicit definitions were uniqueness (n = 24) and usefulness (n = 17). Of interest, all 17 articles that included usefulness in their definition also mentioned uniqueness or novelty.' (Plucker et al., 2004, p. 88)

Novelty is a continuous attribute rather than discrete, being measurable by 'degree of newness' (Rhodes, 1961, p. 309) rather than something being either new or not new (Perkins, 1994; Rhodes, 1961; Kaufman, 2003; Pease et al., 2001; Potts, 1944). Questions arise as to what novelty entails in a creative context (Boden, 2004; Bundy, 1994; Macedo & Cardoso, 2002; Peinado et al., 2010). 'How novel should "novel" be for creativity?' (Perkins, 1994, p. 120).

Another issue is that it is difficult to find domain-independent heuristics to follow when ascertaining the value of products. Usefulness is relative; what is considered useful in products of one domain is not necessarily reproduced in the other and may not apply equally across that individual domain. To recognise the usefulness of a creative product, one must either know the product's domain well enough to appreciate value, or have access to the opinions of people who are experts in that domain.

To exemplify this point, in Jordanous and Keller (2011) we report how the greatest contributor to creativity in musical improvisation is the social communication and interaction that happens between musicians, or between performer(s) and audience during the creative process of improvising.[11] For creativity, improvisers prioritise this over the 'correctness' of the music produced during improvisation. In mathematical proof derivation systems, however, accuracy is paramount.

Zongker's paper entitled *Chicken Chicken Chicken: Chicken Chicken* and published via the *Annals of Improbable Research* (Zongker, 2006) demonstrates how the perception of creativity in a particular domain is not always consistent across all examples of creativity in a domain. *Chicken Chicken*

[11] These findings are reported in greater detail in Chapter 6 for Case Study 1.

Chicken shows creativity, in a domain that emphasises content correctness and usefulness (scientific research papers), because of the extreme absence of any scientifically useful and correct content. Instead the value of Zongker (2006) is as an ironic reflection on academic publications.

As discussed in Section 3.4.1, it is questionable whether the combination of novelty and value is enough to understand creativity (Kaufman, 2009). This reductionist approach provides two tangible attributes with which to evaluate creativity. It is questionable, though, whether creativity can be reduced to just these two components. Other creativity definitions reported in this Chapter incorporate more than merely the novelty and value of products.[12]

Another relevant point, as Section 3.4.2 will emphasise, is that creativity entails more than solely the end product; a creative person or system should not be judged based only on how novel and useful its products are. As early as 1963, Newell, Shaw, and Simon describe four criteria for creativity (Newell et al., 1963), of which the combination of novelty and usefulness is only one. The other three include transformation of views previously accepted by the creative person's peers, motivation and persistence in the creative process and coping with uncertainty during the process.[13]

Creativity as the search for solutions to problems Creativity has often been portrayed as finding solutions to problems, both in computational creativity research (Newell et al., 1963; McCarthy, 1979; Meehan, 1981; Turner, 1994; Wiggins, 2006b; Jennings, 2010b; Ventura, 2011; Veale, 2012, also Pérez y Pérez, 2012, personal communications) and research into human creativity (Weisberg, 1988; Perkins, 1994). Newell et al. (1963) differentiate between creative problem solving and problem solving if the solution is novel and useful, and the original problem was specified in an uncertain manner which required unconventional and persistent and interpretative effort. Cardoso et al. (2009) question Newell et al. (1963)'s approach:

> 'This approach is, of course, a classic AI formulation: there is at its base the implicit assumption that the created artefact is an "answer", and therefore that there must have been a question. In the creative arts and in the less empirically motivated sciences and mathematics, this need not be the case: creative motivation may be altogether less well-defined.' (Cardoso et al., 2009, p. 2)

Reducing creativity to problem solving works when the creator is searching for an ideal solution which is not obvious, or if there is no single ideal solution but several candidates for a reasonable solution (Boden, 1994b). This interpretation of creativity rules out the possibility of being creative if no problem currently exists, though. Creativity must also allow for the idea of being creative to discover what can be done, using what Boden refers to as 'exploratory creativity' and Mandel describes as 'problem-finding creativity' (Boden, 2004; Mandel, 2011).

[12] The work in Chapter 4 will present evidence that there is much more to consider in terms of what creativity is, that the combination of novelty and value does not incorporate.

[13] These four criteria align to some degree with the results in Chapter 4, although those results include other criteria not mentioned by Newell et al. such as domain competence and autonomous freedom on the part of the creative person.

Boden's framework for creativity As discussed in Section 3.6, Boden (2004) identifies three different types of creativity:

- Exploratory creativity - exploring a scope-limited set of possible options within a domain (Boden uses the term *conceptual space* for this set.
- Transformational creativity - transforming the conceptual space (set of options) by identifying where the boundaries of the set can be changed to include new options.
- Combinational creativity - combining two or more concepts within the conceptual space to form a new concept.

Though exploratory creativity seems less important than transformational creativity, exploring a new conceptual space helps to establish the value of the space, by seeing what potential it has for development; a new space is more useful if it has greater potential for exploratory creativity.

With transformational creativity, to reject rules, the rules have to be understood (perhaps via prior explorative creativity). The key point here is that it is not enough to just not be aware of rules or constraints, but that there has to be some intention to transform them. Transformation is usually by dropping constraint(s) and negating constraint(s).

Combinational creativity is inspired somewhat by the associative view of creativity in Koestler (1964), combining concepts together in a Gestalt manner such that the resulting combination provides more of value than the two individual components did on their own. Often part of this creative act is in seeing components that can be combined and how they can be combined.

One potential problem with Boden's three views of creativity is that they all assume the existence of a conceptual space, or constrained set of possibilities, that the creative individual consciously reasons with in order to be creative. Boden's framework therefore does not easily accommodate subconscious processes of creativity, or systems which employ principles of embodiment and emergence (such as McCormack, 2007; Bown & Wiggins, 2007; Bird & Stokes, 2007; McLean, 2009).

The implementation of creativity theory in practical computational creativity systems is currently limited, with most researchers preferring a more direct and hands-on approach. There are however some examples of Boden's three-fold framework being applied as a framework rather than a qualitative discussion of differences in creative practice (Thornton, 2007; Wiggins, 2006b). In particular, Wiggins proposes the Creative Systems Framework (Wiggins, 2006b, 2006a; McLean & Wiggins, 2010; Forth, Wiggins, & McLean, 2010) for this purpose. Ritchie (2006) has focused on and developed transformational creativity.

Domain Individual Field Interaction (DIFI) framework Csikszentmihalyi proposes a systems model of creativity. Figure 3.4, a reproduction of Csikszentmihalyi's Figure 13.1 (Csikszentmihalyi, 1988, p. 329), shows how Csikszentmihalyi sees the creative process unfold over time. Time is a key element in this model, highlighting how creativity is a dynamic process that progresses and develops

as the result of interactions between the three subsystems of Domain, Field and Individual/Person:

> 'what we call creative is never the result of individual action alone; it is the product of three main shaping forces: a set of social institutions, or *field*, that selects from the variations produced by individuals those that are worth preserving; a stable cultural *domain* that will preserve and transmit the selected new ideas or forms to the following generations; and finally the *individual*, who brings about some change in the domain, a change that the field will consider to be creative.' (Csikszentmihalyi, 1988, p. 325).

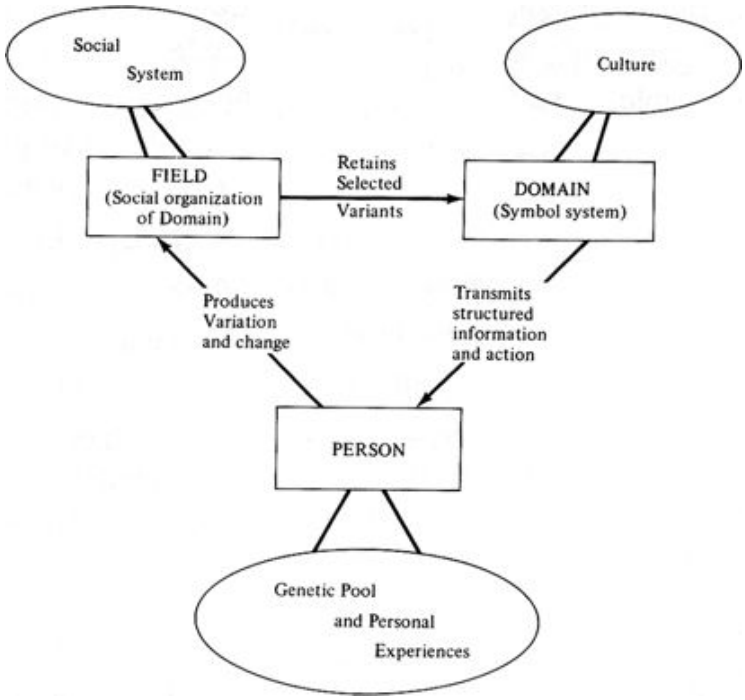


Figure 3.4: Csikszentmihalyi's DIFI framework for creativity: mapping interactions between *Domain*, *Field* and *Person* (Csikszentmihalyi, 1988, Fig. 13.1, p. 329).

For a better understanding of creativity, Csikszentmihalyi argues for a Kuhnian-style scientific revolution in creativity research (Kuhn, 1962), saying that 'we need to abandon the Ptolemaic view of creativity, in which the person is at the center of everything, for a more Copernican model in which the person is part of a system of mutual influences and information.' (Csikszentmihalyi, 1988, p. 336).[14] As a consequence, we cannot determine what is creative (or not) purely by referring to an individual person or their products but must consider people and their products or ideas in the context of the appropriate domain and its field.

The DIFI framework has recently impacted upon computational creativity research (Chordia & Rae, 2010; de Silva Garza & Gero, 2010; Saunders et al., 2010). It is a rare example of a creativity

[14] Kuhn's contributions to scientific method, as well as those of others, have been discussed in Chapter 2 Section 2.2.

framework which has been implemented on more than one occasion as a basis for practical systems, which pays attention to interactions between the creator, the body of knowledge the creator is working in and the audience which critically receives the creative work and provides feedback. There is some information about how these interactions take place, though an examination of Figure 3.4 raises some practical questions, for example: how does the creative individual 'produce Variation and change'? How does the field decide which 'Variants' of creative contributions should be retained and added to domain knowledge?

Reframing creativity by defining something 'creativity-like' Following in the steps of recasting creativity as problem-solving, or a combination of novelty and value, this approach deals with the definitional problem by identifying something similar to creativity that is easier to define. Cohen (1999) defines *behavior X* as the combination of emergence, awareness, motivation and knowledge.[15]

Cohen considers his generative art system AARON in the context of behavior X but refrains from describing AARON as creative (McCorduck, 1991; Cohen, 1999; Pearce & Wiggins, 2001; Boden, 2004; Bird & Stokes, 2007; Colton, 2008b). This restraint is largely because AARON lacks the ability to make aesthetic choices about the quality of produced work. The implication (left implicit by Cohen) is that behavior X and autonomy are two necessary conditions for creativity, though it is unclear whether Cohen sees this combination as sufficient for creativity.

Pease et al. suggest redefining creativity as *creativity2* (Pease et al., 2001, p. 137) so that it is more tightly defined and therefore easier to assess (c.f. Popper's falsifiable hypotheses, Popper, 1972). Creativity2 is however left undefined by Pease et al.

Is redefinition the right approach? Rhodes's emphasis on how we should 'not throw out the baby with the bath water just because the water is cloudy.' (Rhodes, 1961, p. 310) is relevant here. The difficulties in identifying creativity should not lead us to redefine creativity; otherwise we are no longer researching creativity but creativity2 (Pease et al., 2001) or behavior X (Cohen, 1999). This will place limits on what can be achieved, rather than contributing to the bigger picture that is creativity.

The role of evaluation within the creative process As Boden (2004) says, creativity is not just about new ideas but also incorporates the development and refinement of these ideas.[16] Several computational creativity researchers employ in their systems a sequence of *engagement-reflection*, or *generate-and-explore*, where the system takes a cyclic process of engaging with creative production then critically reflecting on what it has produced, to inform the next stage of creative production (Finke et al., 1992; McGraw & Hofstadter, 1993; Gervás et al., 2007; Pérez y Pérez, 1999; Alvarado Lopez & Pérez y Pérez, 2008; Pease et al., 2010; Pérez y Pérez et al., 2010). Others use evolutionary computational approaches, where the fitness of the system is evaluated and evolved (Biles, 1994; Bird & Stokes, 2007; Miranda & Biles, 2007; McCormack, 2007; Jordanous, 2010c).

[15] As previously mentioned in Chapter 2 Section 2.1.4.

[16] Such development/refinement is analogous to Wallas (1945) and Poincaré (1929)'s *validation*; see Section 3.4.2.

Whilst not adopted universally across the computational creativity research community, and whilst there is still some ambiguity as to exactly what the *engagement* and *reflection* stages should entail, this approach satisfies the call of Boden (1999) to treat evaluation as an important part in the creative process itself, so the creative system can recognise the best of the artefacts it generates.

An approach that incorporates evaluation in the creative process faces the same issues and questions as this thesis does; how does one evaluate creativity? To date, the systems mentioned above generally sidestep this issue somewhat, evaluating the *quality* of their output rather than *creativity*.[17] Hence they progress towards more successful output, but not necessarily more creative output.

3.4.2 Defining human creativity

The most widely accepted definition of computational creativity (Section 3.4.1) is a computational system which is perceived to act creatively if treated as a human. This moves the focus from the question *what is computational creativity* to the question *what is creativity*.

There is no shortage of answers offered to this latter question in the research literature on human creativity; a 1988 review mentions over 60 definitions of creativity in the psychology literature (Taylor, 1988) and Plucker et al. (2004) identifies several post-1988 definitional offerings. Here the main definitional contributions of human creativity research are examined, with particular interest in how they relate to computational creativity as well as human creativity.

The Four Ps

One major approach in creativity research is to break down creativity into four perspectives, commonly referred to as the *Four Ps* (Rhodes, 1961; Stein, 1963; Mooney, 1963; MacKinnon, 1970; Simonton, 1988; Tardif & Sternberg, 1988; Odena & Welch, 2009):

- Person: The individual that is creative.
- Process: What the creative individual does to be creative.
- Product: What is produced as a result of the creative process.
- Press: The environment in which the creativity is situated.

Rhodes (1961) was perhaps first to identify the four P perspectives. Rhodes collected 40 definitions of creativity and 16 definitions of imagination. The '*Four P*' dimensions of creativity emerged from analysis of these definitions.[18] Several people seem to have independently identified four similar themes of creativity (MacKinnon, 1970; Stein, 1963; Mooney, 1963; Odena & Welch, 2009), or at least three (Jackson & Messick, 1967, who consider Person, Product and the Response from Press). This repeated pattern of seemingly independent discovery boosts the credibility of the Four Ps.

Simonton (1988) sees discrepancies between combining the Four Ps in theory and in practice:

[17] The distinction between the evaluative aims of creativity and quality was raised in Chapter 2, particularly Section 2.3.

[18] As Rhodes published this work in a relatively unknown journal, many later advocates of a 'Four Ps'-style approach to creativity seem unaware of Rhodes's contribution (e.g. Odena, 2009, personal communications), so do not cite him.

> 'Now, in an ideal state of affairs, it should not matter which one of the four p's our investigations target, for they all will converge on the same underlying phenomenon. ... But reality is not so simple, needless to say. The creative process need not arrive at a creative product, nor must all creative products ensue from the same process or personality type; and others may ignore the process, discredit the product, or reject the personality when making attributions about creativity.' (Simonton, 1988, p. 387)

From this, one conclusion which seems to follow naturally is that an accurate and comprehensive definition of creativity must account for the (potential) presence of all four aspects, in order to be complete. Simonton, however, concludes that '[i]f we cannot assume that all four aspects cohesively hang together, then it may be best to select one single definition and subordinate the others to that orientation' (Simonton, 1988, p. 387), with his natural research inclination leading him to focus his work on *persuasion*, his term for the Press/Environment aspect.

Plucker et al. (2004) conducted a literature survey investigating the use (or absence) of creativity definitions in creativity research. While the key purpose of this analysis is to see how rigorously people define creativity as they work with it, Plucker et al. also use their analysis to derive their own definition by identifying reoccurring themes and forming these into an inclusive definition which happens to account for each of the Four Ps:

> 'Creativity is the interaction among *aptitude, process, and environment* by which an individual or group produces a *perceptible product* that is both *novel and useful* as defined within a *social context*' (Plucker et al., 2004, p. 90)

In reviewing Four Ps research, Kaufman (2009) mentions addendums that have been suggested for the Four Ps: persuasion (Simonton, 1988) and potential (Runco, 2003). In general, however, the Four Ps approach have been adopted as they were originally conceived by various researchers (Rhodes, 1961; Stein, 1963; Mooney, 1963; MacKinnon, 1970).

The Four Ps: Person This perspective addresses human characteristics associated with creative individuals or groups of people. Encouraged by Guilford's call in 1950 for studying the creative person, an abundance of different personal characteristics have been associated with creativity (Rhodes, 1961; Stein, 1963; Koestler, 1964; Tardif & Sternberg, 1988; Odena & Welch, 2009), ranging from personality traits, attitudes, intelligence and temperament to habits and behaviours (for example curiosity, persistence, independence and openness). Some of these are closely related; others are contradictory.

This discrepancy and sheer quantity of attributes together place an obstacle in the way of compiling a definitive list of attributes of a creative person and instead provoke disagreements on exactly which cognitive characteristics should be attributed to creative people. Tardif and Sternberg's review shows that as of 1988, different authors highlight a variety of characteristics, with no general consensus and no characteristics common to all reports (Tardif & Sternberg, 1988, Table 17.1, p. 434).

The Four Ps: Process The creative process has been broken down into a series of sequential or cyclic stages occurring over time (Poincaré, 1929; Wallas, 1945) or subtasks (Odena & Welch, 2009).

Section 3.4.2 gives details of different models of the creative process. In their work on enhancing pupil creativity in school music lessons, Odena and Welch (2009) have broken down the creative process into subtasks, by identifying different types of process (such as different activities, group process, the structuredness or otherwise of a process and composition by improvisation) rather than tracing a linear progression of subprocesses.

Regarding the creative process, Tardif and Sternberg consider how creativity is not just the first flash of inspiration, but is also the activity that validates, develops, and refines that first idea.[19] Rather than occurring at one point in time, say Tardif and Sternberg, creativity occurs over a developmental period of time. Tardif and Sternberg (1988) question whether creativity is a social or an individual process. This point has often been taken forward, most notably by Csikszentmihalyi (1988).

The Four Ps: Product Many authors advocate that *proof* of creativity is necessary to be considered creative (Kagan, 1967; Tardif & Sternberg, 1988; Plucker et al., 2004; Ritchie, 2001).

> 'When an idea becomes embodied in tangible form it is called a *product*. ... Products are artifacts of thoughts.' (Rhodes, 1961, p. 309)

The product-centric view adopted by computational creativity researchers (e.g. Ritchie, 2001), that creative products are both necessary and sufficient for creativity, was present in earlier human creativity research (Kagan, 1967). Influenced by Guilford's seminal 1950 address on creativity research (see Section 3.4.2), emphasis in human creativity research shifted from identifying creative individuals post-production of creative work, to predicting future potential for creativity in individuals. This change in emphasis is demonstrated in the proliferation of psychometric tests for creativity, as reported in Section 3.4.2.

Tardif and Sternberg (1988) consider the creative product more briefly than the other three 'Ps' in their review. They decide that while a creative product is essential for creativity, it is not enough merely to generate a product; the product should also be considered in a domain-specific context.

The Four Ps: Press/Environment The Press perspective encompasses a bidirectional perspective between the environment which influences the creator and receives the creative work, and the creator who publicises their work and is given feedback on what they produce. Tardif and Sternberg (1988) consider both creative domains themselves and the social environments in which creative people are influenced as they employ creative process, advertise their creative products and receive feedback. Rhodes (1961) concentrates on the role that the environment plays on a person during the creative process, rather than how the creative produce is judged by the external world after being created. He reflects on how everyone is different, so everyone perceives the world in a unique way and processes ideas according to their own internal makeup and experiences.

[19] This is again akin to the *validation* stage in Wallas (1945) and Poincaré (1929); see Section 3.4.2.

Of the Four Ps, this is the perspective that is often neglected when one takes an individualistic view of creativity. In general creativity theorists do however acknowledge the influence of the environment in which creativity is situated (Simonton, 1988; Hennessey & Amabile, 1988). If one concentrates on an individual's creativity, however, the Press perspective is often neglected, even if unintentionally. For example, although stating that '[t]o be appreciated as creative. a work of art or a scientific theory has to be understood in a specific relation to what preceded it' (Boden, 2004, p. 74), Boden's treatment of creativity mainly focuses on different cognitive processes of creativity (see Section 3.4.1), rather than a detailed examination of social or environmental influences. Some computational creativity researchers are starting to highlight the importance of the environment in which a creative system is situated (Sosa, Gero, & Jennings, 2009; Saunders et al., 2010; Pease & Colton, 2011b).

Interaction between the Four Ps The mysterious impression often associated with creativity (see Section 3.2.3) can be explained to some extent when one or more of the Four Ps are not accounted:

> 'Each strand [of the Four Ps] has unique identity academically, but only in unity do the four strands operate functionally. It is this very fact of synthesis that causes fog in talk about creativity and this may be the basis for the semblance of a "cult".' (Rhodes, 1961, p. 307)

Rhodes argues that creativity research should follow a very definite path, in a specific direction: 'from product to person and thence to process and to press.' (Rhodes, 1961, p. 309)

> 'Objective investigation into the nature of the creative process can proceed in only one direction, i.e. from product to person and thence to process and to press.' (Rhodes, 1961, p. 309)

Such a statement makes Rhodes's contribution less useful. For example, the Press (environment) in which one is creative has some influence on the creative Process, so one may prefer to study how Press and Person interact before looking at Process issues. For example, Simonton sees creativity as how a person's ideas emerge as influential when that person, by chance, has new ideas and promotes them to influence others. Creative people would not be equivalent to lucky people, by this interpretation, but chance would intervene in their success. Simonton refers to this as the 'chance-configuration theory' that 'outlines the general conditions that favor creativity' (Simonton, 1988, p. 422).

Tardif and Sternberg (1988) discuss each of the Four Ps individually, 'as these really are separate levels of analysis, and it is from comparisons within levels that coherent statements about our knowledge of creativity can be made' (Tardif & Sternberg, 1988, p. 429). Tardif and Sternberg's summary is weakened somewhat by this as it does not make comparisons across the Four Ps, despite highlighting Simonton's emphasis on the interactions and relations between these four views (Simonton, 1988).

Mooney (1963) argues that the four approaches should be integrated in a model of creativity, proposing a model that 'puts together the four approaches by showing them to be aspects of one unifying idea' (Mooney, 1963, p. 333). This model, though, seems to be of life itself, situating 'man' in the 'universe', 'belonging to the whole' whilst 'being integrative to the whole' rather than specifically of creativity. It may be rather too large to be practically useful for modelling creativity.

In his more specific reflections on creativity, though, Mooney's contribution matches neatly with the four Ps approach identified elsewhere at that date (Rhodes, 1961; Stein, 1963)

Confluence Approach

> 'With regard to the confluence of components, creativity is hypothesised to involve more than a simple sum of a person's attained level of functioning on each component.' (Sternberg & Lubart, 1999, p. 11)

This approach follows on from the concept of interaction between the four Ps. The *confluence approach* to creativity works on the principle of creativity resulting from several components converging and looks at what these components are (Sternberg & Lubart, 1999; Mayer, 1999; Ivcevic, 2009). Mayer (1999) describes this as a quantitative approach to creativity. The collection of components may be either a simple or more complex combinations. There may be thresholds for components, above which creativity is said to be present. Components can interact and influence each other, such that a strong component may compensate for a less well represented one. They may have a multiplicative effect, depending on how the confluence approach is defined by a particular author.

As an example, in *investment theory* (Sternberg & Lubart, 1999) creativity is seen as the ability and motivation to trade in ideas; ideas are bought 'low' and sold 'high' (Sternberg & Lubart, 1999, p. 10). Investment theory posits a confluence model consisting of six components:[20]

1. Intellectual ability.
    - Ability to 'think outside the box' (escape conventional thinking).
    - Analytical ability (to recognise worth).
    - Marketing ability (to sell ideas to others).
2. Knowledge (enough to be useful, but not to the extent that domain conventions become constraints).
3. Cognitive styles.
    - Choosing a novel approach.
    - Ability to think globally and locally.
4. Personality.
    - Self-efficacy.
    - Motivation especially in the face of problems.
    - Risk-taking.
    - Ambiguity toleration.
5. Motivation.
6. Environment.

[20] Strong parallels can be drawn between the investment theory confluence model and the Four Ps.

Confluence approaches capitalise on the diversity of creativity that hinders other creativity definitions, reducing creativity to its diverse components. Reductionist approaches can help make a complex concept more understandable, without losing the value of the original concept. Dennett (1995) argues, in a similar discussion about understanding minds, purposes and meanings in the context of Darwinian evolution:[21]

> 'a *proper* reductionist explanation of these phenomena would leave them still standing but just demystified, unified, placed on more secure foundations. We might learn some surprising or even shocking things about these treasures, but unless our valuing these things was based all along on confusion or mistaken identity, how could increased understanding of them diminish their value in our eyes?' (Dennett, 1995, p. 82)

Psychometric tests for creativity

The question of how to evaluate creativity is problematic, in computational or in human creativity:

> 'many people have a gut reaction about studying and measuring creativity ... the thought of grading Faulkner (or Mozart or Picasso) sounds absurd. The idea of "grading" Einstein or Bill Gates is a little absurd, too, yet we still have an awful lot of science and math questions on the SATs, GREs, AP tests and so on that could theoretically provide a reasonable grade for them. Indeed, most of the things that scientists try to measure don't merit this type of reaction.' (Kaufman, 2009, p. 2)

Psychometric tests, or psychological measurement tests, are designed to measure various human skills and psychological attributes. Several tests and test batteries (collections of tests) exist to measure how creative people are, often treating creativity as closely related to divergent thinking, or the ability to think in a broad and expansive way. Some key examples of psychometric tests, along with their underlying interpretation of creativity, are described below.

Guilford's Structure of Intellect battery Prior to 1950, little research of note was being conducted into creativity. J. P. Guilford's presidential address to the American Psychological Association (Guilford, 1950) thrust a spotlight onto creativity research, or more specifically, the lack of thereof. His talk was hugely influential in encouraging more creativity research (Kaufman, 2009), redirecting the focus of research towards attributes of creative people and how to identify who is creative. As part-illustration of this, Figure 3.5 shows how use of the word 'creativity' in books has increased significantly since 1950, according to statistics generated by the Google Ngram project (Michel, Shen, Aiden, Veres, Gray, Pickett, Hoiberg, Clancy, Norvig, Orwant, Pinker, Nowak, & Aiden, 2011).

Concentrating on divergent thinking in creativity, Guilford (1950) proposed the *Structure of Intellect* battery of tests for creativity based around 4 aspects of divergent thinking:

- Fluency (how many ideas in total a person has in a creative situation).
- Flexibility (how many different types of idea the person has).

[21] A similar approach is taken to the work reported in Chapter 4 to derive the key 'building blocks' of creativity (using computational linguistics techniques to analyse and combine several discussions about what creativity is).

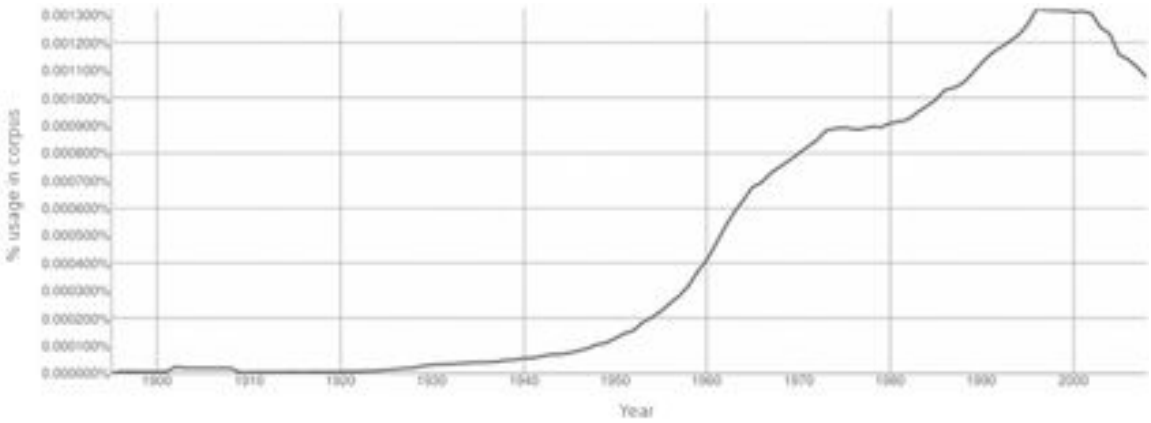


Figure 3.5: Use of the word 'creativity' since 1900 in the 'Google Million' corpus of 1 million books digitised as part of the Google Ngrams project. This image, generated by the Google Books Ngram Viewer at `http://ngrams.googlelabs.com` (last accessed September 2011), shows how use of the word 'creativity' noticeably increased in books published from the mid-1950s onwards.

- Originality (how unusual the person's ideas are).
- Elaboration (how developed the person's ideas are).

Guilford's talk prompted work on several other creativity tests measuring divergent thinking:

- Getzels and Jackson's battery (Getzels & Jackson, 1962).
- Minnesota battery of tests (Goldman, 1964).
- Wallach and Kogan's two test batteries (Wallach & Kogan, 1965).
- Torrance Tests for Creative Thinking (Torrance, 1974).
- Creativity Quotient test (Snyder, Mitchell, Bossomaier, & Pallier, 2004)

Of these, Torrance's test battery is most widely used and studied today (Kaufman, 2009).

Torrance Tests of Creative Thinking Torrance offered three types of creativity definition:

- An *artistic definition*: as depicted in Figure 3.6 (Torrance, 1988).
- A *survival definition*: 'When a person has no learned or practiced [sic] solution to a problem, some degree of creativity is required.' (Torrance, 1988, p. 57).
- An *operational* (Torrance, 1967) or *research definition* (Torrance, 1988):

  > 'the process of becoming sensitive to problems, deficiencies, gaps in knowledge, missing elements, disharmonies, and so on; identifying the difficulty; searching for solutions, making guesses, or formulating hypotheses about the deficiencies; testing and retesting these hypotheses and possibly modifying and retesting them; and finally communicating the results.' (Torrance, 1967, pp. 73-74)

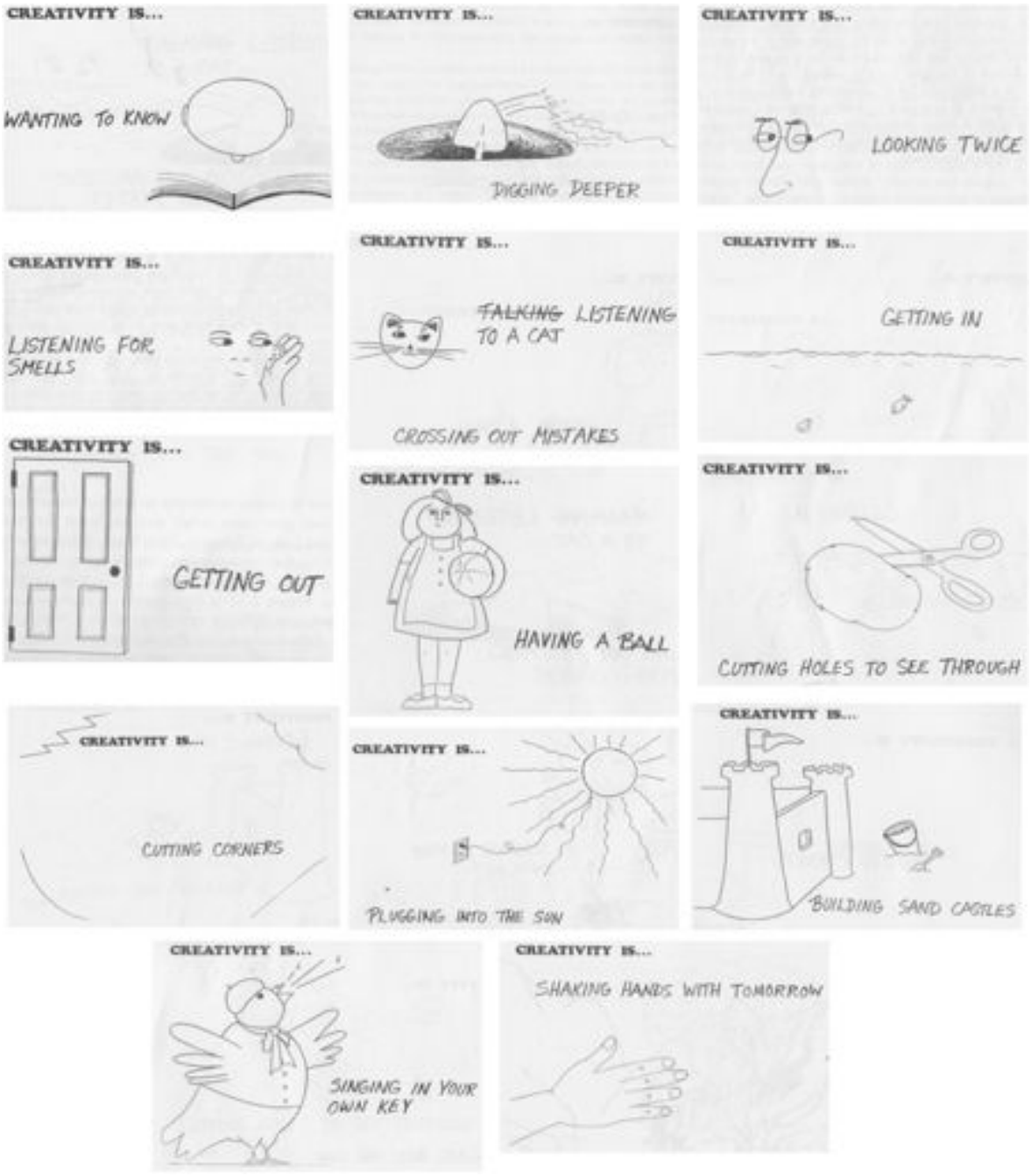


Figure 3.6: Creativity is... 14 pictures originally conceived by Karl Anderson in 1964 and reproduced by Nancy Martin in 1985, as part of E. Paul Torrance's research into what creativity is (Torrance, 1988, Figs. 2.1-2.14, pp. 49-55).

The Torrance Tests of Creative Thinking (TTCT) are based around Torrance's operational definition of creativity, focusing on process. Torrance advocates this definition for two reasons: firstly, it allows further investigation about people who employ these processes, the environments the processes are employed in and what is produced as a result; and secondly, it describes a 'natural process' rather than reporting creativity in a forced manner.

First presented in 1974, Torrance's tests have undergone ongoing development, surviving Torrance himself and becoming an commercial enterprise with copyrighted creativity scoring tables to 'grade' creativity.[22] The TTCT battery is aimed at a wide age range (Torrance, 1967) from toddlers to adults. Influenced by Guilford, the battery tests the same factors of divergent thinking (fluency, flexibility, originality and elaboration) in various tests: *figural* (image-based) and *verbal* (words-based).

Criticisms of divergent-thinking psychometric tests

- The tests look at trivial and inadequate instances of creativity, such as filling in circles with drawings, rather than more meaningful examples of creativity (Sternberg & Lubart, 1999).
- Tests are usually based on the four factors of divergent thinking: fluency, flexibility, originality, elaboration, without firm evidence that divergent thinking fully captures what creativity is. In fact the prominence of divergent thinking in creativity research, as championed by Guilford (1950), has recently been called into question (Baer, 1998; Plucker et al., 2004; Kaufman, 2009; Dietrich & Kanso, 2010). Chapter 4 shows that divergent thinking is only one of 14 different components that are important for creativity.
- The process of scoring responses to tests and analysing results is heavily reliant on the judgment and experience of an individual test administrator.
- The tests assume (without justification) that *everyday creativity*,[23] which they test for, can predict potential future *genius-level creativity* (Kraft, 2005).

From a computational creativity perspective, this last point raises another issue; current computational systems tend to be specialised to a particular domain rather than demonstrating more general creativity in everyday situations. Hence practicalities currently prevent psychometric tests such as those outlined above being used to evaluate the creativity of a computational system.

Other tests for creativity Other tests for creativity in psychology research include:

- Remote Associates Test (Mednick, 1967) (RAT), a psychometric test measuring convergent rather than divergent thinking. Participants are given sets of three words and asked to provide a fourth word that links to all three words semantically.
- Consensual Assessment Test (CAT) (Amabile, 1996), where several experts are asked to rate

[22] This restriction on availability of TTCT material has unfortunately restricted their use in creativity research to some extent (e.g. Kaufman, 2009).

[23] See Section 3.6.1 for further discussion of the distinction between everyday creativity (the concept that everyone can be creative to at least some degree) and creative genius (studying examples of people with exceptional creative achievements).

several aspects of a set of creative products. The various ratings are then combined for an overall comparative judgement of creativity reflecting the general consensus opinion of the experts.[24]

As in computational creativity, no one creativity test is used as standard. Common practice in psychology research tends towards a combination of two or more tests rather than using one test in isolation. For example, Ward, Thompson-Lake, Ely, and Kaminski (2008) uses both Mednick's Remote Associates Test and the Alternate Uses test from Guilford's Structure of Intellect battery.

Creativity as a multi-stage process

Creativity as a process has been subdivided into stages in several different ways, as reported below.

Cognitive stages of creativity Poincaré reflected upon various different stages of his creative cognitive processes during mathematical work (Poincaré, 1929).[25] Later Wallas drew parallels between Poincaré's descriptions and comments from Hermann von Helmholtz, a German physicist of the 19th century (Wallas, 1945). Wallas combined both sets of introspective reflections and recast them as a multi-stage iterative model of the creative process, labelling each stage:

- *preparation* - consciously investigating the problem area and compiling relevant knowledge, without finding a suitable solution.
- *incubation* - exploration of the problem area by the 'subliminal self' (Poincaré, 1929), while conscious thought is concentrated elsewhere, in unrelated areas.
- *illumination* - conscious recognition of a key insight, significant and previously unseen, that points towards a creative solution.
  - *intimation* - an instinctive feeling that inspiration is about to occur. Although *intimation* was originally included as a stage in its own right (Wallas, 1926), Wallas later treated it as part of the *illumination* stage (Wallas, 1945).
- *verification* - conscious exploration and evaluation of the insight in order to produce and refine a creative solution.[26]

A systems perspective to creativity Not unlike Csikszentmihalyi's DIFI framework, described in Section 3.4.1, Hennessey and Amabile (2010) propose a systems model of creativity. Hennessey and Amabile (2010) review creativity literature fairly broadly, comparing different perspectives of

[24]The CAT is one of only a few evaluation methods to cross disciplinary boundaries for use outside psychology research (Wiggins, 2008; Brown, 2010; Masters, 2011).

[25]It is interesting to note that in an alternative translation of Poincaré (1929) (the unattributed 1914 translation quoted by Wallas (1945)), the title of this chapter is translated from 'L'invention mathématique' to 'Mathematical Invention'. In Poincaré (1929), the title is translated to 'Mathematical Creation';'invention' and 'creation' are interchanged.

[26]The *verification* stage is not described by Helmholtz, however Poincaré highlights the need for the *illumination* stage to be followed by a 'second period of conscious work ... in which one verifies the results of this inspiration and deduces their consequences.' (Poincaré, 1929, p. 394).

creativity (mostly within the psychology literature, not taking into account computational creativity but looking at educational and cultural research), and using different levels of analysis.

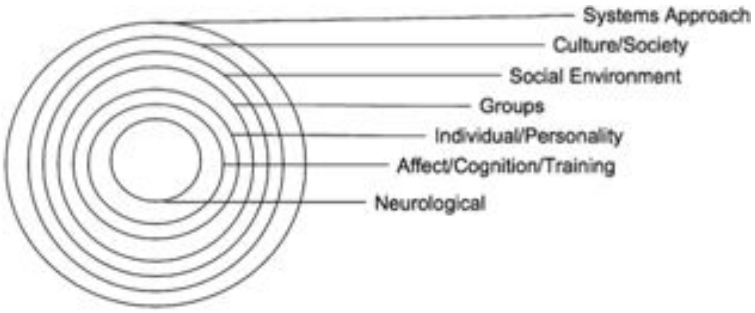


Figure 3.7: Hennessey and Amabile's systems model of creativity, showing different levels of analysis in creativity research (Hennessey and Amabile, 2010, Fig. 1, p. 571).

Their resulting model of creativity is pictured in Figure 3.7, taken from Figure 1 in Hennessey and Amabile (2010, p. 571). Their conclusions criticise the divided nature of creativity research and propose an interdisciplinary 'systems perspective' where 'creativity arises through a system of inter-related forces operating at multiple levels, often requiring interdisciplinary investigation.' (Hennessey & Amabile, 2010, p. 571).

Creative Thinking Spiral In another view of creativity as a cyclical dynamic process, Resnick (2007) proposes the *creative thinking spiral*, pictured in Figure 3.8.

> 'people imagine what they want to do, create a project based on their ideas, play with their creations, share their ideas and creations with others, reflect on their experiences - all of which leads them to imagine new ideas and new projects. As students go through this process, over and over, they learn to develop their own ideas, try them out, test the boundaries, experiment with alternatives, get input from others, and generate new ideas based on their experiences.' (Resnick, 2007, p. 1)

Although authors such as Boden (2004) have long championed a more theoretically informed view of the creative process in computational creativity, this is only now starting to be put into practice (Chordia & Rae, 2010; Saunders et al., 2010; de Silva Garza & Gero, 2010, all using the DIFI framework) (Wiggins, 2006a; Ritchie, 2006; Thornton, 2007; McLean & Wiggins, 2010; Forth et al., 2010, all using Boden's three-fold model of creativity). The recent dynamic models of the creative process proposed by Hennessey and Amabile (2010) and by Resnick (2007) have not had the same length of time to exert influence on computational creativity; it remains to be seen whether these models will be discovered and experimented with by the computational creativity research community.

Figure 3.8: The Creative Thinking Spiral, proposed by Resnick to demonstrate the cyclic way in which people come up with new ideas, develop and share them, then reflect on the ideas as feedback for new ideas (Resnick, 2007, untitled figure on p. 1).

3.4.3 Conceptual analysis of creativity

Creativity can be seen as an example of what Gallie described as an *essentially contested concept* (Gallie, 1956): a subjective concept whose meaning seems to be commonly understood, with a variety of interpretations available to be attached to that concept, but where a fixed 'proper general use' is elusive (Gallie, 1956, p. 167).

Essentially contested concepts can be used for qualitative judgement of worth; however their subjective and abstract nature makes it difficult for people to pin down these judgements in words in a way which is exact, complete, objective and indisputable. Gallie cites the concepts of art, democracy, social justice and the practice of religion as examples of essentially contested concepts (Gallie, 1956, p. 180). Though he does not mention creativity, it seems clear that creativity could also be an example of an essentially contested concept (Torrance, 2012, personal communications).

The value of identifying creativity as an essentially contested concept is seen in Gallie's discussions of understanding these abstract notions better by reflecting on this type of concept more generally, helping us accept and navigate difficulties caused by disputes about a concept's exact meaning:

> 'I hope to show, in the case of an important group of concepts, how acceptance of a single method of approach - of a single explanatory hypothesis calling for some fairly rigid schematisation - can give us enlightenment of a much needed kind.' (Gallie, 1956, p. 168)

Gallie defines essentially contested concepts through several features, such as being internally complex in nature, but amenable to being broken down into identifiable constituent elements of varying relative importance, dependent on a number of factors such as context and individual preference. The existence of this variety of interpretations is contributory to the coining of the descriptor *essentially contested* for such concepts; though there is consensus on the concepts' meaning in very general

terms, exact interpretations differ. There is not a single agreed instantiation of these concepts but instead many reasonable possibilities. The influences of changing circumstances and contexts, and the nature of essentially contested concepts, mean that the interpretation of essentially contested concepts will not reach a point of resolution in their meaning or instantiation but will always continue to be contested. It is more productive, argues Gallie, to acknowledge that these different interpretations will and do exist, and work from that as the basis for understanding words like creativity, rather than expend time and energy in arguing the case for a single interpretation to be the standard meaning. Then we can support evaluative judgements of creativity by referring to 'the respective contributions of its various parts or features' (Gallie, 1956, p. 172).[27]

The varying interpretations of creativity mostly revolve around the type of creativity being displayed, being determined by reflections of the domain in which creative activity is taking place and the demands of that domain on creative products and processes to be used, as well as the background, skills and aims of the person being creative.[28] Different types of creativity manifest themselves in different ways, although they can all be identified. On this note, Wittgenstein has observed how words and language can be interpreted in multiple ways (Wittgenstein, 1958). Giving the example of the concept of a game, and the many differing interpretations of what constitutes a game, Wittgenstein comments on the way in which we identify these different interpretations under the word 'game':

> 'we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. ... I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way. And I shall say: "games" form a family.' (Wittgenstein, 1958, Part 1, Paragraphs 66-67)

Similarly, with creativity, different manifestations of creativity do not necessarily need to all share any common, core elements ('family resemblances') in order to be identified more generally as part of the creativity 'family',[29] but relationships between these different manifestations show shared characteristics similar to Wittgenstein's family resemblances in language. So to understand creativity, we can investigate what resemblances exist across different instantiations of creativity.

Wittgenstein is widely considered to have made a large contribution to our general understanding of the meaning of language. Although Wittgenstein earlier believed that language represents a fixed logical structure or picture that is hooked to a particular part of reality (Wittgenstein, 1922), later

[27] Such a strategy is an important part of the work described in this thesis, as shall be seen in the next Chapters.

[28] This ties in with the discussion of the *Four Ps* of creativity, discussed in Section 3.4.2.

[29] Excepting perhaps the presence of the combination of originality and value that has often been argued as necessary (if not necessarily sufficient) for creativity. For discussions on the interpretation of creativity as originality and value, see Section 3.4.1 of this Chapter. As Chapters 4 and 6 will illustrate, even these two features are not always seen as crucial or even the most important for certain types of creativity, such as the downplaying of value in creative musical improvisation compared to other more important factors such as social communication and interaction, or intention and emotion involvement in the improvisatory process.

Wittgenstein took a more pragmatic view of language - that its meaning is what we make it and how we use it, in our (consensual and collaborative) usage of language (Wittgenstein, 1958). Our public, shared understanding of language makes us interpret it the way we do - and this is not fixed but can evolve and change shape over time as public usage changes.[30] To understand the use of a word, one must know background information and context. For example in paragraph 31 of Wittgenstein (1958), Wittgenstein gives the example where a chess piece is identified to someone as a "king". To understand this usage of words, the person must already know the rules of the game of chess, or must at least know what it means to have a piece in a game which could be called a king (say if this introduction to the king piece comes during explanations of how to play chess). One would not expect the person to take any other interpretation of the word "king" in this context.

What this resolves to is that Wittgenstein sees the meaning of a word or statement is in how that language is used, rather than trying to understand what assigned meaning a word represents. A single word can have multiple uses (and therefore multiple meanings) depending on context (polysemy). Wittgenstein also talks about words that appear similar but have very different functions, using the illustrative example of a set of crank handles for a train engine - which all look the same but perform very different functions when used. His key point is that words' meanings are set by how we use them and our language is grounded in rules set by shared consensual usage.

Similarly, Waismann (1968) criticises the natural tendency (as he perceived) of many of his contemporaries (excluding Wittgenstein) to analyse statements and reduce them wherever possible to underlying rules. This is, Waismann argues, overlooking a crucial issue, that of the "open texture" or "Porosität der Begriffe" of many concepts.[31] Language cannot universally be reduced to logical statements about sense data using generally applicable methods; instead the heterogeneity of language should be taken into account to cope with the 'systematic ambiguity' of language (Waismann, 1946, p. 228). Waismann (1946) sees language as consisting of many layers (strata), with the meaning of statements and concepts being governed by features or 'motives' of each layer. Of these motives, open texture is the most pertinent for our interest in understanding the concept of creativity.[32]

Open texture concerns concepts which are multi-faceted in meaning, where it is difficult to fully

[30] Incidentally, Wittgenstein concluded from these views that any philosophical problems arising from this were actually only linguistic puzzles, arising from the rules and constraints set down by how we use our language and grammar. So if we voice a particular philosophical 'problem' then this problem is only extant because of how we use our language. Rather than being a problem with the language itself, this is instead an issue with our *usage* and *application* of language being inadequate, hence providing us with philosophical puzzles that we try to solve (with language). To Wittgenstein, then, the problem of interpreting creativity becomes a mere puzzle to be solved, arising from the simple problem that our language does not have the rules in place to apply when interpreting what creativity is, due to the way in which we use language (particularly the way we use the word creativity).

[31] Waismann was reported to be strongly influenced, though not completely directed, by the views of Wittgenstein (Edmonds & Eidinow, 2001; Bix, 1991). 'Waismann's work was devoted largely to presenting Wittgenstein's ideas in a more accessible form; however, some of Waismann's concepts were his own extension of Wittgensteinian ideas. The concept of "open texture" belonged to the second group' (Bix, 1991, pp. 55-56).

[32] An interpretation of open texture has also become influential in legal interpretation of subjective concepts (Bix, 1991).

define that concept in all possible concepts and scenarios:

> 'we can never exclude altogether the possibility of some unforeseen situation arising in which we shall have to modify our definition. Try as we may, no concept is limited in such a way that there is no room for any doubt. We introduce a concept and limit it in *some* directions; ... This suffices for our present needs, and we do not probe any farther. We tend to *overlook* the fact that there are always other directions in which the concept has not been defined. And if we did, we could easily imagine conditions which would necessitate new limitations. In short, it is not possible to define a concept like [this] with absolute precision, i.e. in such a way that every nook and cranny is blocked against entry of doubt. That is what is meant by the open texture of a concept.' (Waismann, 1968, p. 120, emphases in original)

Waismann differentiates between open texture and vagueness in a word's meaning. If it is possible to define a word more accurately, such that any possible vagueness can be (eventually) ruled out, then this is not open texture. Open texture is where 'we can never fill up all the possible gaps through which a doubt may seep in. ... Vagueness can be remedied by giving more accurate rules, open texture cannot.' (Waismann, 1968, p. 120).

Crucially for our concerns, the implications of open texture mean that we are unlikely to reach the point where we have a full, complete and static definition of creativity. Other possible interpretations exist, especially for a concept such as creativity which is by its nature likely to extend into new and unpredictable directions:[33]

> 'Every description stretches, as it were, into a horizon of open possibilities: however far I go, I shall always carry this horizon with me.' (Waismann, 1968, p. 122)

From the philosophical reflections outlined above in this Section, we are guided towards accepting the dynamism of language and its usage in different contexts. We should aim towards building a growing and informed understanding of the interpretation of the word 'creativity' as it is used over time and across contexts, rather than imposing a single static definition to be adopted regardless of current context and usage.

3.5 Legal definitions of creativity

One arena in which terms need to be tightly defined and unambiguous, and the exact meaning of specific terms is often fiercely debated, is legal practice. In a legal context, we encounter a by-now-familiar problem with determining what creativity is:

> 'a simple, appropriate definition of the word creativity does not exist' (Clifford, 2004, footnote on p. 260)

> 'Despite the value of facilitating creativity for intellectual property law, understanding creativity is hardly something within the competent domain of law and legal analysis. Not surprisingly, the legislative and judicial development of intellectual property law has paid remarkably little attention to what is known about how to promote creativity.' (Mandel, 2011, p. 1)

[33] As we see throughout this Chapter.

Feist Publications, Inc v. Rural Telephone Service Co.
The *Feist Publications, Inc v. Rural Telephone Service Co.* 1991 Supreme Court case set a legal precedent in the relationship between copyright and creativity. In this case, Feist Publications, Inc. (Feist) were sued by Rural Telephone Service Co. (Rural) for copyright infringement, after Feist copied details from a telephone directory produced by Rural for publication in their own telephone directory service. Rural had previously denied Feist permission to reproduce this information. The Court ruled that as compilation of telephone directories did not 'entail a minimal degree of creativity', they were not 'sufficiently original that Congress may protect such compilations through the copyright laws.' (Feist, 1991, p. 348). 'As a constitutional matter, copyright protects only those constituent elements of a work that possess more than a de minimis quantum of creativity.' (Feist, 1991, p. 363)

The legal implications of the question *what is creativity?*[34] have mainly been discussed due to its connections with originality, an important concern in cases of patent and copyright law. In particular, whether creativity is a necessary condition for originality is an unresolved question that arises in various legal contexts.[35] This next Section looks at these issues and also examines how computational creativity has been discussed in a legal context.

3.5.1 Creativity and Copyright

Copyright is the part of intellectual property law concerned with protecting original work that has been produced in some physical form (Copyright, Designs and Patents Act, 1988; Intellectual Property Office, 2010). A landmark case in US law (Feist, 1991) determined that in granting copyright:

> originality requires independent creation plus a modicum of creativity' (Feist, 1991, p. 346)

The amount of creativity required to be present in original, copyrightable work is 'famously low' (Mandel, 2011, p. 11) - 'even a slight amount will suffice' (Feist, 1991, p. 345). The emphasis is on some (or indeed any) creativity being detected in the work in order to qualify it as original and copyrightable (Karjala, 2008). As the Court's pronouncements in Feist (1991) do not give a definition of what they mean by creativity, however, the legal ramifications of Feist (1991) can make only limited contribution to future cases (Clifford, 2004). Clifford cites a number of examples where decisions made on the creativity present in a work often contradict each other across different cases.

Analysing intellectual property law in the context of creativity research, particularly copyright, Mandel argues that intellectual property law treats copyrightable work as primarily the work of an individual, needing adaptation to adequately accommodate collaborative work.

[34] Given the scope of this thesis, this Section is limited to giving an overview of some pertinent legal discussions of creativity, mostly from UK or US law, rather than attempting a comprehensive presentation of different legal systems.

[35] This question takes the reverse perspective to creativity researchers, who are more interested in originality as a constituent part of creativity rather than vice. versa.

3.5.2 Patent law

> 'The United States Supreme Court has ruled in numerous cases that an invention is an idea rather than an object. If a man can prove that an idea was his by demonstrating or providing evidence that only he had the knowledge from which it was synthesised, he can claim patent to the invention.' (Rhodes, 1961, p. 305)

Patentability is related to how original intellectual property is (Holmes, 2009) and via originality to creativity, given the emphasis on originality for creativity (Feist, 1991). Patents are granted for novel inventions or ideas which may not necessarily have been implemented yet, whereas copyright asserts ownership of produced artefacts (Copyright, Designs and Patents Act, 1988).

> 'a man cannot have a valid patent for something which is merely different from existing knowledge. To the requirement of novelty (i.e. difference) there is superadded the elusive requirement of "subject-matter" or "inventive ingenuity." ' (Potts, 1944, p. 113)

Potts identifies an ambiguity as to what justifies something as an invention of something new rather than just an improvement of something that exists. This relates to the question of 'what is creativity' but focuses on a slightly tangential question (for our purposes): what is an invention? Potts argues that there is a continuous spectrum with inventions at one end and 'mere improvements' at the other, with the level of novelty increasing across the spectrum.

> 'there can be all gradations of novelty, from something different but obvious to anyone having a minimum of practical knowledge and experience, to something which is so brilliant and remarkable that the highest genius (whatever that may be!) is required to produce it.' (Potts, 1944, p. 115)

3.5.3 Computational creativity and the law

The legal ramifications of a computer system being creative have not yet been considered to a great extent. Where this has been investigated, a common response is that a computer cannot be creative, precisely because it is a computer and because creativity is a human behaviour. Clifford (2004) questions whether the absence of human involvement in computational creativity has repercussions in how the creative act is perceived legally, specifically in determining whether a computer-generated piece of work can/should be copyrightable.

For the purposes of copyright, Clifford states the importance of considering process in creativity, to make the case that computer-generated work should by default *not* be considered creative, because 'it would not only be difficult to establish a requisite amount of intellectual creativity, it would be impossible to establish origin creativity. Consequently ... the creativity in the product must be the result of a human-based creative process' (Clifford, 2004, p. 272). Clifford uses previous legal judgements (Computer Associates International, Inc. v. Altai, Inc. (1992), which considered whether one computer program was a copy of another) to suggest a framework for identifying creativity based on the 'abstraction-confirmation-examination' tests proposed in the Computer Associates case:

> ‘ A. Abstraction: Finding the “Expressive Constituents” within the Work ... [expressive constituents refer to the ideas expressed in the work]
>
> B. Confirmation: Insuring[36] that Some of the Expressive Constituents Originated from a Human ...
>
> C. Examination: Did the Author Deliberately Decide to Do It That Way?
>
> D. Applying the Abstraction-Confirmation-Examination Test’ (Clifford, 2004, pp. 291-298)

This would clearly fail to identify a computational creativity system as creative due to step B which requires that all expressive constituents originate from people, not from computers. Similarly, Warner (2010) argues that mechanical processes such as that employed by (all) computational processes demonstrate an absence of creativity, because of the description in Feist (1991) of processes that are ‘so mechanical or routine as to require no creativity whatsoever’ (Feist, 1991, p. 362). Warner arrives at this conclusion by arguing that computation involves ‘mechanical procedures’, and that the use of the phrase ‘so mechanical’ in Feist (1991) is equivalent to the use of the word ‘mechanical’.

Warner perhaps over-interprets the phrasing in Feist (1991), equating ‘mechanical’ with ‘computational’ without justification or acknowledgement. Warner’s ‘careful and exhaustive process of correlation’ (Warner, 2010, p. 10) is also logically flawed, as it is based on the argument that ‘so mechanical’, in a *continuous* sense, correlates directly to ‘mechanical’, in a *discrete* sense:

> ‘*So* in “so mechanical or routine” ... directly qualifies mechanical and indicates that it is to be understood in its most intense sense. *Mechanical* in a “mechanical procedure” is understood in an absolute sense, as matter of preliminary scientific definition. ... A similarity in intensity of meaning then has been established.’ (Warner, 2010, p. 5)

Even within the pages of the *Computer* journal, Holmes questions whether machines can generate patentable products (Holmes, 2009), due to problems with originality. Holmes claims that ‘originality must lie in the personality of the inventor’ (Holmes, 2009, p. 99) and asks ‘what becomes of personality’ (Holmes, 2009, p. 100) when original products are automatically generated. He concludes:

> ‘Computing professionals do great social harm when they portray their machinery as anything but tools to amplify human accomplishment and originality.’ (Holmes, 2009, p. 99)

Some examples of computational creativity have been recognised. Computer programs and databases are legally treated as literary works (Copyright, Designs and Patents Act, 1988; Holmes, 2009). *Gaussian-Quadratic*, a computer-produced artwork, was successfully registered for copyright in the 1960s, (Noll, 1994) though Noll (the human artist involved) struggled to obtain this copyright. The copyright application was repeatedly rejected due to the role of the computer and the use of random generation in generating the artwork. The application was eventually accepted because a human (Noll) had written the program and had used an algorithmic process to generate ‘randomness’.

[36] Perhaps *Ensuring* was intended here, rather than *Insuring*.

3.6 Different perspectives on creativity

Sections 3.3, 3.4, 3.4.3 and 3.5 have examined some key contributions towards defining the concept of creativity. Abstracting from specific individual definitions, some higher-level issues emerge that affect how one views creativity, subsequently shaping how a definition is formed. This Section examines such issues and states what decisions this thesis work makes on these issues, to acknowledge the effect of these choices.

3.6.1 Creative genius or everyday creativity?

Early study of creativity assumed a focus on creativity at the *genius* level, such that one must be particularly gifted and make a substantial contribution to their domain in order to be described as creative (Poincaré, 1929; Hadamard, 1945; Wallas, 1945). More recently, Sternberg (1988) takes a 'genius' approach to creativity, distinguishing creative and non-creative individuals.

In the same volume as Sternberg (1988), Weisberg draws attention away from 'the genius view of creativity' (Weisberg, 1988, p. 150) to look at the thought processes of 'ordinary' people (Weisberg, 1988, p. 151). Koestler (1964) and Rhodes (1961) echo this focus. Boden (2004) also focuses on everyday layperson (laycomputer?) creativity rather than specialist expert creativity, arguing that creativity is not just a special attribute of the few. Despite Boden (2004) concentrating on everyday creativity, ideas in Boden (2004) are often illustrated with examples of genius creativity, such as citing Bach, Newton and Shakespeare (Boden, 2004, p. 278) to exemplify how we admire creative work.

The key question in this debate is whether creativity can be learnt or developed, or if it is an attribute only for the selected few to possess. The work of Odena and Welch (2009) and others in educational creativity development is built around the belief that it is possible to nurture and develop an individual's capacity for creativity (their 'little-c' creativity (Gardner, 1993), or 'P-creativity' (Boden, 2004), as will be explained in Section 3.6.3.

Everyday creativity is becoming more prevalent in recent research on creativity.[37] This preference may be because research aimed at developing everybody's creativity is more inclusive and contains more perceived value than research which is aimed at relatively few individuals only.

3.6.2 Cognitive approaches or embodied creativity?

As discussed in Section 3.4.1, the three-fold framework of creativity outlined in Boden (2004) has had some influence in computational creativity. Boden's framework and its successors are based on 'conceptual spaces' and the cognitive operations of exploration, transformation and combination.[38] It

[37] As demonstrated in the adoption of 'everyday creativity' as the theme for the 2009 ACM conference in Creativity and Cognition, strongly aiming the conference away from genius or domain-specific creativity (Bryan-Kinns, 2009).

[38] Chris Thornton and Jens Holger-Streck at the University of Sussex are in the initial stages of a Leverhulme Trust project examining how this translates to current research, looking at the cognitive foundations of computational creativity.

is questionable how well this theory sits with an embodied view of creativity situated in an interactive environment, with bi-directional influences exchanged rather than the creator having a one-way influence on the conceptual space. This debate is reminiscent of the similar debate in artificial intelligence (A.I.) between symbolic and non-symbolic approaches to A.I. (Minsky, 1982).

Rather than preferring one perspective over the other, this thesis takes a more inclusive approach, enabling the spectrum of possibilities between these two views. The components identified in Chapter 4 represent both cognitive and embodied aspects of creativity. As a brief and introductory example of this, the cognitive processes referred to in the *Thinking and Evaluation* component co-exists with the *Social Interaction and Communication* component.

3.6.3 Levels of recognition of creativity

Boden (2004) distinguishes between P-creativity (psychological/personal creativity) and H-creativity (historical creativity). P-creativity is where a person creates something that is original to them; it may already exist in the world but the creator is not aware of it. H-creativity, on the other hand, is a creation that does not exist anywhere else. Anything H-creative is P-creative by definition; H-creativity is P-creativity with the addition of historical novelty.

A similar debate resounds in psychological research, in slightly different terminology.

- P-creativity (Boden, 2004) $\approx$ *new* (Elliott, 1971) or *little-c* creativity (Gardner, 1993).
- H-creativity (Boden, 2004) $\approx$ *traditional* (Elliott, 1971) or *big-C* creativity (Gardner, 1993).

This thesis focuses on P-creativity, as is common in computational creativity research, as a simplifying assumption to avoid the practical problems of modelling all human knowledge in a computer-accessible format (see Chapter 5 Section 5.1.3 for further discussion).

3.6.4 The generality of creativity in different domains

There is some debate over whether there is such a thing as 'creativity' in general, shared between different types of creativity and independent of domain specifics (as argued by Plucker (1998). The alternative view is that creativity is domain-specific (Baer, 1998; Csikszentmihalyi, 1988; Gabora, 2011); e.g. musical creativity is independent of artistic creativity, or scientific creativity. If this latter view is taken, then a further issue is to identify an appropriate level of detail with which to define domain boundaries. As illustration, the domain for the case study systems in Chapter 6 could be musical creativity, or could be refined further to 'microdomains' (Baer, 1998, p. 174) such as musical improvisational creativity, or jazz musical improvisational creativity (as all the systems evaluated in Chapter 6 could be said to improvise some form of jazz - see Chapter 6 Section 6.2). All these labels describe the musical systems analysed in Chapter 6 to differing degrees of specificity.

In debate in the *Creativity Research Journal* in 1998, both Plucker and Baer concede a lack of evidence to conclusively decide if creativity is domain-general or specific to individual domains. Plucker

(1998) explains this through bias in the methods used for detecting creativity. Tests for specific domain achievements, such as assessments of performance quality, show creativity as domain-specific, whereas psychometric tests show evidence for creativity being domain-general. Baer (1998) argues it is safer for researchers to assume creativity is domain-specific:

> 'if the generic hypothesis is correct, then the content of the exercises one uses does not matter and nothing is lost by making the incorrect (domain specificity) assumption. But if the specificity hypothesis is correct and one chooses all exercises from the same domain (which the generic hypothesis allows), then the loss will be significant, as any improvement in creative thinking will be limited to the single content domain from which the exercises are chosen.' (Baer, 1998, p. 176)

Psychometric tests for creativity make the opposing assumption that creativity is domain-general; creativity in one context, such as devising multiple uses for a household object, can predict a person's general creative abilities. Baer (1998) offers several examples where creativity in one domain does not necessarily correlate to creativity in another, as judged by experts in each domain (Amabile, 1996).

Difference in creative ability in different domains may be down to variances in domain knowledge and competence (Plucker, 1998; Amabile, 1996; Rowlands, 2011). Amabile (1996) notes that *Creativity-Relevant Skills* such as cognitive and exploration skills must be accompanied by two other skillsets necessary for creativity: *Domain-Relevant Skills* specific to a given domain and *Task Motivation*, the motivation that the creative person has for the task.

Plucker and Beghetto dismiss the notion that people need to reach a certain level of knowledge in order to make 'meaningful contributions' to a domain (Plucker & Beghetto, 2004, p. 163). Instead they argue that a person can achieve levels of creativity in a domain that are relative to the knowledge they have of that domain. One might question how useful such achievements would be to a broader community, however Plucker and Beghetto's point allows for *little-c* or *P-creativity* creativity (see Section 3.6.3) that is useful at an individual level, independently of its importance in a wider context.

Plucker and Beghetto (2004) and Baer (2010) disagree with both Baer (1998) and Plucker (1998), representing a shift in opinions for both Baer and Plucker since 1998. Plucker and Beghetto (2004) and Baer (2010) now adopt a hybrid view, acknowledging that some aspects of creativity transcend domains and others are specific to that domain. This view has been taken in several creativity models (Csikszentmihalyi, 1988; Sternberg, Grigorenko, & Singer, 2004; Wiggins, 2006a; Kaufman, 2009).

> 'Rather than artificially split creativity, domain-specific *and* domain-general features of creativity can be examined. In doing so, a richer question can be addressed: What aspects of creativity are domain general and which aspects are domain specific?' (Plucker & Beghetto, 2004, p. 159)

In this spirit, this thesis adopts a hybrid position similar to Plucker and Beghetto (2004) and Baer (2010). Creativity shall be treated as being independent of individual domains to some extent[39] but with some elements of creative behaviour specific to the relevant domain.[40] Therefore both domain-

[39] As will be explored in Chapter 4.

[40] Domain-specific elements of creativity in a given domain shall be investigated in Chapters 6 and 7

general and domain-specific elements of creativity should be used to evaluate creativeness.[41]

3.7 Summary

The difficulty of capturing in words an adequate definition of creativity should not discourage us from such an attempt (Rhodes, 1961; Plucker et al., 2004; Colton, 2008b), even though other researchers have been swayed away from this task (as reported in Sternberg & Lubart, 1999; Veale et al., 2006). The goal of this thesis is an evaluation methodology for computational creativity; Section 3.1 explained how a rigorous and comparative evaluation process needs clear standards to use as guidelines or benchmarks (Torrance, 1967; Kaufman, 2009). Creativity is generally ill-defined (as discussed in Section 3.2.2) with no standard definition agreed upon (Section 3.2.1). There are complications in constraining this seemingly mysterious term (Section 3.2.3) to definition. Dictionary definitions are restricted by their format to provide only cursory, shallow insights into what creativity truly is (Section 3.3). In research definitions, though, many researchers have helped 'demystify' creativity, both in computational and human creativity research (3.4.1 and Sections 3.4.2, respectively).

Generally computational creativity is defined by those in the field as human creativity simulated by computational systems (Section 3.4.1), without further clarifying what human creativity is perceived to be. Consequently it is important to examine how creativity is defined in human-focused research areas. Although some progress has been made in defining creativity in a computational context, this is often by redefining creativity as something closely related, for example: searching for solutions to problems, combining novelty and value, combining exploration, transformation and association of concepts in a conceptual space, or defining a 'creativity-like' concept (Section 3.4.1).

These approaches are practically useful and more computationally malleable; however they all suffer the same problem; their definition of creativity may not actually be definitions of creativity as a whole, but of some subset of creativity as seen from a particular perspective. To avoid similar problems, Section 3.4.2 delved into psychological and other research on human creativity to see what can be learnt from the frameworks, psychometric tests and models of creativity examined.[42] In such research, the *Four Ps* construct (Section 3.4.2) ensures we pay attention to four key aspects of creativity: the creative Person, the generated Products, the creative Process and the Press/Environment hosting and influencing the creativity. This framework helps to see creativity in a wider context. Models of the creative process and psychometric tests of creativity based on the creative person generating products are useful in their own limited sphere, but creativity is complex and multi-dimensional, requiring broader treatment. The confluence approach to creativity (Section 3.4.2) takes a reductionist approach, understanding creativity as a whole by breaking it down into smaller constituent parts.[43]

[41] The SPECS methodology presented in Chapter 5 uses both general and domain-specific aspects of creativity.

[42] The work in Chapter 4 will take similar steps by using multiple sources from different academic viewpoints.

[43] Chapter 4 will adopt a confluence-style approach, seeking to capture a wider disciplinary spectrum of perspectives on creativity that has previously been attempted (Sternberg & Lubart, 1999; Mayer, 1999; Ivcevic, 2009).

Such multi-dimensional approaches to understand the complexities of creativity are supported by the philosophical literature on how to interpret and analyse concepts like creativity, as reported in Section 3.4.3. 'Creativity' must be understood in context of the scenarios and circumstances it is used in; an exhaustive and complete definition of the term is somewhat 'utopian' (Waismann, 1965, p. 76,223).

In the search for a more precise and accurate definition of creativity, Section 3.5 explored how creativity has been defined and used in the law, via selected relevant UK and US examples. Here too, though, there was no standard definition of creativity to be found (Clifford, 2004; Mandel, 2011), despite the need to detect the presence of creativity for legal reasons (Feist, 1991; Karjala, 2008; Copyright, Designs and Patents Act, 1988). What did appear in this review of legal research was an interesting conflict of opinions on how computational creativity is perceived legally (Section 3.5.3).

Several competing views of creativity exist (Section 3.6). Sometimes differences of opinion do not need to be directly resolved but can co-exist, such as whether creativity is centred around cognitive function or whether it is embodied and situated in an interactive environment (Section 3.6.2), or whether creativity is domain-independent or domain-specific. Other conflicts arise where a previously narrow view of creativity has been widened in perspective. An inclusive view of creativity should adopt the wider perspective. For example rather than focus just on creative geniuses one should focus on human everyday creativity, of which genius is a special case (Section 3.6.1). Similarly, P-creativity encompasses H-creativity (Boden, 2004) (Section 3.6.3).

To satisfy the need for clear and defined benchmarks by which to evaluate progress in creativity research, particularly computational creativity research, there is much useful contributory material towards a satisfactory definition, as reported in this Chapter. What remains to be done is to draw this assortment of material together, unifying the perspectives where possible to remove disciplinary separation and boundaries. Chapter 4 will describe how this task has been undertaken, reporting the results of this work and reflecting on the implications for the study of computational creativity.

Chapter 4

Identifying key components of creativity

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012) and peer-reviewed conference papers (Jordanous & Keller, 2012, 2011; Jordanous, 2011a, 2010a).

Figure 4.1: *Wordle* word cloud of this Chapter's content

Overview

As a concept, creativity has proved resistant to a satisfactory and comprehensive summative definition, despite numerous attempts. Section 4.1 motivates the need for a working definition of creativity for evaluation. Rather than adding another to the mass of existing definitions, Section 4.1 outlines this approach's aims to extract the common themes of creativity across disciplinary or domain specifics. Combining several viewpoints across various perspectives in creativity research leads to a more inclusive summary of how we define creativity. This is inspired by the use of ontologies to represent knowledge about a concept or concepts (Chandrasekaran, Josephson, & Benjamins, 1999).

Cognitive linguistics advocates that the meaning of a word is dependent on the context it is used in (Evans & Green, 2006). In particular, the premise exists in cognitive linguistics that the study of language helps reveal how people think (Lakoff, 1987; Lakoff & Johnson, 1980). Words used frequently in discussions of the nature of creativity provide the context for use of 'creativity' and are therefore linked to our interpretation of creativity (Oakes, 1998; Kilgarriff, 2001, 2006). Also, in Chapter 3, Section 3.4.3 described how the meaning of words like creativity can best be understood by seeing how they are used and by identifying individual components that collectively contribute to the meaning of creativity. In a similar vein, Section 4.2 of this Chapter reports on an empirically-based approach for harnessing the most discussed aspects of creativity. A corpus of academic papers is collated, representing sixty years of research into the nature of creativity, from research perspectives such as psychology, education, computational creativity and others. This corpus is analysed against a corpus of matched papers on subjects unrelated to creativity, using the log likelihood ratio (LLR) statistic on word frequencies in the two corpora. 694 *creativity words* are identified that appear significantly more often in the creativity papers corpus than expected.

These creativity words are clustered into groups of words with similar meanings. Through inspection of these clusters, a total of 14 different components are identified, representing different aspects of creativity that collectively contribute to the overall meaning of the word 'creativity'. Section 4.3 presents these components. These components are offered to address the clear need within the computational creativity research community (Chapter 2) and beyond (Chapter 3) for a comprehensive, accurate and cross-disciplinary understanding of creativity.

4.1 Creativity is what we say it is: Motivation and aims for this work

The need for a universally-accepted definition of creativity is clear. Chapter 3 Section 3.1 illustrates how this point has been made by several creativity researchers, concluding that although it is problematic to define creativity, the benefits of a more informed and objective approach outweigh these concerns. As Hennessey and Amabile point out (Hennessey & Amabile, 2010), an accurate and encompassing definition assists our understanding of creativity and further research. It also smooths

out individual variances in interpretations of creativity, to highlight the agreed-upon universal components of creativity as a concept, transcending any disciplinary or domain-specific biases (Plucker & Beghetto, 2004). Chapter 3 Section 3.2 argues that the difficulty of capturing in words an adequate definition of creativity should not discourage us from such an attempt, due to the need for clear evaluative criteria that reflect the nature of creativity as closely as possible (Chapter 3 Section 3.1).

As discussed in Chapter 3 Section 3.2.2, we assume an intuitive understanding of the concept of creativity but lack a universally accepted and comprehensive definition of the concept. There have been many attempts to capture the nature of creativity in words; indeed the work in this Chapter is based around thirty such examples. Despite these attempts, Chapter 3 demonstrates that no definitive consensus has yet been reached on exactly what creativity is. Multiple viewpoints exist, many of which prioritise different aspects of creativity. This recalls the perspectives on creativity discussed in Chapter 3 Section 3.4.3, where differing interpretations of concepts like creativity will necessarily remain unresolved because of their complex, context-dependent nature. Instead of seeking a single definitive realisation of creativity, it is more effective to understand creativity by identifying constituent parts and features that influence how 'creativity' is interpreted, taking into account the many effects of context on how these parts interact and exert their relative influence.

In the academic literature on creativity, many repeated themes have emerged in the literature as important components of creativity. Differing opinions can be found, though, in what are considered primary contributory factors of creativity. For example psychometric tests for creativity (e.g. Guilford, 1950; Mednick, 1967; Torrance, 1974; Kraft, 2005; Runco, Dow, & Smith, 2006; Kaufman, Kaufman, & Lichtenberger, 2011) focus on *divergent thinking* and *problem solving* as key attributes of a creative person. In contrast, computational creativity research (e.g. Wiggins, 2006b; Peinado & Gervas, 2006; Pease et al., 2001) places emphasis on the *novelty* and *value* of creative products. Whilst there is some agreement across academic fields, the differing emphases contribute to subtle variances in the interpretation of creativity.

Cognitive linguistics advocates that the meaning of a word is dependent on the context it is used in (Evans & Green, 2006). In particular, the premise exists in cognitive linguistics that the study of language helps reveal how we think (Lakoff, 1987; Lakoff & Johnson, 1980).

Identifying what contributes to our intuitive understanding of creativity can guide us towards a more formal definition of what creativity is. The work presented here adopts a confluence-style reductionist approach (Chapter 3 Section 3.4.2), understanding creativity as a whole by breaking it down into smaller constituent parts. This approach works on the principle that creativity emerges as a result of several components converging (Boden, 1994a; Sternberg & Lubart, 1999; Mayer, 1999; Ivcevic, 2009) and investigates what these components are. Similar approaches have been applied to define other concepts that are difficult to define in words, for example: consciousness (Seth, 2009), personality (McCrae & Costa Jr, 1999), flow (Romero & Calvillo-Gamez, 2011) and

musical perceptual preferences (Huron, 2001).

Some of the issues surrounding creativity definition have been debated for decades; clearly these cannot all be resolved satisfactorily in the scope of this thesis. Hence the current requirement becomes a *working understanding* of creativity in computational systems which is practical, accurate and complete enough to be used as current evaluative standards.

This work aims to include a broader spectrum of perspectives on creativity than has previously been considered. The intention is to avoid being restricted by previously learned disciplinary boundaries or constraints by employing empirical methods where possible over a wide and cross-disciplinary range of sources. This empirical approach draws together several academic opinions across disciplinary divides, for a more universally acceptable definition of creativity.

There is much useful contributory material towards a definition of creativity.[1] This assortment of material can now be drawn together to unify different perspectives where possible, to avoid the disciplinary separation that is often seen in creativity research (Hennessey & Amabile, 2010). This Chapter describes how this task has been tackled (Section 4.2) and reports the results (Section 4.3).

4.2 Methodology for identifying components of creativity

This Section describes the steps taken[2] to identify components of creativity by analysing corpus data:

1. Employing natural language processing methods and a corpus-based statistical analysis to identify words that are associated with creativity: *creativity words*.
2. Using a statistical measure of word similarity and a clustering algorithm to help group the creativity words into semantically-related groups identifying different aspects of creativity.

4.2.1 Identifying words significantly associated with creativity

Compiling a creativity corpus

To identify words significantly associated with creativity, a representative *creativity corpus* was assembled, consisting of academic papers discussing the nature of creativity. The creativity corpus contains thirty such papers from a variety of academic stand-points, ranging from psychological studies to computational models. Appendix B lists these papers.[3]

The creativity corpus papers were chosen to cover a wide range of years (1950-2009) and academic disciplines. Figure 4.2 show the disciplinary make-up of the corpus as it changes over this time period.

[1] Selected key contributions were outlined in Chapter 3.

[2] Much of the natural language processing work was undertaken in conjunction with Bill Keller, a computational linguist at the University of Sussex. In this interdisciplinary project, working with a computational linguistics expert ensures that my ideas and motivation for this work are supported by a thorough background and up-to-date knowledge in computational linguistics research findings and methods. In what follows, where the work undertaken is joint rather than my own, I acknowledge Bill's contribution. Where there is no such acknowledgement, the work is my own.

[3] Due to practical considerations in some of the methods employed in this Chapter, the size of the creativity corpus was limited to thirty papers, to ensure the overall task remained manageable.

Creativity corpus:
a collection of thirty academic papers which explicitly discuss the nature of creativity

A paper was included if it was considered particularly influential, as measured by the number of times it had been cited by other academic authors. For papers published in very recent years and which have therefore not yet accrued many citations, selection was based on intuitive judgement.

Academic papers were used as the source of information, for several reasons:

- Ability to access timestamped textual materials over a range of decades.
- Ease of locating relevant papers: e.g. availability of tools to perform targeted literature searches, electronic publication of papers for download, tagging of paper content by keywords, citations in papers to other related papers.
- Publication of academic papers in an appropriate format for computational analysis: most papers that are available electronically are in formats such as PDF or HTML, which can be converted to text fairly easily.
- Availability of citation data as a measure of how influential a paper is on others: whilst not a perfect reflection of a paper's influence, citation data is often used for measuring the impact of a journal (Garfield, 1972) or an individual researcher's output (Hirsch, 2005).
- Availability of provenance data, such as who wrote the paper and for what audience (from the disciplinary classification of the journal).

Generating comparative data for statistical analysis

For statistical comparison purposes, a second corpus was compiled containing sixty papers on topics unrelated to creativity: the *non-creativity corpus*. Appendix B lists these sixty papers.

Non-creativity corpus:
Sixty academic papers on topics unrelated to creativity, from the same range of academic disciplines and publication years as the creativity corpus papers

The non-creativity corpus papers were selected to match the creativity corpus make-up as closely as possible (except for the creativity content), covering a similar range of disciplines over the same time period, as is pictured in Figure 4.2. A literature search retrieved, for each paper in the creativity corpus: the *two* most-cited papers in the same academic discipline, published in the same year, that did not contain any words starting "creat-" (creativity, creative, and so on). The non-creativity corpus is twice the size of the creativity corpus, to acknowledge that in general the set of academic papers on creativity is a small subset of all academic papers.

Pre-processing of the corpora

Both corpora were processed by Bill Keller (Jordanous & Keller, 2012) using the RASP toolkit (Briscoe, Carroll, & Watson, 2006) to perform *lemmatisation* and *part-of-speech* (POS) tagging.

- Lemmatisation converted words in the text from their existing form to a root form or lemma. This allowed the grouping together of words which are different forms of the same basic root. For example, the words *creates*, *created* and *creating* all occur in the corpus as distinct variants of the lemma *create*. During lemmatisation, RASP converted these words to a form "create+[RASP tag indicating the exact form of this word]". In later stages of text processing, only the text before the plus sign was used for this work.
- Each word was assigned a POS tag identifying its grammatical category (e.g. noun, verb, etc.), using the local context. This distinguished between words that fit more than one category, usually with different associated meanings according to category. For example, with POS-tagging, *novel* as a noun (*a good novel*) was distinct from the adjective (*a novel idea*).

RASP automatically performed other processing as it ran, placing each new sentence on a new line headed by a new line indicator `^_^` and annotating the punctuation used. These steps were inconsequential for the work described below so are mentioned only for completeness.

Figure 4.2: The disciplinary make-up of the corpora over time. The y-axis represents the count of papers per datapoint for the creativity corpus; this count will be doubled for the non-creativity corpus which is twice the size (1 paper in the creativity corpus is matched to 2 papers in the non-creativity corpus. Some disciplines have increased in output over time, for example computational creativity publications started to appear around the 1990s. This is reflected in the papers chosen. N.B. Some papers fit in more than one discipline.

To illustrate the form of each corpus after pre-processing, here is the first sentence from Guilford's seminal paper on creativity (Guilford, 1950), the first paper in the creativity corpus:

> I discuss the subject of creativity with considerable hesitation, for it represents an area in which psychologists generally, whether they be angels or not, have feared to tread.

After RASP processing, this became:

```
^_^ I_PPIS1 discuss_VV0 the_AT subject_NN1 of_IO creativity_NN1 with_IW considerable_-
JJ hesitation_NN1 ,_, for_IF it_PPH1 represent+s_VVZ an_AT1 area_NN1 in_II which_-
DDQ psychologist+s_NN2 generally_RR ,_, whether_CSW they_PPHS2 be_VB0 angel+s_-
NN2 or_CC not+_XX ,_, have_VH0 fear+ed_VVN to_TO tread_VV0 ._.
```

Without going into the exact meanings of each RASP tag,[4] this exemplifies the current form of the corpora.[5] Some additional pre-processing steps were employed by Bill Keller at this stage.

- Each word (POS-tagged lemma) was mapped to lower case, so capitalised forms (e.g. *Novel*) would be treated the same as lower case forms (e.g. *novel*).
- The data was filtered so that only words of the major categories (noun, verb, adjective and adverb) were represented. The major categories contain the semantic content of the papers making up the creativity corpus. Minor categories or 'function words', such as prepositions (e.g. *upon*, *by*), conjunctions (e.g. *but*, *or*) and quantifiers (e.g. *many*, *more*) offer little semantic content so have limited value for this work and can be removed from the data.

Generating word-frequency data

After RASP processing, word frequency information was generated by Bill Keller, consisting of two frequency counts for each word (POS-tagged lemma) in the creativity corpus, indicating how often that word appeared in each corpus. The frequency data informed the statistical tests described next.

Identifying significant creativity words using the LLR statistic

The word frequency data derived from the two corpora was used to establish which words occur significantly more often in the creativity corpus than in the non-creativity corpus, hence identifying words commonly used to discuss creativity. For this the log likelihood ratio (LLR) statistic (also referred to as G^2 or G-squared) was used. The log likelihood ratio statistic compares two corpora to measure how well data in one corpus fits a model distribution based on both corpora (Dunning, 1993; Kilgarriff, 2001; Rayson & Garside, 2000; Oakes, 1998). The higher the log likelihood ratio statistic, the greater deviation there is between the observed usage of a word and what was expected given the model distribution. This study compares the actual frequency with which words are used in the creativity corpus against a model distribution based on both the creativity corpus and non-creativity corpus. The log likelihood score for a given word w is calculated as shown in equation 4.1:

[4] Briscoe et al. (2006) gives details of RASP tags and further information

[5] From this point onwards, references to a 'word' in the corpus data indicate a lemma tagged with POS information.

$$LLR = 2\sum o_i(\ln o_i - \ln e_i) \tag{4.1}$$

where for $i \in \{1,2\}$, o_i is the observed number of occurrences of w in corpus i and e_i is the expected number of occurrences of w in corpus i. The expected count e_i of a word in corpus i is:

$$e_i = \frac{N_i \times (o_1 + o_2)}{N_1 + N_2} \tag{4.2}$$

where N_i is the total number of words in corpus i (i.e. sum of frequencies of all words in corpus i).

The log likelihood ratio is an alternative to the chi-squared test (χ^2). It has been advocated as the more appropriate statistical measure of the two for corpus analysis as it does not rely on an assumption of normality in word distribution (Dunning, 1993; Kilgarriff, 2001; Oakes, 1998), a particular issue when analysing smaller corpora, such as those used in the current work.[6] The LLR statistic is more accurate in its treatment of infrequent words in the data, which often hold useful information. The χ^2 statistic tends to under-emphasise such outliers at the expense of very frequent data points.

Filtering the results

Using LLR, 12169 words were found more often than expected in the creativity corpus. This set of 12169 words was filtered to remove words with an LLR score less than 10.83, representing a χ^2 significance value for $p = 0.001$ (at one degree of freedom). This filtering process reduced the set of words to only those that occurred significantly more often than expected in the creativity corpus.

The LLR statistic measures to what extent the observed distribution of a word in a corpus deviates from expected; however it does not indicate whether the word is more or less frequent than expected in the creativity corpus. Only those words which appear more frequently than expected in the creativity corpus were retained; the rest were discarded. To avoid extremely infrequent words disproportionately affecting the data, any word occurring fewer than five times was removed. Finally, the words were inspected to remove any spurious items such as proper nouns (author citations, for example).

These filtering steps resulted in the identification of a total of 694 *creativity words*: a collection of 389 nouns, 72 verbs, 205 adjectives and 28 adverbs that occurred significantly more often than expected in the creativity corpus. The creativity words are listed in Appendix C.

4.2.2 Clustering the creativity words semantically

694 creativity words were identified through the steps taken in Section 4.2.1. Many of the creativity words have similar meanings to each other and there may be overlap between the meanings of different words. Although a much smaller set than 12169 words, a set of 694 words is not tractable to work

[6]The creativity corpus ($\approx$ 300,000 words) and non-creativity corpus ($\approx$ 700,000 words) are small in comparison to corpora such as the Lancaster-Oslo/Bergen corpus('LOB corpus', $\approx$ 1 million words), the British National Corpus ('BNC', $\approx$ 89 million words) and more recent web-based text collections containing billions of words.

with, giving little relief from the current definitional confusion identified in Chapter 3. To extract the underlying semantic themes underpinning these 694 words, the creativity words were clustered into semantically related groups and analysed to identify distinct components of creativity.

As part of collaborative work for this project, Bill Keller provided a measure of *distributional similarity* for pairs of words in the creativity set.[7] The set of creativity words was separated into four sets: nouns, verbs, adjectives and adverbs.[8] Taking each of the four sets, for each pair of words in a set, a value, or *similarity score*, between 0.0 (not at all similar) and 1.0 (identical) was calculated.

To calculate similarities, an information-theoretic measure of similarity (Lin, 1998) was employed that has been widely adopted for calculating the similarity between a pair of words (Lin, 1998; Weeds, 2003). Lin defines similarity formally using information theory, measuring how much two things, A and B, have in common and how much overlap there is when describing A and B. In Lin (1998) this Similarity Theorem is derived and practically applied to measure similarity in different case studies:

> 'The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are.' (Lin, 1998, p. 2)

Lin (1998) provides several examples of this definition being applied to measure similarity in different case studies. In this work, the amount of information is quantitatively measured using a point-wise mutual information statistic: distributional similarity.

Distributional similarity is based on the Distributional Hypothesis (Harris, 1954) which states that two words which tend to appear in similar contexts will tend to have similar meanings. 'Context' can be interpreted in different ways, for example being in the same document, or the same sentence. In this work, 'context' is interpreted as the grammatical relations and words surrounding the target words. Similarity is based on how often two words appear in the same grammatical relationship to some other word. For example, some evidence that the words *concept* and *idea* are similar in meaning would be provided by occurrences such as the following:

(1) the *concept/idea* involves	(subject of verb 'involve')
(2) applied the *concept/idea*	(object of verb 'apply')
(3) the basic *concept/idea*	(modified by adjective 'basic')

It would be difficult to see the same pattern of occurrences in sentences for, say, the words *concept* and *chorale*. Intuitively we can understand that although *concept* and *chorale* are both nouns, and both occur in the set of creativity words, they express dissimilar things. Even though it is possible to conceive of sentences where these words both occur in the same grammatical relationship, e.g. the *concept/chorale* was interesting, this is not as simple as it was for the words *concept* and *idea*.

[7] Bill Keller's work is summarised here; Jordanous and Keller (2012) gives more details from a computational linguistics perspective.

[8] As described in Section 4.2.1, no other word types were used in this work.

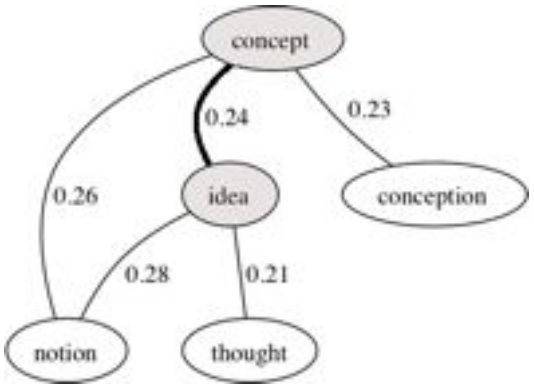


Figure 4.3: Graph representation of the similarity of the nouns *concept* and *idea* and closely semantically related words. Each word is drawn as a node in the graph, linked together by a weighted edge representing the similarity of the two words (maximum similarity strength is 1.0).

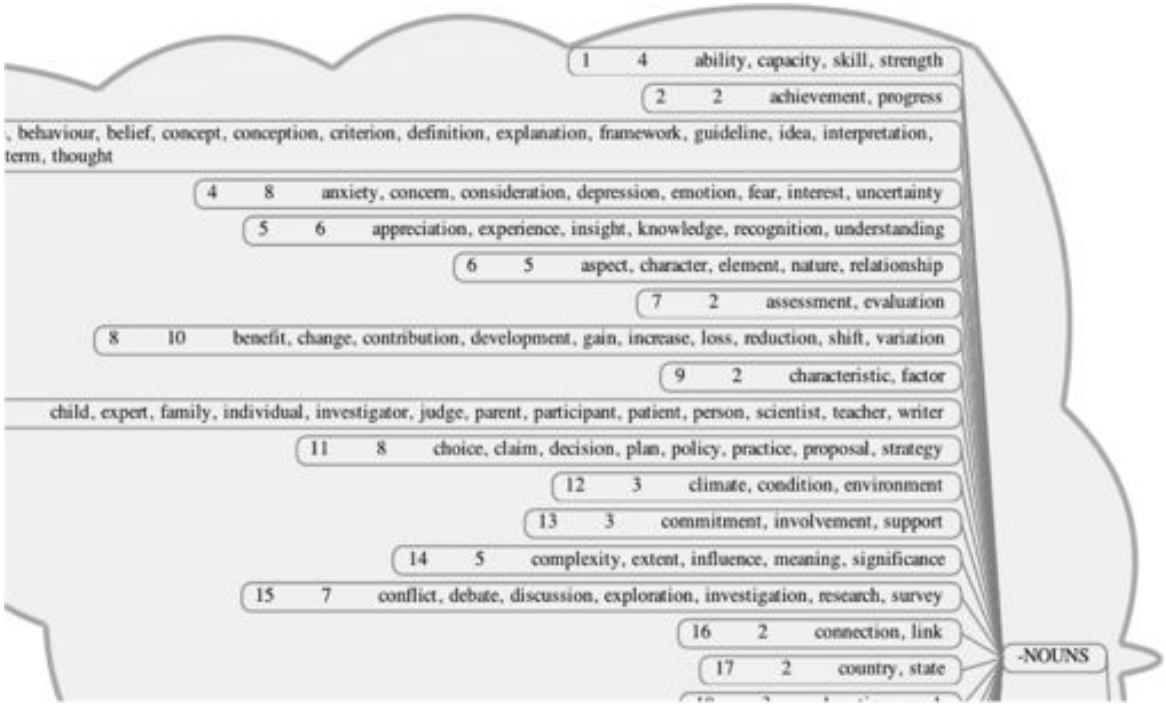


Figure 4.4: Some of the word clusters produced using the Chinese Whispers algorithm.

Using Lin's similarity measure (Lin, 1998) with the written portion of the British National Corpus (BNC Consortium, 2007) as source data for the similarity calculations, the words *concept* and *idea* scored 0.2365 for similarity (to 4 d.p.). This score is relatively high; out of 388 nouns,[9] only 3 nouns

[9]The 388 nouns exclude the word *concept* itself, which would score a similarity score of 1.0 when paired with itself.

had a similarity score with *concept* of 0.2 or greater. This similarity pair and other highly similar related words are illustrated in Figure 4.3. As expected, using the same measurements, the two words *concept* and *chorale* scored much lower for similarity, at 0.0026 (to 4 d.p.). Similarly, out of 388 nouns, *idea* is ranked 2nd in terms of similarity to *concept*,[10] whereas *chorale* is ranked 315th. Lin's similarity measure was used to calculate similarity scores between all pairs of words of the same category (nouns, verbs, adjective, adverbs) in the creativity words set.[11] The word pair similarity data for each category was represented as an edge-weighted graph as illustrated in Figure 4.3, where the nodes of the graph are words and the edges are weighted by similarity scores (similarity score > 0).

The graph clustering algorithm *Chinese Whispers* (Biemann, 2006) identified word clusters (groups of closely interconnected words) in the dataset. This algorithm iteratively groups together graph nodes that are located close to each other, as represented by edge weights. By grouping together words with similar meanings, the number of data items was reduced and themes in the data could be recognised more readily, from each distinct cluster. Figure 4.4 shows some of these clusters. The Chinese Whispers algorithm is nondeterministic, potentially generating different results each time. The algorithm was therefore run a number of times. Figure 4.5 gives an example of the clusters produced in one run.

An alternative projection of the data

To take a different view of the data and focus on the words most closely related to creativity, the 20 words with highest LLR score were each plotted as the root node on individual subgraphs, using the graph drawing software *GraphViz*.[12] For each of the 20 words, a strongly connected component graph was plotted, representing all the words connected directly to that word. To reduce the amount of data to be examined, similarity scores were discarded if they fell below a threshold value (so that each graph only contained the most strongly connected words). This kept the graphs smaller and easier to inspect and analyse visually. An example (simplified for readability) is given in Figure 4.6.

4.2.3 Analysing the results by inspection

At this point the data could be expressed in three different projections, each of which could be visually inspected. The projection of the data as word clusters, the strongly connected component word graphs described above and the original LLR scores were collated and manually analysed by inspection.

Inspecting the clusters produced as a result of Chinese Whispers, several possible groupings were identified which grouped the clusters further together to reduce numbers of clusters. Figure 4.5 shows the process of identifying links between the original clusters (on the left-hand side of the figure), indicating components (on the right-hand side) that represent a cluster or clusters at a higher level of abstraction. The graph projections of the data and the LLR scores were used to inform this process

[10] Again the word itself, *concept* is ranked first with a perfect similarity score of 1.0.

[11] So a similarity score would be calculated between two nouns, for example, but not between a noun and an adjective.

[12] `http://www.graphviz.org/`

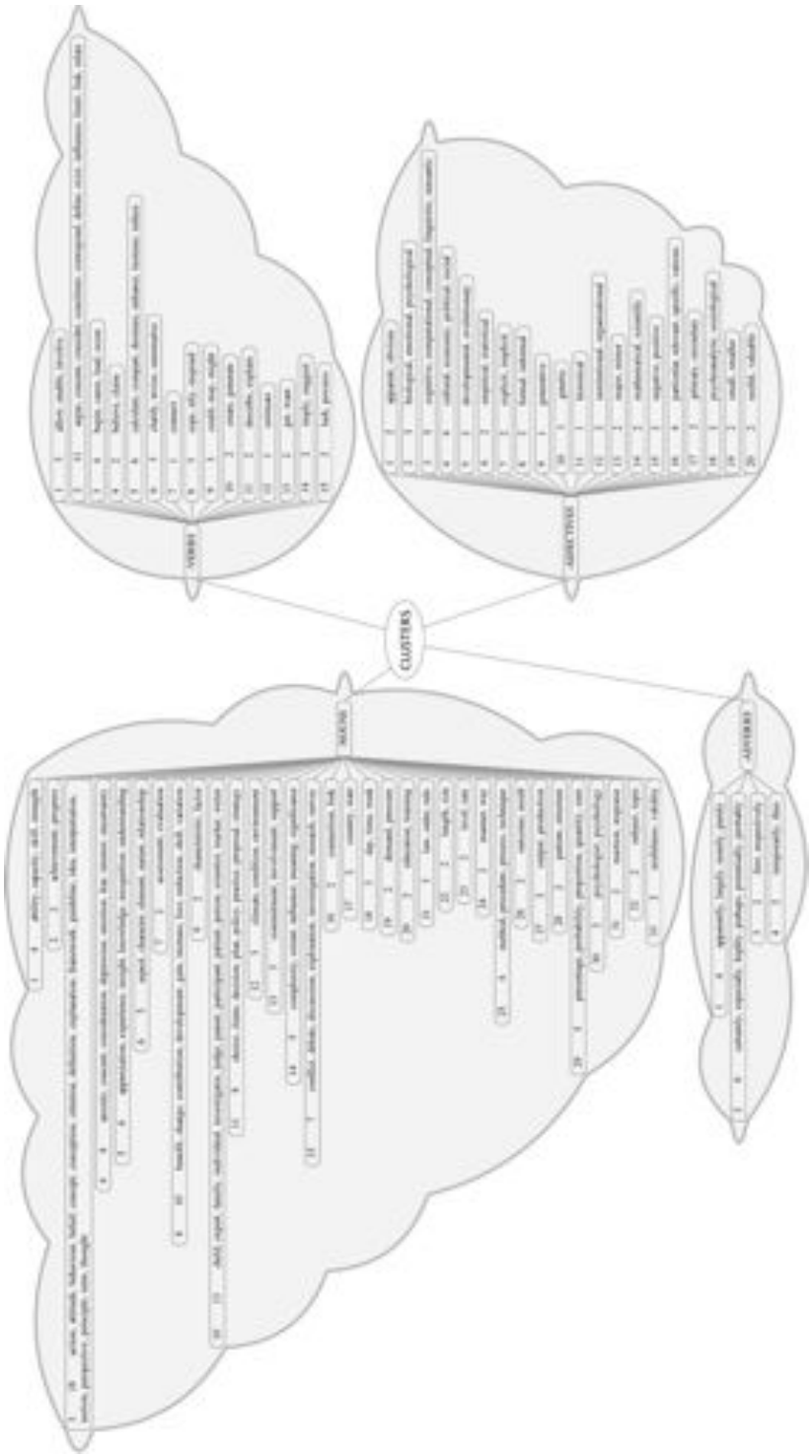


Figure 4.5: The process of manually identifying components (on the right) from clusters (on the left).

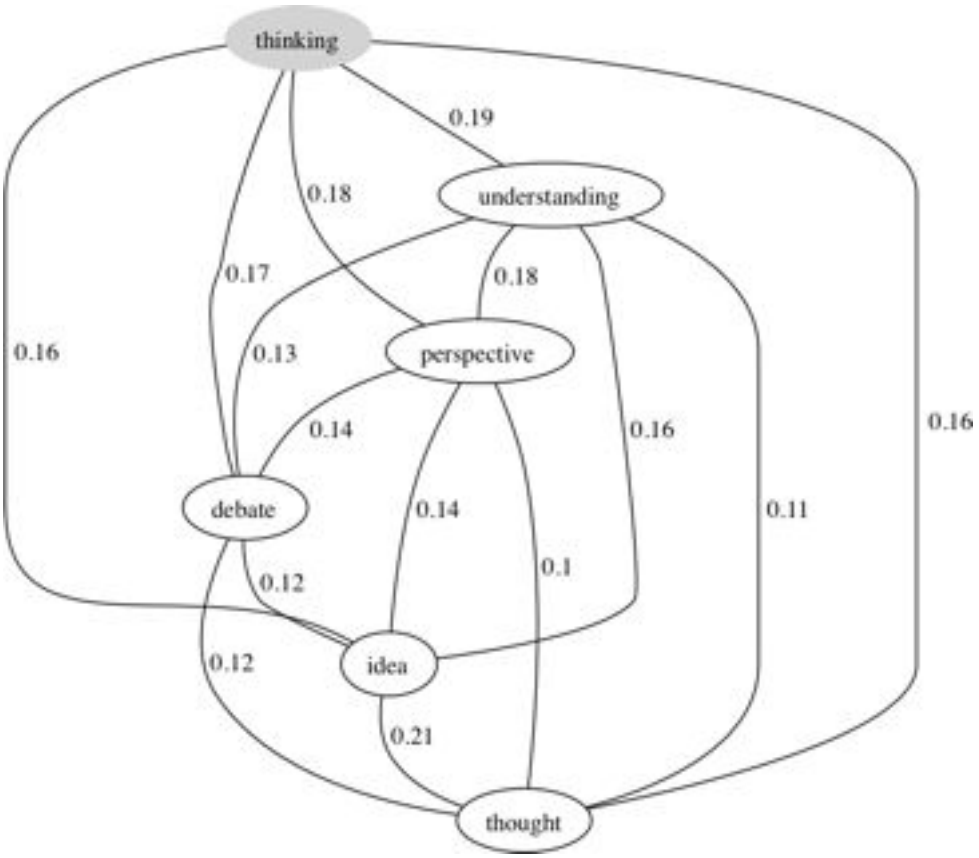


Figure 4.6: *thinking*, a high frequency word in the creativity corpus, is the root node for this graph. The other nodes represent the most semantically similar words in the creativity corpus. The edges between two nodes are labelled with the similarity score between those two words and the length of the edge is roughly proportional to that score.

and supplement the data as represented in Figure 4.5.

To see different perspectives and identify less obvious (but still important) aspects of creativity, each element in the analysis data was considered in terms of each of the *Four Ps* of creativity.[13] The *Four Ps* framework helps to highlight different aspects of creativity, to portray creativity in a fuller context. For example, *novelty* is commonly associated with end products or results of creative behaviour: how novel is the end product? A creative process can also take a novel approach, be performed by a new person or be new in a particular environment. Viewing *novelty* from the perspectives of *product*, *process*, *person* and *press* uncovers these interlinked interpretations.

In this way, components were manually identified and refined, in an iterative process of analysis through inspection. An alternative way of generating the components may have been to employ more

[13] As described in Chapter 3 Section 3.4.2: Person, Product, Process and Press.

objective methods to manipulate the results, for example using principal components analysis or more computational linguistics tools. At this point, though, the size of the results set was considered sufficiently small to be manageable for manual analysis. Using human judgement at this point may have unwittingly introduced some bias, though many efforts were made to avoid this and to be guided by the data as much as possible. It was felt, however, that the benefits of introducing human judgement at this stage would outweigh these concerns, especially given the amount of empirical processing of the data prior to this late stage. There were also a number of anomalies in the results set, as a consequence of 'noise' in the data[14] which could be identified and removed through manual inspection.

Another concern which arose during this work could be dealt with through manual inspection; the *sentiment* (positivity or negativity) attached to the words in relation to creativity. Take for example, the two phrases 'improvisation is creativity' and 'improvisation is not creativity'.[15] This would link together the two words 'improvisation' and 'creativity', despite the two different sentiments.[16] Although improvisation may be associated with creativity in one perspective but seen as distinct from creativity in another, the meanings of the two words are still connected together. This connection arises through the two words being mentioned in the same context when discussed. On occasion during the manual analysis, the original creativity corpus was consulted to investigate how some of the less obvious words were connected to creativity. Component construction was guided accordingly.

4.3 Results: Fourteen components - or *building blocks* - of creativity

Following the above analysis steps, the following set of concepts has been generated. These are proposed as a set of key *components of creativity*.

No claims are made for this set to be a necessary and sufficient definition of creativity in all possible manifestations, for two reasons. Firstly, some components are logically inconsistent with others, for example the balance between autonomous independent behaviour in *Independence and Freedom* and the reliance on information exchange in *Social Interaction and Communication*.[17] Additionally, creativity can manifest itself slightly differently across different domains (Plucker & Beghetto, 2004) and components will vary in importance, according to domain requirements. As an illustration of this second point, creative behaviour in mathematical reasoning is focused towards finding a correct solution to a problem than creative behaviour (Colton, 2008b), but in contrast, such a focus is less important in musical improvisation creativity (Jordanous & Keller, 2011).[18]

Instead, this set is presented as a collection of aspects which contribute to defining what creativity

[14] See as example 'words' such as "ativity" or "blind-variation-and-selective-retention" in the list in Appendix C.

[15] These phrases did not occur in the corpora but are used to introduce improvisation as a creative act (see Chapter 6).

[16] As mentioned earlier, function words such as 'is' are filtered out of the data.

[17] Inconsistencies and overlap between components were expected to some extent, given the subjectivity and complexity of creativity.

[18] See also Chapter 6.

is. The set items, presented in alphabetical order, should be treated as components of creativity that, in combination, act as *building blocks* for creativeness. The set is akin to an *ontology*[19] of creativity (Chandrasekaran et al., 1999): a collection of knowledge about the concept of creativity.[20]

Figure 4.7: Key components of creativity identified in the empirical work presented in this Chapter.

1. Active Involvement and Persistence
   - Being actively involved; reacting to and having a deliberate effect on a process.
   - The tenacity to persist with a process throughout, even at problematic points.
2. Generation of Results
   - Working towards some end target, or goal, or result.
   - Producing something (tangible or intangible) that previously did not exist.
3. Dealing with Uncertainty
   - Coping with incomplete, missing, inconsistent, uncertain and/or ambiguous information. Element of risk and chance, with no guarantee that problems can or will be resolved.
   - Not relying on every step of the process to be specified in detail; perhaps even avoiding routine or pre-existing methods and solutions.
4. Domain Competence
   - Domain-specific intelligence, knowledge, talent, skills, experience and expertise.
   - Knowing a domain well enough to be equipped to recognise gaps, needs or problems that need solving and to generate, validate, develop and promote new ideas in that domain.
5. General Intellect
   - General intelligence and intellectual ability.
   - Flexible and adaptable mental capacity.

[19] *Ontology* is used in the computer science sense (Chandrasekaran et al., 1999) rather than the philosophical sense.

[20] In collaborative work with Bill Keller, the components have recently been published on the Semantic Web, as a *creativity ontology* (Jordanous & Keller, 2012). The creativity ontology is available at *http://purl.org/creativity/ontology*, last accessed December 2012.

6. Independence and Freedom
    - Working independently with autonomy over actions and decisions.
    - Freedom to work without being bound to pre-existing solutions, processes or biases; perhaps challenging cultural or domain norms.
7. Intention and Emotional Involvement
    - Personal and emotional investment, immersion, self-expression, involvement in a process.
    - Intention and desire to perform a task, a positive process giving fulfilment and enjoyment.
8. Originality
    - Novelty and originality - a new product, or doing something in a new way, or seeing new links and relations between previously unassociated concepts.
    - Results that are unpredictable, unexpected, surprising, unusual, out of the ordinary.
9. Progression and Development
    - Movement, advancement, evolution and development during a process.
    - Whilst progress may or may not be linear, and an actual end goal may be only loosely specified (if at all), the entire process should represent some developmental progression in a particular domain or task.
10. Social Interaction and Communication
    - Communicating and promoting work to others in a persuasive, positive manner.
    - Mutual influence, feedback, sharing and collaboration between society and individual.
11. Spontaneity / Subconscious Processing
    - No need to be in control of the whole process - thoughts and activities may inform a process subconsciously without being fully accessible for conscious analysis.
    - Being able to react quickly and spontaneously during a process when appropriate, without needing to spend time thinking about options too much.
12. Thinking and Evaluation
    - Consciously evaluating several options to recognise potential value in each and identify the best option, using reasoning and good judgment.
    - Proactively selecting a decided choice from possible options, without allowing the process to stagnate under indecision.
13. Value
    - Making a useful contribution that is valued by others and recognised as an influential achievement; perceived as special; ‘not just something anybody would have done’.
    - End product is relevant and appropriate to the domain being worked in.

14. Variety, Divergence and Experimentation
    - Generating a variety of different ideas to compare and choose from, with the flexibility to be open to several perspectives and to experiment with different options without bias.
    - Multi-tasking during a process.

4.4 Summary

The goal of this work is an evaluation methodology for computational creativity. As shown in Chapter 3, creativity is complex to define; however a rigorous and comparative evaluation process needs clear standards to use as guidelines or benchmarks (Torrance, 1988; Kaufman, 2009).[21] This current Chapter 4 reported how techniques from the field of computational linguistics have been used to empirically derive a collection of *components* of creativity. These components represent a comprehensive, multi-perspective definition of creativity.

The set of components was generated through an analysis of what is considered important in talking about creativity. Words were identified which appear significantly often in connection with discussions of creativity. To identify these words, the log likelihood ratio statistic (Section 4.2.1) was used to compare two corpora: a *creativity corpus* consisting of thirty academic papers about creativity (Section 4.2.1 and a *non-creativity corpus* consisting of sixty academic papers matched to the creativity corpus by year and subject area, on subjects unrelated to creativity (Section 4.2.1).

Lin's semantic similarity measure (Lin, 1998) and the Chinese Whispers clustering algorithm (Biemann, 2006) were used to cluster the resulting words according to similarity, as reported in Section 4.2.2. Through these clusters and similarity data, a number of distinct themes emerged. The themes identified in this linguistic analysis have collectively provided a clearer 'working' understanding of creativity, in the form of components that collectively contribute to our understanding of what creativity is. Together these components act as building blocks for creativity, each contributing to the overall presence of creativity; individually they make creativity more tractable and easier to understand by breaking down this seemingly impenetrable concept into constituent parts.

The components can be used as several aspects with which to examine computational creativity systems by, for a more informed and detailed evaluation of how creative these systems are. Crucially, this level of detail can help us identify what aspects a system is creative and how the system's perceived creativity can be improved. Chapter 5 will investigate the use of the components as evaluation standards for computational creativity.

[21] Chapter 8 Section 8.1.1 will report a survey asking people to evaluate the creativity of musical improvisation computer systems, where several of the participants called for the term 'creativity' to be defined before they could feel comfortable and confident evaluating for creativity. This goes against the general assumption that people have a working intuitive sense of what creativity is, at least in the context of judging and assigning creativeness.

Chapter 5

Operationalising tests of creativity

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012) and a peer-reviewed conference paper (Jordanous, 2011a).

Figure 5.1: *Wordle* word cloud of this Chapter's content

Overview

This Chapter aims to identify a more systematic and rigorous approach to computational creativity evaluation. The point is to understand to a greater level of detail exactly why a system can justifiably be described as creative and to inform future development work on that system.

Section 5.1 looks at the main points of contention and practical issues when evaluating computational creativity, considering how existing methodologies address these issues. Any negative preconceptions about computational creativity need to be overcome if a system is to be fairly evaluated (Section 5.1.1). Questions arise about what to evaluate (Section 5.1.2), what interpretations of creativity should be used (Section 5.1.3), who should perform the evaluation (Sections 5.1.4 and 5.1.5), when evaluation should be performed (Section 5.1.6) and what types of tests should be used (Section 5.1.7). There is also a distinction to be drawn between the aim of evaluating creativity or evaluating quality (Section 5.1.8); as the survey of evaluative practice in Chapter 2 Section 2.3 showed, these aims have become blurred to some extent. Following from the Section 5.1 discussions, Section 5.2 proposes Evaluation Guidelines for computational creativity evaluation that some of these issues using heuristics.

Steps to implement the heuristics in the *Evaluation Guidelines* are considered in Section 5.3. The SPECS methodology is derived as a result: the *Standardised Procedure for Evaluating Creative Systems*. Section 5.3 presents the SPECS methodology and accompanying text discussing implementation details. Section 5.4 presents decision-tree-style diagrams offering practical guidance on applying SPECS in various scenarios (Figures 5.6(a) - 5.6(g)).

SPECS is presented as the methodological solution to the question: How should we evaluate the creativity of a computational system? As Section 5.5 discusses, SPECS is related to scientific method in some ways but there are clear ways in which SPECS handles the specific requirements of computational creativity evaluation more appropriately than is possible with scientific method.

5.1 Practical issues involved in evaluating computational creativity

Computational creativity evaluation is affected by a number of different issues and challenges, which have hindered the acceptance of any one evaluation methodology as standard. Many of these issues are a matter of significant, long-standing and/or continuous debate both in the computational creativity community and in research on human creativity. Chapter 3 examined the definitional problems in creativity in depth, leading to the derivation in Chapter 4 of a set of components that can be used to collectively construct a working definition of creativity. This Section discusses some other important practical issues, in the context of existing computational creativity evaluation methodologies as reported in Chapter 2 and also from a wider perspective, that guide the direction of this work.

5.1.1 Overcoming negative preconceptions about computational creativity

Cardoso et al. stress in their review of computational creativity progress that creativity is 'in the eye of the beholder' (Cardoso et al., 2009, p. 17), as does Widmer et al. (2009). As emphasised in the Four Ps approach to creativity (Chapter 3 Section 3.4.2), the opinions of the audience are crucial in making, distributing and maintaining creativity judgements.

As is examined in greater detail in Chapters 1 and 8, people's evaluation of computational creativity can be influenced by preconceived notions and beliefs. This perception has been discussed from several different perspectives (e.g. Minsky, 1982; Cohen, 1999; Pearce & Wiggins, 2001; Boden, 2004; Colton, 2008b, 2008a; Nake, 2009; Reffin-Smith, 2010). People may be reluctant to accept the concept of computers being creative, either through conscious reticence or subconscious bias (Moffat & Kelly, 2006). On the other hand, researchers keen to embrace computational creativity may be positively influenced towards assigning a computational system more credit for creativity than it perhaps deserves. Hence our ability to evaluate creative systems objectively can be significantly affected once we know (or suspect) we are evaluating a computer rather than a human.

Perception of the creativity of a computer system is a recurring point throughout Colton (2008b), primarily with respect to the negative preconceptions that people may have about computational creativity.[1] Colton offers his framework as a means for people to discuss the creativity of a computer system, free of any restrictions from what Colton describes as:

> 'the default position in the popular perception of machines that software cannot be creative' (Colton, 2008b, p. 18)

While Colton (2008b) does not address the problem of overcoming negative prejudices, he does provide the creative tripod framework for description of creativity by those who are receptive to the idea of computational creativity. To add to this tool he recommends that computational systems should be described in high-level terminology rather than at algorithmic level or in source code, to avoid arousing preconceptions about computer creativity developing. This is the approach Colton is taking with his *Painting Fool* artist system,[2] which he hopes will be considered as an artist in its own right rather than as a computer system which is programmed to generate images (Colton, 2008b, 2008c). This approach has its advantages but there are implications as to whether a system presented solely in this way could be replicated, as is expected in scientific work.

5.1.2 The product/process debate in computational creativity evaluation

> 'As a research community, we have largely focussed on assessment of creativity via assessment of the artefacts produced.' (Colton, 2008b, p. 1)

[1] Section 1.4 of Chapter 1 reflects on this in more detail.

[2] `http://www.thepaintingfool.com`

One important debate in computational creativity evaluation is about whether evaluation of a creative system should focus exclusively on the output produced by the system, or whether the processes built into the system should also be taken into account. Of the major contributors to this debate, Colton (2008b) and Pease et al. (2001) argue that both product and process should be included in evaluation, whereas (Ritchie, 2007) concentrates solely on the product of systems, stating that examining the process is unimportant for creativity (Chapter 2).

Ritchie argues that humans normally judge the creativity of others by what they produce, because one cannot easily observe the underlying process of human creativity. Ritchie therefore advocates a black-box testing approach, where the inner program workings are treated as unknown and evaluation concentrates on the system's results. While it is true that we can only use the material we have available to form an evaluation, the evaluation experiments in Pearce and Wiggins (2001) show that people often make assumptions about process in their judgements on product.[3] Pearce and Wiggins discuss how our interpretation of how something was produced is important, even if the actual method is unknown, and that such an interpretation can be derived if people are repeatedly exposed to the compositional systems (human or computational) that they are evaluating. Collins (2008) discusses how making reasonable assumptions can assist the reverse-engineering[4] of program code from output, in scenarios where white-box testing (evaluation with access to the program code) is not possible.

Colton (2008b) acknowledges Ritchie's arguments but quotes examples from art to demonstrate that process is as important as the end product when evaluating creativity, at least in the artistic domain. As evidence, Colton cites conceptual art,[5] where the concepts and motivations behind the artistic process are a significant contribution of the artwork.

Colton (2008b) also poses a thought experiment that considers two near-identical paintings presented at an exhibition. In the first painting, the dots are placed randomly, whereas in the second, the dots' locations represent the artist's friendships with various people. Colton argues that the second painting would be more appealing to purchase than the first, even though the end product is very similar, due to the process by which it was created. This thought experiment illustrates how process can impact on our judgement of creative artefacts, though it is questionable whether this thought experiment describes perception of creativity, or of quality or appeal.

The thought experiment described in Ventura (2008) gives further evidence on how knowledge of the creative process affects how we evaluate creativity.[6] The RASTER and iRASTER systems (see Chapter 2 Section 2.1.6) were designed by Ventura to be decidedly non-creative. If these systems were implemented and their generated images were given to people to evaluate without telling the

[3]This finding was replicated in the Chapter 8 Section 8.2.1 survey carried out to obtain ratings of typicality and value of improvised music for Case Study 1.

[4]Reverse-engineering is the process of identifying (and possibly replicating) how a product is generated, through analysis of that product.

[5]See the panel *What is Conceptual Art?* for details on conceptual art in the context of this debate.

[6]This point was not made explicitly in Ventura (2008), however.

What is Conceptual Art?

Sol LeWitt defined Conceptual Art (LeWitt, 1967) as an art form where 'the idea or concept is the most important aspect of the work. When an artist uses a conceptual form of art, it means that all of the planning and decisions are made beforehand and the execution is a perfunctory affair. The idea becomes a machine that makes the art.'

If assessing how creative a piece of conceptual art is, solely by evaluating the product, then there are two negative consequences:
1. The primary intentions of the artist are ignored (the artist's focus is on how the art is made rather than what the result is).
2. The level of creativity presented will probably be underestimated, especially if the art results in producing something that might seem commonplace outside the context of that art installation.

Two examples are Tracey Emin's controversial exhibit *My Bed* (1999), pictured in Figure 5.2 and Duchamp's Fountain (1917), pictured in Figure 5.3.

evaluators how they were produced, the evaluators may well rate the creativity of the system highly. Supplying the evaluators with details of how a program works, though, could have a detrimental impact on the subsequent evaluations (Cope, 2005; Colton, 2008b).

Ritchie (2008) concedes that it can be important to consider the system's processes:

> 'If a humour-generating program is needed for some practical purpose, then how it is constructed may be of little theoretical interest. However, in the case of more theoretical research, where the program is there to test an abstract model, the mechanisms are the whole point.' (Ritchie, 2008, p. 147)

One issue with creativity can be thought of as analogous to the adage that a magician never reveals their secrets. This adage is based on the fact that tricks do not appear so impressive once you have found out how the magician performed the trick. Similarly things can appear to be less creative when you know how they were produced:[7]

> 'it is not unknown for critics of AI to refuse to accept programs as creative (or intelligent) once the mundane mechanistic nature of the inner workings are revealed' (Ritchie, 2001, p. 4)

As mentioned above, Colton (2008b) intentionally avoids this by reporting on his artistic system in high-level terms only, rather than giving details of the program (Colton, 2008b, p. 8).

A final point to consider in this debate comes from a definition of creativity as consisting of the Four Ps: *process*, *product*, *person* and *press* (or environment) (Chapter 3 Section 3.4.2). Current computational creativity evaluation methodologies either look solely at a system's *products* or at a combination of the *products* and the *process*, with the possible exception of Colton (2008b) whose

[7] If the inner workings of a program are very impressive, complex or novel, then we may still be impressed by the program, but this is a different perspective to whether or not we think the program is creative.

Figure 5.2: Tracey Emin's *My Bed* (1999). The intended significance of this piece was not the end result of displaying a unmade bed in a cluttered bedroom, but to document Emin's troubled state of mind and lifestyle at the time of creating this work. Image retrieved (28th July 2010) from `http://www.saatchi-gallery.co.uk/artists/artpages/tracey_emin_my_bed.htm`.

evaluative framework is influenced by how an audience perceives the creativity of a system. Chapter 3 considers observations about the creative *person* operating in a press/environment, seeing how such observations are relevant for evaluation purposes.

5.1.3 Boden's distinction between P-creativity and H-creativity

Boden (1990) distinguishes between *P-creativity* (psychological creativity) and *H-creativity* (historical creativity). This distinction is to do with the context in which creative products are considered novel. If a creator produces an artefact which is unknown to the creator but had already been discovered elsewhere, then this would be deemed P-creative. If the creative artefact is unknown not only to the creator, but has never been discovered before by anyone, then this is H-creative. A number of researchers have commented on the need to clarify which type of creativity is being addressed (e.g. Pearce & Wiggins, 2001; Ritchie, 2007, also Edmonds, 2009, personal communications).

The computational creativity community has focused on P-creativity, mainly for practical reasons; it is far less problematic to compare output against knowledge accessed by the system than to compare system output against the sum total of all knowledge in the world.

As mentioned in Chapter 3 Section 3.6.3, Pearce and Wiggins justify a focus on P-creativity through noting that H-creativity can be thought of as a subset of P-creativity, where set membership is determined by historical and social factors. In concentrating on P-creative achievements, by

definition this would include H-creative achievements as well. As Boden expresses:

> 'the point at issue here is not "Who thought of X first?" but "Is X a creative idea, and if so, how creative is it?" Our concern ... must be with P-creativity in general, of which H-creativity is a special case.' (Boden, 1994a, p. 112)

5.1.4 Practical issues in using human judges

Figure 5.3: Duchamp's *Fountain* (1917). Duchamp submitted the Fountain to an art exhibition for the Society of Independent Artists, where every submission would be accepted and exhibited. The Fountain sparked debate with the judging panel (of which Duchamp was himself a member) as to whether it should be exhibited; as a result the Fountain was included in the exhibition but hidden from sight (and Duchamp resigned from the Society). Since then, the Fountain has been judged the most influential modern art work of all time, according to a poll of 500 art experts commissioned as part of the Turner Prize 2004. Image retrieved (28th July 2010) from `http://arthistory.about.com/od/dada/ig/DadaatMoMANewYork/dada_newyork_07.htm`.

To cope with subjectivity and changes over time in definitions (Chapter 3), human standards should be incorporated in evaluation. Incorporating human opinion in evaluation raises various practical issues to resolve. A seemingly simple method of evaluating creativity is to ask people to say how creative the system is. Generally these evaluations do not provide judges with a definition of creativity but rely on the judges' own intuitive interpretation of what creativity is.

There are practical concerns which hinder us from using human judges as the sole source of evaluation of a system. Human evaluators can say whether they think something is creative but may only be able to give limited explanation of their opinions. As described in Chapter 3, it is hard to define why something is creative; this is a tacit judgement rather than one we can easily voice.[8] A more informed idea of what makes a system creative helps understand both why a system is creative and what needs to be worked on to make the system more creative.

There is a place for soliciting human opinion in creativity evaluation, not least as a simple way to consider the system's creativity in terms of those creative aspects which are overly complex to define empirically, or which are most sensitive to time and current societal context. The process of running

[8] This was a major finding in research reported in Chapter 8 Section 8.1.1.

adequate evaluation experiments with human participants, though, takes time and effort.[9]

Human opinion is variable; what one person finds creative, another may not (León & Gervás, 2010; Jennings, 2010a).[10] This may also be affected by the relative level of expertise of the judges (Section 5.1.5), previous experience of the systems or acquaintance with similar systems[11] or preconceived notions about computational creativity.[12] Therefore large numbers of participants may be needed to capture a general consensus of opinion, and large numbers do not necessarily guarantee conclusive results, as Chapter 8 Section 8.1.1 demonstrates.

In addition to the time and resources needed to devise and run suitable evaluation experiments with large numbers of people, other issues are introduced such as obtaining ethics approval, attracting enough suitable participants, costs associated with paying participants or a reliance on goodwill. Many of these issues may adversely affect the research process but are out of our direct control to resolve. Although these issues are often accepted as a natural part of the research process, that has to be dealt with, it would be useful if this outlay of research time and effort could be reduced.

5.1.5 The expertise and bias of the evaluators

In the Chapter 2 Section 2.3 survey on contemporary evaluation practice, a minority of systems (25 out of 75) were evaluated by people other than the implementors of that system. 8 of these 25 systems were evaluated by domain experts or the target users, 4 papers used evaluators with a range of expertise in that domain and 4 papers used novice evaluators only (for the remaining 9 systems, the level of expertise of those evaluating the system was left unstated).[13]

The issue of who should evaluate a system tends to be overlooked when discussing creativity evaluation methodologies. The two key contributions thus far to computational creativity evaluation (Ritchie, 2007; Colton, 2008b) do not make any recommendations on this matter, except for the underlying implication that their methodologies are tools for researchers to evaluate their own systems. Questions of impartiality arise when self-evaluating work; as mentioned earlier, both Ritchie's criteria and Colton's creative tripod can be adapted (intentionally or unintentionally) to portray an evaluation in a desired way, as shown in Ventura (2008). In cases like this, transparent methods are crucial for holding the researcher to account on their evaluative decisions.

Boden (1994a) argues that both experts and novices can make an intuitive assessment of creativity, but that experts are better placed to explain their judgement, especially if they are used to discussing or analysing that domain. Hence, she argues, their [expressed] opinions will generally be more infor-

[9] This was found to some extent in the human evaluation studies for Case Study 1, reported in Chapter 8.

[10] See also Chapter 8 Section 8.1.1.

[11] See Chapter 7 for the DARCI system.

[12] This is discussed in Chapter 1 Section 1.4. See also Chapter 8 Section 8.1.1, which investigates how people perceive computational creativity, specifically after they had evaluated the creativity of some computational systems (Section 8.1.1).

[13] Expert evaluator(s) have extensive knowledge and competence in the given creative domain. Novice evaluator(s) have only a shallow knowledge with the given domain, lacking in depth.

mative and more grounded in fact.

Availability and willingness to participate can often dictate who becomes involved in evaluating a system. Aside from issues of bias, if evaluation is done by those who implement the system, then it could be assumed that the systems are being evaluated by people with domain knowledge.[14]

Using external evaluators helps to avoid accidental or intentional bias, such as in Pearce and Wiggins (2007) or as recommended in Pease et al. (2001). Pearce and Wiggins (2007) use a version of the Consensual Assessment Technique (Amabile, 1996)[15] to obtain expert judges' evaluation of their musical system. Pease et al. (2001) advocates using external judges as part of their *Emotional Response measure* but makes no comment on the level of expertise of these judges.

5.1.6 Distinguishing between post-hoc system evaluation and system self-evaluation

In her keynote address to the AISB'99 convention on Creativity, Boden argued that work in computational creativity should be judged by the authors' consideration of how their systems evaluate their own products, referring to the problem of automatic evaluation as 'the Achilles' heel of AI research on creativity' (Boden, 1999, p. 11).

> 'To enable AI-models to evaluate their own ideas, especially those produced as a result of transformational creativity - which, by definition, breaks some of the accepted rules, is a tall order. Not because computers can't "really" value anything (that may be so, but it is a philosophical rather than a scientific point), but because we find it even more difficult to articulate our aesthetic values clearly than to define the rich conceptual spaces concerned. ... One of the dimensions on which these [AISB99] papers should be judged is the extent to which they are aware of these sorts of problem.' (Boden, 1999, p. 11)

Boden stops short of requiring that creative systems should incorporate some implementation of self-evaluation, but calls for authors to reflect upon the issues involved as part of their papers. As the Chapter 2 Section 2.3 survey shows, this call has generally not been responded to in much depth.[16]

Wiggins (2000) distinguishes between two different 'senses' of computational evaluation:

> 'The question of evaluation, highlighted in one sense by Margaret Boden at her AISB'99 conference keynote, actually has two senses: how can a computer program evaluate its output; and, how can we evaluate strategies for modelling creativity.' (Wiggins, 2000, p. iii)

A number of creative systems do include some self-evaluation as part of the creative process, particularly those using an *engagement-reflection* process or evolutionary computing approaches (see Chapter 3 for examples). A creativity evaluation methodology such as in Ritchie (2007), Colton (2008b), Pease et al. (2001), takes the second interpretation raised by Wiggins (2000): how to evaluate

[14] Boden makes the somewhat controversial statement that 'only an expert in a given domain can write interesting programs modeling that domain' (Boden, 1994a, p. 115). As pointed out by Collins (2011, personal communications) though, perhaps developing a program to model a domain helps the programmer develop expert domain knowledge.

[15] As mentioned in Chapter 3 Section 3.4.2.

[16] There are some exceptions e.g. Wiggins (2006b), or the appreciation 'leg' of Colton's creative tripod (Colton, 2008b).

the creative *strategies* used in the system. This is particularly the case should we want to compare systems against each other (as is one of the central aims in this thesis), as post-hoc evaluation allows us to compare previous systems with current work, regardless of when the system was implemented.

This thesis takes the second sense of creativity evaluation, that of post-hoc evaluation of a system, whilst acknowledging system self-evaluation as an important part of the creative process.

5.1.7 Using quantitative and qualitative methods

> 'Mapping measurements of creativity quantitatively is an attractive approach to assessing creativity.' (Haenen & Rauchas, 2006, p. 3)

Can creativity evaluation be reduced to a set of quantitative measurements? Comparing two numeric measurements is generally simpler and less open to interpretation than comparing two opinions and quantitative measurements can result in an overall creativity score, should that be desired.

The creative tripod framework in Colton (2008b) uses qualitative evaluation only; using this framework, a system is creative if it is perceived to behave skilfully and imaginatively, showing appreciation of its work. It is problematic to employ this framework to perform useful comparison between systems; if system x is skilful in one way and system y is skilful in another way, which is more skilful? Or more creative?

If wishing to evaluate to human standards, the use of (at least some) human input seems necessary, either prior to evaluation in capturing those standards or during evaluation itself through the use of human judges (Gervás, 2000; Pease et al., 2001; Pearce & Wiggins, 2001; Ritchie, 2007; Colton, 2008b; Ritchie, 2008).[17] Despite the quantitative form of Ritchie's criteria, the framework in Ritchie (2007) depends heavily on the two rating schemes of typicality and value. Although Ritchie makes very few recommendations on how to carry out these ratings, all the examples cited in Ritchie (2001, 2007) use subjective assessment by human judges to obtain the majority (or all) of this data (Binsted et al., 1997; Gervás, 2002; Pereira et al., 2005; Haenen & Rauchas, 2006; Jordanous, 2010c). Ritchie (2001) acknowledges that subjective ratings and appraisals depend on the particular judge's opinions, background knowledge and even perhaps on their current mood, but offers no practical solutions such as using a number of different judges. Qualitative data given by judges is discarded.

Most of the tests in Pease et al. (2001) are presented as formal statements similar to Ritchie's criteria, except for one test (*Emotional Response*) that uses external judges to provide qualitative evaluation. A criticism of Pease et al. (2001) is that the tests are not linked together and results are not combined (e.g. Jick, 1979; Sauro & Kindlund, 2005) for overall feedback (Ritchie, 2007).

The case studies in this thesis employ both quantitative and qualitative evaluation methods, to allow for quantitative comparison of systems whilst incorporating human judgement. Where pos-

[17] The latter option accounts for the fact that evaluation standards may alter over time, for example something may develop over time to have a greater contribution than it first appeared (Ventura, 2008; Colton et al., 2000) or may need some time to become accepted as creative (Boden, 1994a).

sible, the results of individual evaluative tests should be viewed in combination. Chapters 6 and 7 demonstrate the application of these principles.

> 'In general, one cannot assess creative ideas by a scalar metric. ... The appropriate method of assessment would have to take into account the fact that conceptual spaces are multidimensional structures, where some features are 'deeper', more influential, than others. ... To compare the degree of creativity of two ideas, we would have to weigh depth against number: novelty in one deep feature (a core dimension of the space) might outweigh several simultaneous novelties in more superficial features.' (Boden, 1994a, p. 113)

5.1.8 Evaluative aims: Creativity or quality?

The survey in Chapter 2 Section 2.3 highlights an issue that researchers face when evaluating their system. Should systems be evaluated solely on the value and correctness of their output, or should there be some assessment of the creativity demonstrated by the system (which incorporates quality judgements on the output)? Both are important, though the quality of output is often easier to define and test for, especially in the absence of a standard definition or creativity evaluation methodology.

From the 75 surveyed creative systems in Chapter 2 Section 2.3, only 35% of systems were evaluated according to how creative they were; the rest of the systems were evaluated solely by the quality of the system's performance. Two systems (Riedl, 2008; Collins et al., 2010) were described as being assessed for creativity but were actually assessed only on the accuracy of the system. Of the 18 papers making practical use of creativity evaluation methodologies such as Ritchie (2007) or Colton (2008b), only 10 papers used the methodologies for creativity evaluation, with the rest adapting the methodologies to evaluate the quality of their system output.

This shows some confusion about the distinction between creativity and quality; as Chapter 3 investigates, our interpretation of creativity includes reflections on quality but encapsulates more than just how correct or valuable the creative output is. A pertinent example of such confusion can be found in Ventura (2008) (see Chapter 2 Section 2.1.6). Ventura aims to critically analyse creativity evaluation methodologies but actually addresses quality (or 'recognisability') evaluation only.

It is not that there is no interest in techniques for evaluating the creativity of creative systems; there is plenty of evidence to the contrary in publications such as Ventura (2008), Colton (2008b), Ritchie (2007), Pease et al. (2001), Wiggins (2000), as well as positive interest in this work from several researchers.[18] From these sources and from the spread of different approaches used for creative system evaluation, the suggestion is that people evaluate systems on their quality and accuracy because an evaluation methodology for creativity has not yet emerged as an accepted standard. To date Ritchie (2007) and Colton (2008b) have come closest to providing this standard, but as Chapter 2 details, there are issues with each methodology that if left unaddressed will prevent wide-scale adoption of these approaches for practical use in evaluation.

[18] Chapter 9 Section 9.3 will discuss how earlier versions of this work have been received to date by other researchers.

5.1.9 The difficulty of evaluating creative systems for creativity

> 'As we all know, creative systems are not easy to evaluate' (Oliveira et al., 2007, p. 54) 'Comparing the behaviour of artificial creative systems is a difficult task.' (Saunders et al., 2010, p. 107)

With no universally accepted definition of creativity and no baseline to use as a basis for creativity judgements, the question of how to evaluate creative systems presents many challenges which have only just started to be addressed. As Chapter 3 investigated, there are many issues surrounding the evaluation of concept with no definitive right answer, different subjective interpretations of creativity and disciplinary research separation, most of which are not immediately resolvable.

It may be possible to borrow evaluation methods from research into human creativity, which has a longer history than computational creativity research, arguably starting from the seminal presentation in Guilford (1950) on creative personalities. To date there has been very little in the way of crossover between human and computational creativity assessment. One rare example mentioned in Chapter 3 is the use of Amabile's Consensual Assessment Technique or CAT (Amabile, 1996), where several human experts evaluate creative products. CAT was used for system evaluation by Pearce and Wiggins (2007), although as described in Chapter 2 Section 2.1.1 the authors adapted the technique to evaluate 'stylistic success' of their system, not creativity. Chapter 3 Section 3.4.2 discusses how although psychometric tests exist to assess human creativity, these tests are not generally transferable to computational subjects. They test everyday general creativity rather than specialised creative abilities; usually computational creativity systems are designed for specific tasks.

If wanting to compare systems against each other, another practical consideration arises, as to what material is available for evaluation. This matter arose in both case studies, either in obtaining information on older systems for comparative analysis or in what information was made available on current systems.[19] Similar problems have also been reported by other researchers working in evaluation. For example, Colton (2011, personal communications) noted how even though there were a number of similar systems to his own mathematical discovery system HR, it was difficult to perform any meaningful comparisons as alternative systems either worked in a different domain, or source code or output data was not available. A research talk by Robey (2011) discussed problems of availability of older programs and software.[20]

It is worth noting that if a method is over-complex, not specified clearly enough or impractical to implement, this may prejudice people against using it, especially if the return from using the method is not deemed worth the effort of using it, or if the method has not been adopted by others in that research community. For example, the lack of specific recommendations in Ritchie (2001, 2007) on criteria inclusion, weights and parameter values leaves open many questions which need to be investigated before Ritchie's criteria could be used as a standard evaluative tool. Although there has been some

[19] Chapter 9 Section 9.1.1 considers this more.

[20] The points made by Robey are returned to in Chapter 9 Section 9.1.1.

research using Ritchie's criteria with fixed parameters (Gervás, 2002; Pereira et al., 2005; Haenen & Rauchas, 2006), this approach was criticised in Ritchie (2007) for not using the full customisability of the criteria approach. This thesis takes a similarly customisable approach but offers examples and suggestions for best use of the methodology.

5.1.10 Comparing systems across different domains

The ability to compare two creative systems in terms of how creative they are has often been raised during the course of this Chapter and elsewhere (Boden, 1994a; Wiggins, 2006a; Colton, 2008a). If the two systems operate in the same domain and/or perform similar tasks, then much can be learnt from comparing the relative success of the two approaches. Systems operating in different domains may however demonstrate creativity differently.

Colton et al. (2001) warn that comparison between creative systems should only be under 'special circumstances' and even then such comparisons may be 'unfair' (Colton et al., 2001, p. 1). Ritchie gives mixed comments on this issue. He criticises Pereira et al. (2005) for their comparison of two creative systems which work in slightly different aspects of linguistic creativity (comparing a conceptual blend generator[21] to a system that generates paraphrases) for over-interpreting their findings.

> 'these conclusions may be over-interpretation of the findings, given the degree of arbitrariness in the choice of constants (α, γ) (which Pereira et al. remark on) and the various interpretations of notions such as "typicality" and "quality". As Pereira et al. comment, care is needing [sic] in deciding how to assess these basic factors.' (Ritchie, 2007, pp. 88-89)

The criteria framework is however intended as an attempt to make different domains more amenable to fair comparisons across the criteria (Ritchie, 2007). Ritchie has expressed a contrasting view to the above quote when talking about the 'intellectual apartheid' (Ritchie, 2001, p. 3) that even mediocre artistic activities can be described as creative, whereas scientific activities are only viewed as creative if done very well. Ironically, even in the midst of talking about avoiding 'prejudice in favour of "art" against "science" ', one paragraph later he discusses how the quality of a creative product is related to its ' "artistic" value' rather than using more domain-neutral terms (Ritchie, 2001, p. 4).

One point to note in this debate is that humans are capable of comparing two creative people (or systems) to some degree, even if they work in different areas. For example, it would not be unreasonable to comment that 'System *x* is more creative generally than System *y*' although we might struggle to justify our reasoning for this view.

It is true that more extensive comparison can be done if systems work in a similar domain, and that otherwise it is not so meaningful to compare systems except at a more general level.[22]

[21] Conceptual blending, the combining of two different concepts for a composite meaning to emerge in linguistic expression, is defined and explained in Fauconnier and Turner (2002)

[22] As will be seen in the two case studies in Chapters 6 and 7, especially the latter case study, such a general comparison can still be useful though, to see which systems are more creative in general and see how creativity is modelled in different

> 'In the problem solving paradigm, if a new program solves a previously unsolvable problem, or solves a bunch of problems faster than all other programs, then clear progress has been made. As creativity is such a subjective notion (is your child really as creative as you say?), it's much more difficult for us to compare the creative abilities of different programs. However, much progress has been made towards telling whether we should use the word creative to describe a program and telling whether one artefact generation program is performing more creatively than another.' (Colton, 2008a, p. 6)

5.2 Guidelines for evaluation

Being guided by the above consideration of a number of issues surrounding the application of an evaluation methodology, this Section moves towards the goal of a standardised methodology for computational creativity evaluation.

5.2.1 Preliminary conclusions and directions

The discussions in Section 5.1 and in this thesis so far lead to conclusions which guide the directions taken in the rest of this Chapter and in the thesis as a whole:

- The extent to which a system is creative is important to computational creativity researchers.
- What is needed is a methodology that can be used as a standard, practical approach to evaluate and compare computational creativity systems. This will allow the research community to measure its progress.
- Inspection of the creative process should be included in evaluation of a computational system, as well as examining the products generated by the system. Inspired by the *four Ps* of creativity (Chapter 3 Section 3.4.2), factors will also be investigated relating to the system implementation and to the environment in which the system operates.
- Both quantitative and qualitative evaluation methods should be incorporated in the overall evaluation, to allow for quantitative comparison while including human-set standards in the evaluation which may not be quantifiable.
- The evaluation approach proposed in this thesis is intended to be used after systems have been implemented, for assessment and comparison of systems. Although the ability of systems to self-evaluate their own creativity is a salient part of the creative process, and creativity evaluation methodologies could form part of a creative system,[23] self-evaluation during the creative process is not the focus of this work.
- Creativity is manifested in different ways according to the domain; it should be clear what is meant by 'creativity' in the application of the evaluation methodology.

systems and for different domains. This highlights in particular cases where it may be possible to improve a program's creativity by cross-applying approaches used by creative systems operating in a different domain.

[23] As explored in (Jordanous, 2010c) which incorporates the criteria of (Ritchie, 2007) into the fitness function of a genetic algorithm. See Chapter 6 for more details.

- Evaluation methods should be clearly stated in enough detail to be repeatable by other researchers and available for critique.

5.2.2 Formative decisions towards constructing heuristics for evaluation

Heuristics and guidelines were considered on for how to approach evaluation, in a way that clarifies what is being evaluated under the term 'creativity'. The heuristics should be applicable across a wide range of domains but customisable to reflect the specifics of the domain of the system in question.

Taking a general definition of creativity and adapting it for a particular type or domain of creativity is preferred to defining creativity from the specific perspective of a given domain, to avoid losing sight of what creativity is in general. Taking a domain-specific perspective on creativity means running the risk of over-specialising to that domain. The Chapter 2 Section 2.3 survey of current evaluative practice shows the tendency to move evaluative focus away from creativity to domain value.

Knowledge of a domain is important, however, in terms of understanding it well enough to model creativeness in that domain. Whilst not going as far as Boden's belief that 'only an expert in a given domain can write interesting programs modeling that domain'(Boden, 1994b, p. 115), the Evaluation Guidelines assert that those building a system should be appreciative of requirements for creativity in that domain, and should evaluate their systems accordingly.

Depending on how creativity is defined by the researcher(s), previous evaluation frameworks (Ritchie, 2007; Colton, 2008b; Pease et al., 2001, and other discussions) may be accommodated if appropriate for the standards by which the system is being evaluated. For example if skill, appreciation and imagination are identified as key components of creativity for a creative system or domain, it would be appropriate to use the creative tripod (Colton, 2008b).

The evaluator should be able to choose the most appropriate existing evaluation suggestions, as long as their choices are clearly documented. This is preferable to the evaluator being tied to a fixed definition of creativity that may not apply fully in the domain they work in. At this point no recommendations are made on what tests to include; this will be considered further in Section 5.3.3. What is emphasised here is that for systematic evaluation we must clearly justify claims for the success or otherwise of research achievements. This approach affords such clarity.

It should be reiterated here that this thesis does not actively pursue the formation of a 'creativity function' or measurement system that produces a creativity score or rating; such a measurement system does not exist in a satisfactory format for creativity of any kind, human or computational. Instead, taking treatment of human creativity as a guide, creativity evaluations should be relative and comparative, rather than meeting some arbitrary absolute threshold values. The evaluation methodology should be able to provide information towards answering the following questions:

- Is this system creative? Why?

- Is this system more creative than other similar[24] systems? In what ways?
- How successfully does this system demonstrate creativity as expected in that domain?
- How could the creativity of this system be developed and improved upon?

5.2.3 Evaluation Guidelines for recommended evaluative practice

In addressing the issues raised in the survey, three *Evaluation Guidelines* are proposed here as heuristics for good evaluative practice to follow in computational creativity evaluation:

1. Identify a definition of creativity that your system should satisfy if it is to be considered creative:
   (a) What does it mean to be creative in a general context, independent of any domain specifics?
   (b) What aspects of creativity are particularly important in the domain your system works in (and conversely, what aspects of creativity are less important in that domain)?
2. Using step 1, clearly state what standards you use to evaluate the creativity of your system.
3. Test your creative system against the standards stated in step 2 and report the results.

The intention of the Evaluation Guidelines

The Evaluation Guidelines guide us in investigating in what ways a system is being creative and how research is progressing in this area, using an informed, multi-faceted approach that suits the nature of creativity. The Evaluation Guidelines allow comparison between a creative system and other similar systems, by using the same evaluation standards. A clear statement of evaluation criteria makes the evaluation process more transparent and makes the evaluation criteria available to other researchers, avoiding unnecessary duplication of effort.

There is a time-specific element here; a creative system is evaluated according to standards at that point in time, where a creative domain is at a certain state, viewed by society in a certain context.[25] Hence detailed comparisons can be made using each standard, to identify areas of progress.

This is not an attempt to offer a single, all-encompassing definition of creativity, nor a unit of measurement for creativity where one system may score x %. The Evaluation Guidelines are not intended as a measurement system that finds the most creative system, or gives a single summative rating for the creativity system (though people may choose to use and adopt the approach for these purposes if it is relevant in their domain). Such a scenario is usually impractical for creativity, both human and computational. There is little value in giving a definitive rating of computational creativity, especially as we would be unlikely to encounter such a rating for human creativity. Nor is this a suggestion of how to derive a single necessary and sufficient definition of creativity across all domains, especially as decades of creativity have so far failed to identify such a definition.

[24] Intuitively, it is more appropriate to compare two systems in the same domain. As domains become less related, the comparisons become less meaningful because there is less common ground on which to base the comparisons.

[25] It should be borne in mind, that these standards may change over time.

This approach is in no way intended to hinder researchers such that they are forced to target other goals and justifications for their research rather than the pursuit of making computers creative. For those researchers whose intention is to implement a computer system which is creative, the approach outlined here offers a methodological tool to *assist* progress.

5.3 From guidelines to the SPECS methodology

The Evaluation Guidelines give heuristics for how best to perform creativity evaluation. These heuristics can now be translated into a set of methodological steps to perform, to apply the Evaluation Guidelines for evaluation. These methodological steps are collectively referred to as the *SPECS* methodology: the *Standardised Procedure for Evaluating Creative Systems*. Each step is introduced and discussed individually and the full version of the methodology is presented in Table 5.1.[26]

5.3.1 Step 1: Defining creativity

Identify a definition of creativity that your system should satisfy to be considered creative:

(a) What does it mean to be creative in a general context, independent of any domain specifics?

(b) What aspects of creativity are particularly important in the domain your system works in (and what aspects of creativity are less important in that domain)?

Identifying general aspects of creativity

Step 1a of SPECS requires a general definition of creativity. As Chapter 3 notes, there are a plethora of definitions of creativity and choosing an appropriate definition is non-trivial.

In Chapter 4, a general and interdisciplinary definition of creativity is derived in the form of a set of 14 components that act as building blocks for creativity.[27] These components represent common themes across creativity in general and can be used to fulfil the requirements of Step 1a.

Figure 5.6(c) of Section 5.4 illustrates how to approach Step 1a and offers practical guidance to take the evaluator through this step.

Identifying the relative contributions of different aspects in a creative domain

While some aspects of creativity are shared by all types of creativity, some aspects of creativity will be prioritised (or de-prioritised) more in some domains than others (Chapter 3 Section 3.6.4). For Step 1b, the researcher needs to be aware of how creativity is demonstrated in the creative domain they focus on, adjusting the definition from Step 1a accordingly.

It is recommended that for Step 1a, the components of creativity from Chapter 4 are used. If this is the case, the researcher should investigate the relative important of each component in their

[26] The next Section, Section 5.4, gives a practical 'walkthrough' of the steps of SPECS, in the form of interlinked decision-tree-style diagrams that illustrate different paths through the application of SPECS.

[27] See Chapter 4, especially Figure 4.3.

Table 5.1: SPECS, the Standardised Procedure for Evaluating Creative Systems

1. **Identify a definition of creativity that your system should satisfy to be considered creative:**
   (a) What does it mean to be creative in a general context, independent of any domain specifics?
      - Research and identify a definition of creativity that you feel offers the most suitable definition of creativity.
      - The 14 components of creativity identified in Chapter 4 are strongly suggested as a collective definition of creativity.

   (b) What aspects of creativity are particularly important in the domain your system works in (and what aspects of creativity are less important in that domain)?
      - Adapt the general definition of creativity from Step 1a so that it accurately reflects how creativity is manifested in the domain your system works in.
2. **Using Step 1, clearly state what standards you use to evaluate the creativity of your system.**
   - Identify the criteria for creativity included in the definition from Step 1 (a and b) and extract them from the definition, expressing each criterion as a separate standard to be tested.
   - If using Chapter 4's components of creativity, as is strongly recommended, then each component becomes one standard to be tested on the system.
3. **Test your creative system against the standards stated in Step 2 and report the results.**
   - For each standard stated in Step 2, devise test(s) to evaluate the system's performance against that standard.
   - The choice of tests to be used is left up to the choice of the individual researcher or research team.
   - Consider the test results in terms of how important the associated aspect of creativity is in that domain, with more important aspects of creativity being given greater consideration than less important aspects. It is not necessary, however, to combine all the test results into one aggregate score of creativity.

particular domain, and weight the contribution of each component accordingly. This may be done by quantifying the importance of each component.[28] Alternatively, the components could be categorised according to level of importance for that domain.[29]

The importance of each component can be investigated in many ways, such as consulting the

[28] In Case Study 1 (Chapter 6), components will be weighted by how often they were mentioned in written discussions about musical improvisation.

[29] This will be demonstrated in Case Study 2 (Chapter 7), where components were classified either as *Crucial for creativity*, *Quite important*, *A little important* or *Not at all important*.

opinion of experts and/or the general public, analysing prior research or consulting general knowledge about that field. The researcher is advised to support their choices using relevant knowledge. Figure 5.6(d) of Section 5.4 illustrates how to approach Step 1b, again offering practical guidance to help the evaluator perform this step.

As the Chapter 4 components were derived from general discussions about creativity, and identify common general themes across creativity, it is strongly recommended that the components be used for Step 1a of SPECS (what does it mean to be creative in general). The components can then be customised according to importance during Step 1b (what is more/less contributory to creativity in a specific creative domain of interest). If one chooses a different interpretation of creativity, this choice should be clearly stated and justified as to why it forms a base definition of creativity, both in general and for the particular domain of interest.

Figures 5.6(b) of Section 5.4 gives practical guidance through the performing of Step 1 for evaluation, linking to Figures 5.6(c), 5.6(d) and 5.6(e) for further illustration of (respectively) Steps 1a and 1b, plus some examples of creativity definitions that would not be suitable for use as the underlying model of creativity for Step 1.

5.3.2 Step 2: Identifying standards to test the systems' creativity

Using step 1, clearly state what standards you use to evaluate the creativity of your system.

For Step 2 of SPECS, the definition of creativity from Step 1 is transformed into an equivalent (or as close as possible) set of standards for testing the system. If using the 14 components of creativity from Chapter 4, as is strongly recommended in this thesis, each component becomes an aspect of the system to be tested. Little analysis is required here, except perhaps to re-express the components in a form more relevant to that particular domain, as is done in Case Study 1 (Chapter 6).

Further analysis may be required, if using other definitions, to convert the definition into standards for testing. In prose definitions, the conversion from definition to standards is not so direct. Take for instance 'Creativity is the ability to come up with ideas or artefacts that are *new*, *surprising and valuable*' (Boden, 2004, p.1): does Boden require a system to actually produce these ideas/artefacts before it can be deemed creative, or merely have the ability to do so? Careful analysis of the specific definition is needed.

No further detail is given in this present discussion on heuristics for this conversion; this process will be specific to the definition used. Instead this acts as another reason for recommending the Chapter 4 components for Step 1. Figure 5.6(f) of Section 5.4 gives guidance on approaching Step 2 and provides specific examples on how to extract evaluation standards from different types of creativity definitions.

5.3.3 Step 3: Testing systems using the components

Test your creative system against the standards stated in step 2 and report the results.

The choice of what tests to use for evaluation is heavily dependent on the standards chosen to be tested and the preferences, capabilities and equipment/facilities of the researcher(s) involved. Section 5.1 discussed several issues surrounding evaluation of computational creativity, which should be taken into account at this stage:

- The product/process debate in computational creativity evaluation.
- Boden's distinction between P-creativity and H-creativity.
- Practical issues in using human judges.
- The expertise and bias of the evaluators.
- Using quantitative and qualitative methods.

Other issues are also discussed in Section 5.1, however these are addressed or partially solved through SPECS and other findings in this thesis:

- Post-hoc system evaluation or system self-evaluation? - *post-hoc evaluation*.
- Evaluative aims: creativity or quality? - *creativity*.
- The difficulty of evaluating creative systems for creativity - *using the SPECS methodology*.
- Comparing systems across different domains - *define creativity with reference to the domain being worked in, and look for common priorities between domains*.

Figure 5.6(g) of Section 5.4 provides guidance on how to approach and carry out Step 3, leading to the generation of evaluation results based on findings from Steps 1 and 2.

As the purpose of SPECS is to provide detailed feedback on system performance rather than an overall 'creativity score', it is not necessary to recombine the test results for a single aggregated measure of the whole system. It is important, however, is to consider individual test results relevant to individual aspects' importance, with results for more important aspects for a particular domain being given more emphasis than less important aspects.[30]

Identifying the coverage of components in existing creativity tests

To assist the researcher in choosing tasks, it is useful to see how the components have been tested previously. The evaluative tests used or discussed in the surveyed papers were collated during the survey of evaluation methods in Chapter 2 Section 2.3. These tests were annotated according to what component or components they covered. This included all evaluative tests performed or discussed, whether they evaluated creativity or other aspects of the system. Table 5.2 and Figure 5.4 give details of how often each component was considered in evaluative tests in the survey.

Looking at these results, it is clear that *Domain Competence* and *Value* are prioritised in testing, far above the other 12 components, each appearing in more than half of the papers surveyed. This is

[30] Two ways of identifying and weighting component importance shall be explored in Chapters 6 and 7.

Table 5.2: A count, for each component, of papers in the Chapter 2 survey (total 75 papers) that tested or discussed their creative system's performance on that particular component. The components were reviewed in a mean of 11.6 papers with a standard deviation of 14.4.

Component	Occurrences in survey
Domain Competence	47
Value	38
Originality	18
Thinking and Evaluation	13
Variety, Divergence and Experimentation	13
Social Interaction and Communication	9
Progression and Development	8
Generation of Results	5
Intention and Emotional Involvement	4
Dealing with Uncertainty	2
Independence and Freedom	2
Spontaneity / Subconscious Processing	2
General Intellect	1
Active Involvement and Persistence	0

as expected given the findings found in the Chapter 2 Section 2.3 survey, that much evaluation examined the worth of the system or its technical proficiency, rather than creativity. Tests for *Originality* appeared fairly frequently, in 18 out of 75 papers, although the emphasis on originality and novelty in creativity (Chapter 3 Section 3.4.1) alongside value suggests that tests for *Originality* would be thought more important than this figure implies.

Theories of creativity and implementation mechanisms played their part in various components being discussed. *Variety, Divergence and Experimentation* was discussed in 13 out of the 75 evaluated systems, boosted by some employment of theories of conceptual space exploration (Boden, 2004) and the view of creativity as search (see Chapter 3 Section 3.4.1). *Thinking and Evaluation* appeared in 13 papers, influenced by a number of systems employing an evaluation-reflection process (also covered in Chapter 3 Section 3.4.1). 9 evaluations tracked the system's performance for *Social Interaction and Communication* and 8 for *Progression and Development*, often reflecting the use of agent-based and evolutionary methods respectively.

Somewhat surprisingly given emphasis on the end product by research such as Ritchie (2007),[31] only 5 papers measured their system's productivity at generating end products (*Generation of Results*). The ability to produce results was assumed implicitly in several evaluation methods but not often considered directly as a measure of performance. *Active Involvement and Persistence* was not directly included in any evaluations.

[31] See Chapter 2 Section 5.1.2.

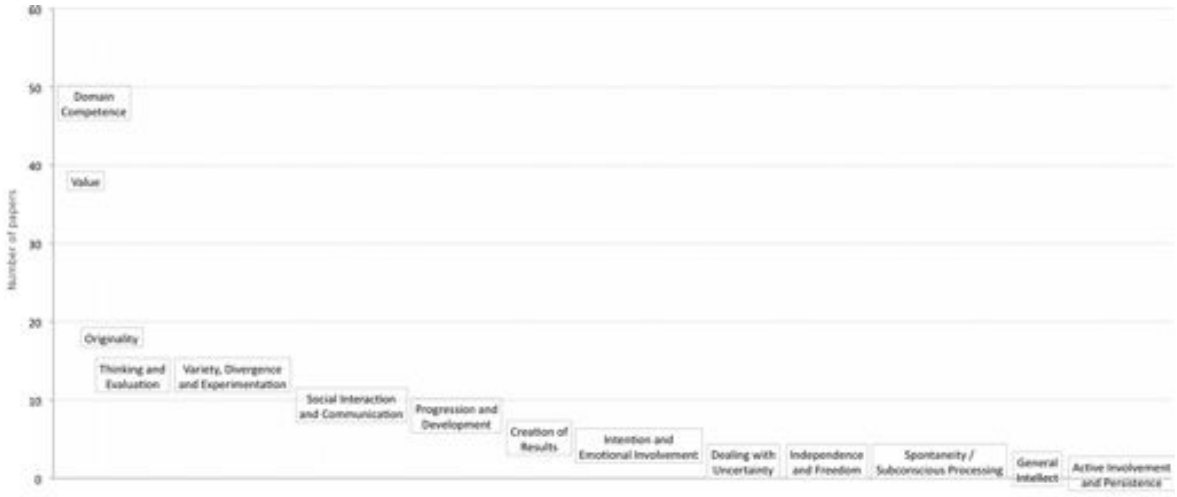


Figure 5.4: A visual representation of how many papers tested for each component in the survey from Chapter 2. N.B. There is no significance to the *x*-axis positioning of components apart from layout for ease of readability.

The survey data gives us some examples on the types of tests that have been applied to evaluate different components. Ways of testing each component are summarised below. Components are considered in descending order of the number of relevant evaluative tests found.

Domain Competence (in 47 papers) Many papers evaluate the appropriateness of results relevant to the domain. Appropriateness can be as rated by judges, through user feedback or by seeing if the results are fit for a practical purpose through practical application or user judgement. Comparisons were often made to see how similar a system's results were to existing examples in that domain. Some authors critically discussed the appropriateness of their results in the paper.

For some systems, competence in their domains was based on their performance in comparisons or competitive contests with similar systems or with humans doing a similar task. Systems using the empirical criteria in Ritchie (2007) for evaluation rated the products of their system for their typicality (and value) in a domain. This was done through ratings of typicality by human judges or by merging typicality and value ratings together for one rating of 'recognisability' (see Chapter 2 Section 2.1.6). Other systems used Colton's creative tripod framework for evaluation (Colton, 2008b), describing their system's demonstration of *skill*, *imagination* and *appreciation* to varying levels of detail. *Domain Competence* corresponds to the quality of *Skill* in the creative tripod framework.[32]

Value (38) Evaluations of *Value* of the surveyed systems are typically based on aspects of the end product(s) rather than any of the other Four Ps: Process, Person or Press (see Chapter 3 Section 3.4.2).

[32]Chapter 8 Sections 8.2.2 and 8.2.4).

While there were many examples of empirical measurements of *Value*, as described below, several systems' *Value* was assessed through user evaluations. Evaluation data was either directly provided by the user or provided indirectly through studies, such as through audience reactions and feedback at exhibitions or through qualitative tests with target users for usability and effectiveness of the system. Feedback about the appeal of systems' products and personal preferences about the products was also provided through user evaluations.

Many systems were evaluated by the correctness and validity of their products, such as calculating the percentage of material produced during runtime that can actually be used, or statistical tests for validity. Some systems were measured in terms of how interesting their products were, for example seeing if the products performed at a level above a given threshold for *Originality* in the Wundt curve function[33] or using variables representing domain-specific interest or complexity measurements.

The usefulness of a system's products could also be quantified, through the percentage of a user query which is satisfied by system output (Pereira & Cardoso, 2006), or the percentage of results that are valid. Human ratings of usefulness were also used. Usefulness ratings were not all quantitative, with use of post-implementation discussions on usefulness or the interpretation of value as serving an intended purpose. Other definitions of *Value* were less generally applicable across several types of creative system, using domain-specific metrics for value.

Originality (18) As for *Value*, *Originality* evaluations were based on the system output rather than on reflections about the process or the system itself. The newness and novelty of products were evaluated using ratings supplied by human judges or by using measures of similarity to compare system products against artefacts already known by the system. Not all systems that discussed evaluation of *Originality* actually gave details of how this evaluation was, or should be done. In a handful of cases the only evaluation done for this component was post-implementation discussion of the system's originality by the paper authors or brief comments.

Similarity metrics and comparison methods were used to measure originality and novelty compared to examples already known to the system. One interesting metric for originality and novelty was the use of the Wundt curve function (e.g. Saunders et al., 2010), which models the relationship between the amount of original material and 'hedonism' or pleasure; a threshold is reached beyond which more new information has a detrimental rather than positive effect.

Thinking and Evaluation (13) Contributions to this component were largely through systems that employ engagement-reflection or similar self-evaluation approaches (see Chapter 3 Section 3.4.1) to evaluate their products during runtime. Usually this evaluation involved metrics relating to other components such as *Originality* and *Value*. *Thinking and Evaluation* was either performed by the whole system on its operations, or by individual parts of the system in a low-level evaluation, for example

[33] See the comments on *Originality* for more details about the Wundt curve.

agents evaluating the work of others. Evaluative tests for this component were also performed via critical discussions about the system's ability (or lack of) to self-evaluate.

Analogies can be drawn between *Thinking and Evaluation* (primarily Evaluation) and the creative tripod quality of *appreciation* (Colton, 2008b) (Chapter 8 Sections 8.2.2 and 8.2.4). Other replications of thought processes were offered in the surveyed papers, although without much rigorous testing, such as an ability to justify findings and to apply lateral-thinking-style processes and transformative methods, or by a system using clustering algorithms to organise and evaluate its own output.

Variety, Divergence and Experimentation (13) Colton's creative tripod framework (Colton, 2008b) includes imagination, an important constituent of *Variety, Divergence and Experimentation* (Chapter 8 Sections 8.2.2 and 8.2.4). This component incorporates exploring and experimenting with a divergent set of possibilities, to find artefacts that other computational systems cannot, or by encouraging evolution and exploration, unrestricted by an end goal or overriding aim. Several papers critiqued the imaginative capacities of their systems, either as part of applying the creative tripod for evaluation of creativity to varying levels of detail or as a standalone key attribute.

Quantitative metrics were applied to measure *Variety, Divergence and Experimentation*, such as statistical tests like the SEGS method (Mozetič et al., 2010) (statistical tests for measuring the significance of differences between gene sets) or cross entropy calculations. Some systems employed user interaction to evaluate the diversity in their system. Whilst this can be a reasonable method for post-implementation evaluation, a 'fitness bottleneck' (Biles, 1994; Jordanous, 2010c) can result, as the system has to wait for user input before continuing the evolution process.[34]

Social Interaction and Communication (9) Social interaction capabilities of a system can be measured outside the system boundaries. Testing of user reactions to interactions was performed by measuring variables associated with users' engagement, interest, emotional reactions and self-assessed interaction levels. Certain elements of interactions were measured, for example counting how many different features of a system are used and how often. The other main contributor to this component was through agent-based systems, observing whether multi-agent systems outperformed single-agent systems on a task; and tracking how agents interactively evaluate each others' creative works.

Progression and Development (8) To see how system output develops over time, longitudinal studies compared subjective judgements made on the same system several months apart. In shorter time spans, there were brief discussions of how systems operate over time, or how system output improved over time. Several systems used evolutionary methods to progress their systems forward, measuring specific improvements through fitness functions or allowing undirected progress.

Generation of Results (5) Showing an influence from psychology research in creativity (Chapter 3 Section 3.4.2) metrics for fluency were employed similar to those in Guilford (1950). From infor-

[34] Automating the fitness function helps guard against this.

mation theory, measures of precision and recall quantified the accuracy and coverage of the results. Other systems' requirements were less precise, merely defining the production of results as a criterion for success (presumably measurable by observing system output).

Some systems were tested on the ability to detect that it had produced results and should therefore terminate, or to achieve end goals (regardless of whether the system was aware of this or not). Whilst program termination and goal completion are not always desirable properties, for example if the system is intended to develop over time and progress in creativity, in certain cases this type of dichotomous observation was useful; if a system had a target that it achieved, it could terminate.

Intention and Emotional Involvement (4) Some papers critically discussed the desire to be creative that was demonstrated by their system or talk about how the system's operation is driven by emotion. In one case a number of variables representing emotions were incorporated in the system, including domain-specific measures of danger perception, love, tension, empathy and emotive events.

Dealing with Uncertainty (2) Two papers discussed how curiosity was modelled in their system or how their system attempted to create curiosity, to deal with and reduce uncertainty by explorative methods, without reporting how this could be evaluated. One can speculate on ways to examine this. For example qualitative methods could be used, or information-theoretic measures could be employed to measure information that is uncertain or contradictory prior to system execution, which becomes known during the system's runtime.

Independence and Freedom (2) This was discussed rather than implemented in the two papers. Monitoring how much information is required from a user was a suggested way of measuring the system's independence; minimal levels of intervention would contribute to system success.

Spontaneity / Subconscious Processing (2) For this component, no papers talked about levels of cognitive processing or simulating consciousness or subconsciousness in their system, but spontaneity was hinted at. Spontaneity implies near-instantaneous results or changes in state. One proposed test (in two papers) therefore related to spontaneity as it involved the speed at which a result was produced by the system.

General Intellect (1) No papers tested their system in an IQ-style test of general intelligence, but one paper considered aspects of intelligent behaviour that were general rather than specific to that domain: forming hypotheses, spotting causality links between events and recognising appropriate temporal order for events.

Active Involvement and Persistence (0) This was not directly considered in any papers surveyed. The time taken to produce work partly corresponds to *Active Involvement and Persistence*, in terms of how the system avoids becoming stuck in the process (stagnation). Two systems measured how quickly results are produced but did not consider whether the systems stagnated or not, presumably

recording a value of infinity or similar if the system became caught in a situation such as a never-ending loop, never producing an end result. This test could be adapted though, to record how often the system became stuck compared to the number of times it produced a product.

5.4 Practical walkthrough guidance on applying SPECS for evaluation

Figure 5.6(a) and its child Figures 5.6(b) - 5.6(g)) give a guide on how to apply SPECS to evaluate how creative a system is, either in isolation or in comparison to other systems.[35] These decision-tree-style diagrams show how to approach the evaluation of systems exhibiting different types of creativity, in different domains, using SPECS. These figures give guidance[36] as to how to identify (or produce) creativity definitions, how to incorporate different types of definitions into the SPECS methodology and pointers towards how they may be evaluated.

It is emphasised that at each stage of the evaluative process, evaluators are expected to state what is being done or chosen and to justify why such actions or decisions are being made (with supporting evidence where appropriate), for a clear and reasoned evaluative process. Such transparency contributes to an overarching aim for SPECS: to encourage standardised, relevant and repeatable evaluation processes to be proposed, critiqued and developed by communications and progress in the field as a whole.

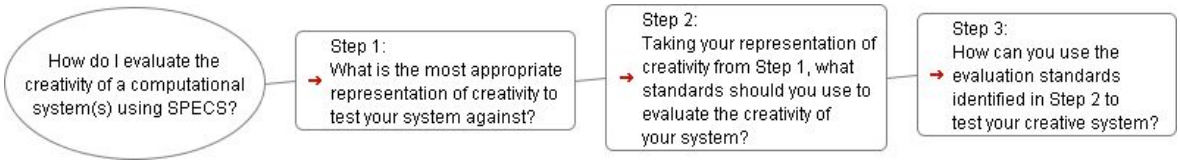
Additionally in this vein, before conducting SPECS, evaluators should be aware of any previous approaches to evaluate the creativity of comparable systems. Should any such creativity evaluations exist, or if evaluators wish to apply an existing evaluation method, the evaluation method should be critically reviewed as to whether it it appropriate for use to evaluate the current system(s) (or complement a larger evaluation). This critical review should consider the following points (and may require evaluators/peer critics to conduct Step 1 (both Steps 1a and 1b) to inform their choices):

- Does their approach actually evaluate creativity?
- Is their approach suitable for the type of creativity you are evaluating? (either as it is or with some customisation)
  - Can their representation of creativity be applied directly for your system's domain?

[35] The value of an evaluative comparison of one system with other comparable systems has been emphasised more than once in this thesis, most notably in the value of such feedback for the Case Study 1 systems, for measures of progress within the field of related research (and by the field). 'Comparable systems' refers to those systems which may generate useful feedback in such a comparison; for example you may wish to compare systems operating in similar domains or related domains, or perhaps explore whether comparison to different types of system could yield interesting results. It is noted that a key part of creativity (as discussed in Chapter 3 and elsewhere) is originality; hence (as highlighted by Simon Colton and others in discussions at the ICCC'11 conference) it is unlikely that there will be an existing creative system that is the same or nearly the same as the system being evaluated for creativity, though there often exist systems operating in similar domains. For example the systems in Case Study 1 (Chapter 6 all operate in the creative domain of musical improvisation, although they employ different processes and the musical output of these systems varies in style. References to 'comparable' or 'similar systems' in the current discussions are made with this point borne in mind.

[36] References are made to useful content in other parts of this thesis where appropriate, for further specific guidance.

Figure 5.5: A series of decision-tree-style diagrams, for illustrative detail and practical guidance through the steps of SPECS for evaluating the creativity of a computational system. Throughout this series, the small arrows in the left hand side of a node indicate a link to that same node in a separate figure. Where there is a choice of options and one path is strongly recommended, this is given in bold type.

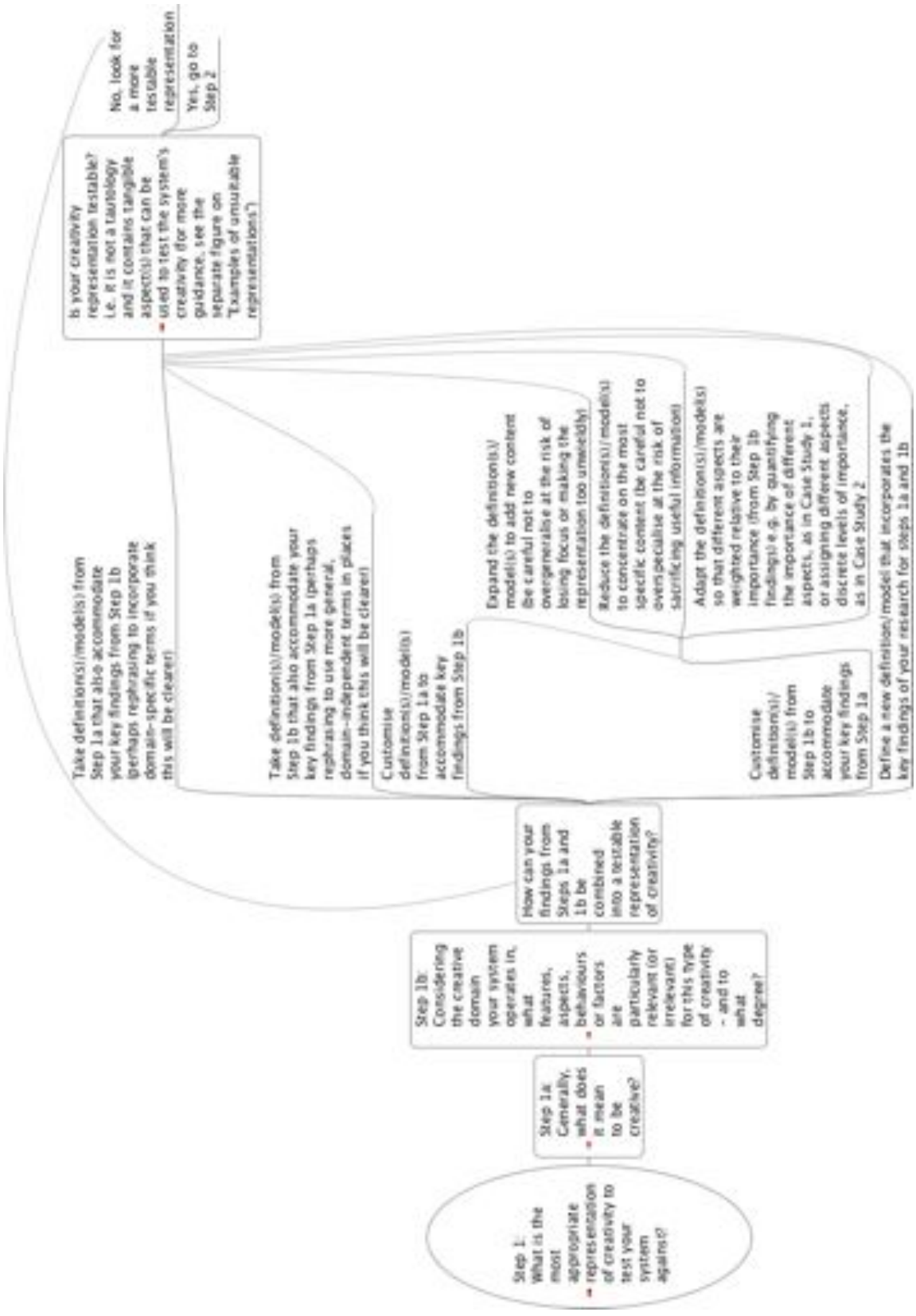


(a) SPECS evaluative process.

  - ∗ Does their representation of creativity accommodate general aspects of creativity? (Carry out Step 1a if necessary to inform this decision)
  - ∗ Does their representation take into account what we know about that type of creativity? (Carry out Step 1b if necessary to inform this decision)
- – If their representation of creativity is not appropriate as is for your system, can it be customised to your domain successfully, without missing important details or being too general or too specific?
  - ∗ Is their definition overly specific to their system, or to their domain, without taking into account more general aspects of creativity? (Carry out Step 1a if necessary to inform this decision)
  - ∗ Is their definition overly general, not taking into account specific requirements for creativity in that domain (Carry out Step 1b if necessary to inform this decision)
- Was their approach suitable for the type of creativity they are evaluating?[37]
- Similarly, can their evaluation methods be applied directly or customised in an appropriate way for your system?

The considered critique and reuse of suitable existing evaluation approaches and models emphasise the overriding aim towards a standardised approach to evaluation of computational creativity that encourages comparison and the placing of an individual system(s) within a wider research context, as measures of progress of the research area and the field as a whole.

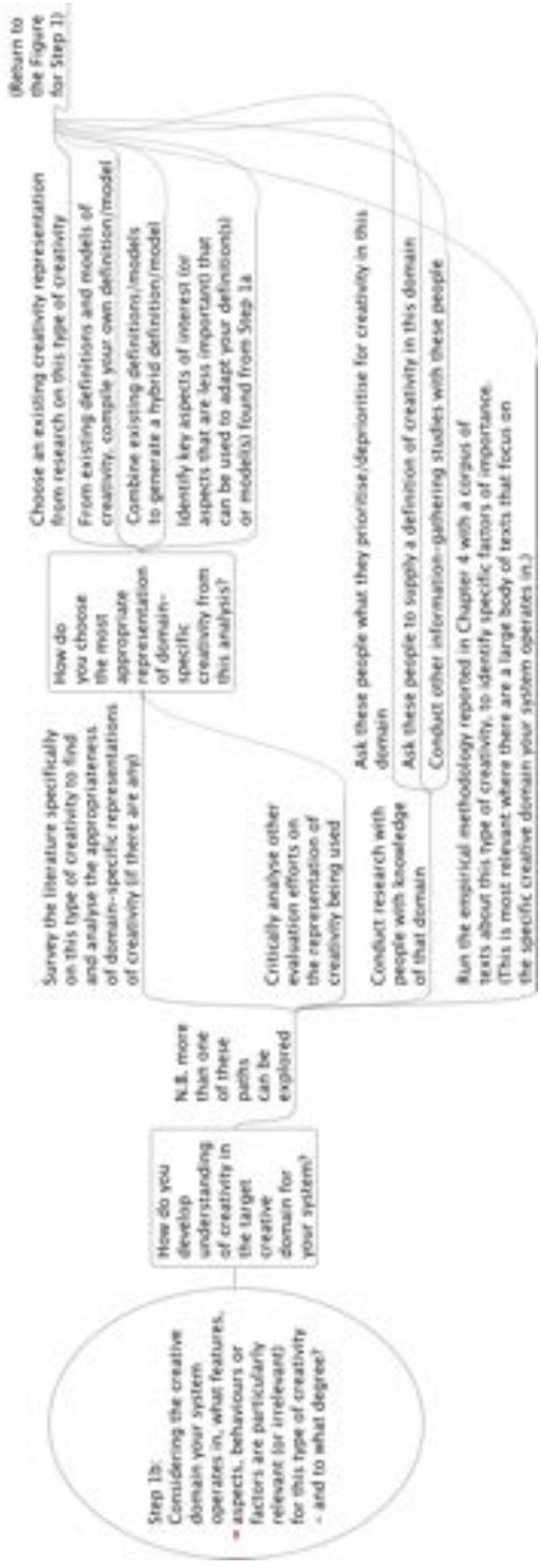
[37]This question focuses more on the appropriateness of their evaluation and only indirectly assists the current evaluation task; however after investigating the existing evaluation, the current evaluator(s) will be well placed to provide formative feedback to the original evaluators, as a valuable peer-review contribution.


Step 1:
What is the most appropriate representation of creativity to test your system against?
Step 1a:
Generally, what does it mean to be creative?
Step 1b:
Considering the creative domain your system operates in, what features, aspects, behaviours or factors are particularly relevant (or irrelevant) for this type of creativity - and to what degree?
How can your findings from Steps 1a and 1b be combined into a testable representation of creativity?
Is your creativity representation testable? I.e. it is not a tautology and it contains tangible aspect(s) that can be used to test the system's creativity (for more guidance, see the separate figure on 'Examples of unsuitable representations')
No, look for a more testable representation
Yes, go to Step 2
Take definition(s)/model(s) from Step 1a that also accommodate your key findings from Step 1b (perhaps rephrasing to incorporate domain-specific terms if you think this will be clearer)
Take definition(s)/model(s) from Step 1b that also accommodate your key findings from Step 1a (perhaps rephrasing to use more general, domain-independent terms in places if you think this will be clearer)
Customise definition(s)/model(s) from Step 1a to accommodate key findings from Step 1b
Expand the definition(s)/model(s) to add new content (be careful not to overgeneralise at the risk of losing focus or making the representation too unwieldly)
Reduce the definition(s)/model(s) to concentrate on the most specific content (be careful not to overspecialise at the risk of sacrificing useful information)
Adapt the definition(s)/model(s) so that different aspects are weighted relative to their importance (from Step 1b findings) e.g. by quantifying the importance of different aspects, as in Case Study 1, or assigning different aspects discrete levels of importance, as in Case Study 2
Customise definition(s)/model(s) from Step 1b to accommodate your key findings from Step 1a
Define a new definition/model that incorporates the key findings of your research for steps 1a and 1b

(b) Step 1.

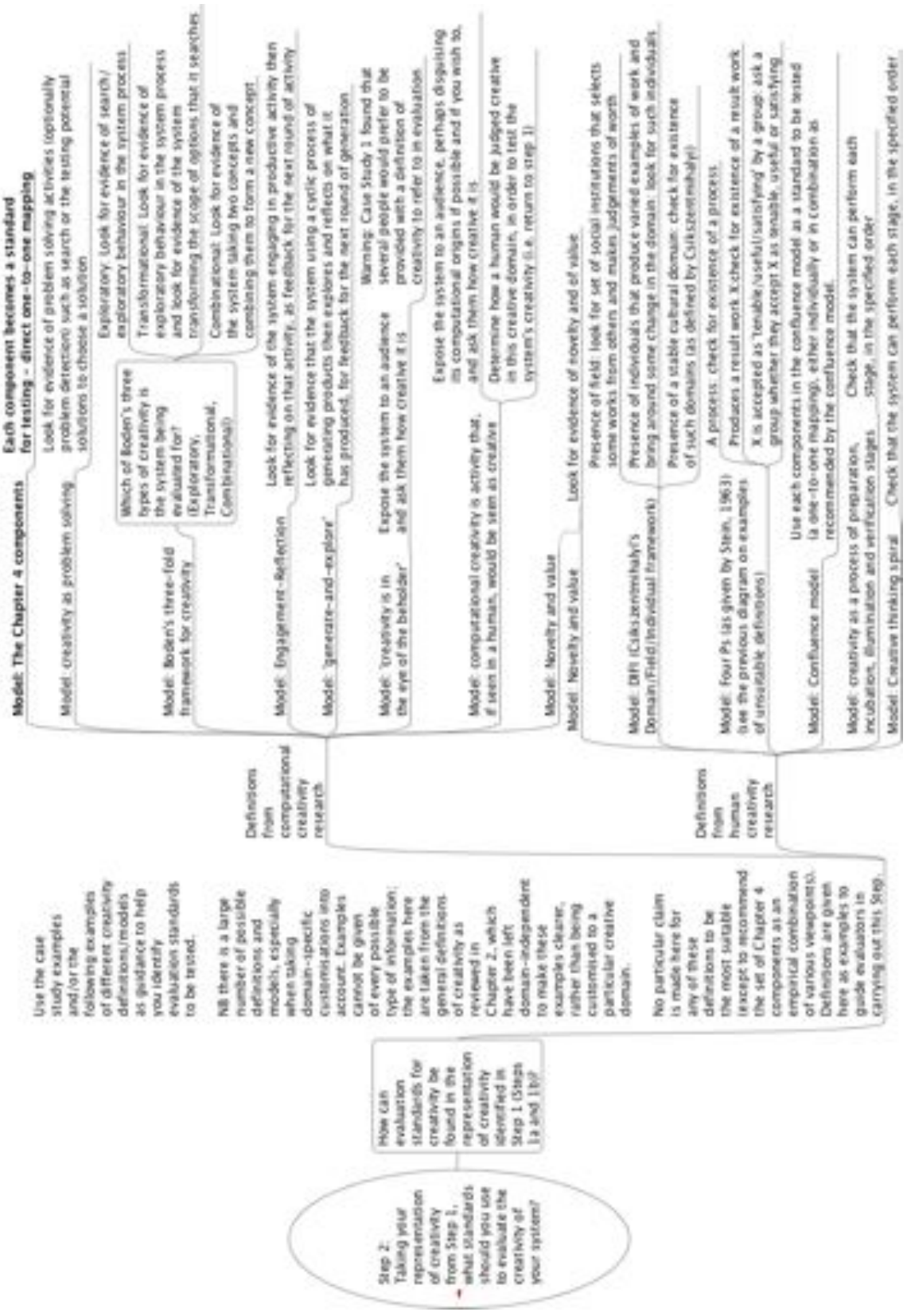

Step 1a: Generally, what does it mean to be creative?
How can you develop your general understanding of creativity?
N.B. more than one of these paths can be explored
Survey the literature on creativity to find and analyse the appropriateness of domain-independent representations of creativity
Critically analyse other evaluation efforts on the representation of creativity used
Critically analyse the relevance of applying existing creativity evaluation methodologies to evaluate the creativity of your system (e.g. Rhodhie's criteria, Colton's creative tripod, the FACE/IDEA model) (see Chapter 2)
How do you choose the most appropriate representation of creativity from such analysis?
Adopt the building blocks as a criteria-based representation of general creativity (recommended)
Run the empirical methodology reported in Chapter 4 with a new creativity corpus, to generate a new set of creativity components. (This will be most relevant in the future, where a corpus based on papers dated from 1950–2009 may not be sufficiently up-to-date.)
Choose an existing creativity representation from creativity research
From existing definitions and models of creativity, compile your own definition/model
Combine existing definitions/models to generate a hybrid definition/model
to Step 1b

(c) Step 1a.

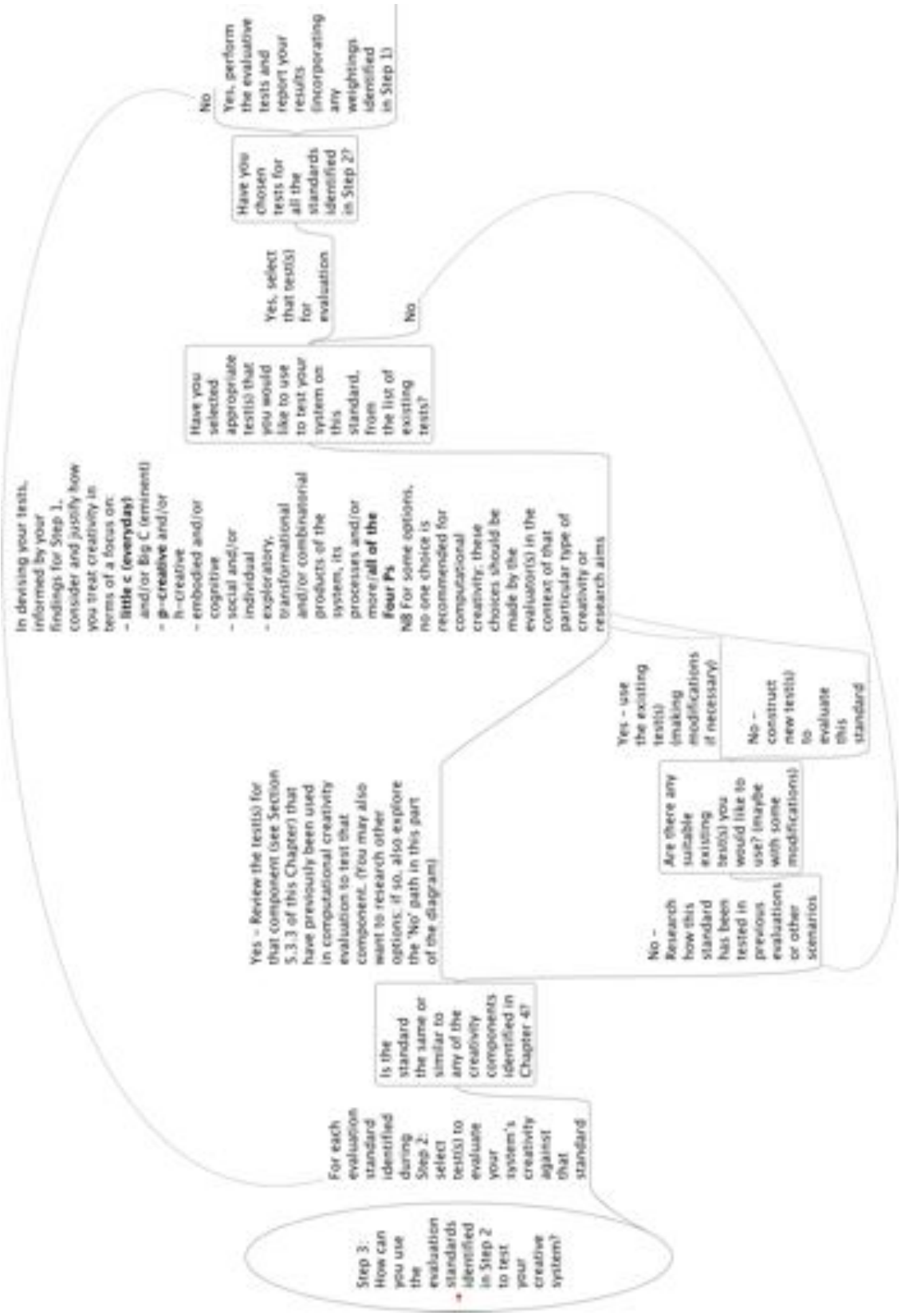

Step 1b:
Considering the creative domain your system operates in, what features, aspects, behaviours or factors are particularly relevant (or irrelevant) for this type of creativity - and to what degree?
How do you develop understanding of creativity in the target creative domain for your system?
N.B. more than one of these paths can be explored
Survey the literature specifically on this type of creativity to find and analyse the appropriateness of domain-specific representations of creativity (if there are any)
How do you choose the most appropriate representation of domain-specific creativity from this analysis?
Critically analyse other evaluation efforts on the representation of creativity being used
Conduct research with people with knowledge of that domain
Ask these people what they prioritise/deprioritise for creativity in this domain
Ask these people to supply a definition of creativity in this domain
Conduct other information-gathering studies with these people
Run the empirical methodology reported in Chapter 4 with a corpus of texts about this type of creativity, to identify specific factors of importance. (This is most relevant where there are a large body of texts that focus on the specific creative domain your system operates in.)
Choose an existing creativity representation from research on this type of creativity
From existing definitions and models of creativity, compile your own definition/model
Combine existing definitions/models to generate a hybrid definition/model
Identify key aspects of interest (or aspects that are less important) that can be used to adapt your definition(s) or model(s) found from Step 1a
(Return to the Figure for Step 1)

(d) Step 1b.

(e) Examples of unsuitable representations.


Step 2: Taking your representation of creativity from Step 1, what standards should you use to evaluate the creativity of your system?
How can evaluation standards for creativity be found in the representation of creativity identified in Step 1 (Steps 1a and 1b)?
Use the case study examples and/or the following examples of different creativity definitions/models as guidance to help you identify evaluation standards to be tested.
NB there is a large number of possible definitions and models, especially when taking domain-specific customisations into account. Examples cannot be given of every possible type of information; the examples here are taken from the general definitions of creativity as reviewed in Chapter 2, which have been left domain-independent to make these examples clearer, rather than being customised to a particular creative domain.
No particular claim is made here for any of these definitions to be the most suitable (except to recommend the set of Chapter 4 components as an empirical combination of various viewpoints). Definitions are given here as examples to guide evaluators in carrying out this Step.
Definitions from computational creativity research
Definitions from human creativity research
Model: The Chapter 4 components
Each component becomes a standard for testing – direct one-to-one mapping
Model: creativity as problem solving
Look for evidence of problem solving activities (optionally problem detection) such as search or the testing potential solutions to choose a solution
Exploratory: Look for evidence of search/exploratory behaviour in the system process
Transformational: Look for evidence of exploratory behaviour in the system process and look for evidence of the system transforming the scope of options that it searches
Combinational: Look for evidence of the system taking two concepts and combining them to form a new concept
Model: Boden's three-fold framework for creativity
Which of Boden's three types of creativity is the system being evaluated for? (Exploratory, Transformational, Combinational)
Model: Engagement-Reflection
Look for evidence of the system engaging in productive activity then reflecting on that activity, as feedback for the next round of activity
Model: 'generate-and-explore'
Look for evidence that the system using a cyclic process of generating products then explores and reflects on what it has produced, for feedback for the next round of generation
Model: 'creativity is in the eye of the beholder'
Expose the system to an audience and ask them how creative it is
Warning: Case Study 1 found that several people would prefer to be provided with a definition of creativity to refer to in evaluation
Model: computational creativity is activity that, if seen in a human, would be seen as creative
Expose the system to an audience, perhaps disguising its computational origins if possible and if you wish to, and ask them how creative it is
Determine how a human would be judged creative in this creative domain, in order to test the system's creativity (i.e. return to step 1)
Model: Novelty and value
Model: Novelty and value
Look for evidence of novelty and of value
Presence of field: look for set of social institutions that selects some works from others and makes judgements of worth
Model: DIFI (Csikszentmihalyi's Domain/Field/Individual framework)
Presence of individuals that produce varied examples of work and bring around some change in the domain: look for such individuals
Presence of a stable cultural domain: check for existence of such domains (as defined by Csikszentmihalyi)
A process: check for existence of a process
Model: Four Ps (as given by Stein, 1963) (see the previous diagram on examples of unsuitable definitions)
Produces a result work X: check for existence of a result work
X is accepted as 'tenable/useful/satisfying' by a group: ask a group whether they accept X as tenable, useful or satisfying
Use each component in the confluence model as a standard to be tested (a one-to-one mapping), either individually or in combination as recommended by the confluence model.
Model: Confluence model
Model: creativity as a process of preparation, incubation, illumination and verification stages
Check that the system can perform each stage, in the specified order
Model: Creative thinking spiral
Check that the system can perform each stage, in the specified order

(f) Step 2.


Step 3: How can you use the evaluation standards identified in Step 2 to test your creative system?
For each evaluation standard identified during Step 2: select test(s) to evaluate your system's creativity against that standard
Is the standard the same or similar to any of the creativity components identified in Chapter 4?
Yes - Review the test(s) for that component (see Section 5.3.3 of this Chapter) that have previously been used in computational creativity evaluation to test that component. (You may also want to research other options: if so, also explore the 'No' path in this part of the diagram)
No - Research how this standard has been tested in previous evaluations or other scenarios
Are there any suitable existing test(s) you would like to use? (maybe with some modifications)
Yes - use the existing test(s) (making modifications if necessary)
No - construct new test(s) to evaluate this standard
In devising your tests, informed by your findings for Step 1, consider and justify how you treat creativity in terms of a focus on: - little c (everyday) and/or Big C (eminent) - p-creative and/or h-creative - embodied and/or cognitive - social and/or individual - exploratory, transformational and/or combinational products of the system, its processes and/or more/all of the four Ps NB For some options, no one choice is recommended for computational creativity; these choices should be made by the evaluation(s) in the context of that particular type of creativity or research aims
Have you selected appropriate test(s) that you would like to use to test your system on this standard, from the list of existing tests?
No
Yes, select that test(s) for evaluation
Have you chosen tests for all the standards identified in Step 2?
No
Yes, perform the evaluative tests and report your results (incorporating any weightings identified in Step 1)

(g) Step 3.

5.5 SPECS in the context of standard scientific methodology

The requirements of SPECS are to define what it means for the system in question to be creative, identify from this definition evaluation standards and perform the evaluation. These requirements are comparable to the general requirements of scientific method as identified in Chapter 2 Section 2.2,[38] to identify an appropriate hypothesis and relevant empirical consequences of that hypothesis which can be used to test the hypothesis's validity.

Like SPECS, scientific method is generally perceived as focusing not on the discovery of hypotheses, but on how hypotheses are tested and validated or falsified. SPECS makes the assertion that that the creativity of a system should be measured in terms of how creativity is manifested generally and in domain-specific ways, as informed by the body of research on the nature of creativity. SPECS focuses on how appropriate definitions can be tested, without imposing instructions on exactly how such definitions should be arrived at (though guidance is offered in the form of the decision-tree-style diagrams in Figures 5.6(a) - 5.6(g) in this Chapter).

Parallels between SPECS and scientific method can hence be drawn. There are however a number of ways in which SPECS diverges from scientific method, due to the nature of creativity and of computational creativity research:

Computational creativity research is not (solely) a scientific endeavour Scientific method 'is what distinguishes *science* from other ways of acquiring knowledge or belief.'(Bird, 1998, p. 237, my emphasis added). This thesis treats computational creativity research as inclusive of a wider range of disciplines and motivations (Chapter 1 Section 1.5), rather than casting computational creativity research as purely scientific. SPECS is therefore inclusive of these 'other ways of acquiring knowledge or belief', rather than discarding them for being motivated by, say, artistic or engineering concerns. Sloman disagrees with a demarcation of science as distinct from non-science (e.g. metaphysics, pseudoscience, philosophy):

> 'Rather, they are all aspects of the attempt to discover what is and is not possible in the world and to understand why.' (Sloman, 1978, p. 26)

Sloman (1978) does not consider, though, how areas such as the arts, humanities or engineering fit in with this broader view of 'what is and is not possible in the world'. He also examines the issues of how we acquire, develop and understand new knowledge and theories under the specific question: 'What are the aims of *science*?' (Sloman, 1978, Part One, Chapter 2 title, pp. 22-62, my emphasis added), consistently referring to 'science' and 'scientific' practice rather than using more general terms. Elsewhere, such emphasis on the 'scientific' is a defining part of many scientific methods

[38] Chapter 2 Section 2.2 found that rather than seeking one universally agreed standard scientific method, it is more appropriate to treat scientific method as several possible methods (Bird, 1998; Hacking, 1981b; Feyerabend, 1993). Therefore here SPECS is considered in the context of some generalisations identified about scientific methods, rather than one scientific method in particular.

(Popper, 1972; Hacking, 1981b; Bird, 1998), despite the complications that it introduces (Thagard, 1988, p. 48). This emphasis on *scientific* knowledge distances scientific method somewhat from computational creativity evaluation. Many of the points mentioned next arise as a consequence of this fundamental point.

Transience of definitions of creativity over time and in context The dynamic nature of creativity means that a static, context-free definition is inappropriate (Chapter 3). In particular, Chapter 3 showed a number of reasons why it is unlikely that a universal definition of creativity will be arrived at; instead manifestations of creativity vary according to the domain and contextual and/or time-dependent factors. Crucially, in scientific method, hypotheses are treated as static statements whose validity (or non-validity) would not be affected by changes in time and context. This is inappropriate for the evaluation of creativity and creates extra issues to be dealt with if scientific method is to be used for computational creativity evaluation. SPECS, on the other hand, can accommodate creativity's dynamism and context-sensitiveness (particularly with the focus on domain-specific aspects of creativity in Step 1b). Specifically this occurs when appropriate definitions - and therefore appropriate evaluation criteria - have been identified. As can be seen from the emphasis throughout this thesis on understanding creativity, in conceiving SPECS, a priority has been to perform these steps well.

In the review of scientific method (Chapter 2 Section 2.2) it was identified that Popper (1972) saw scientific knowledge as being developed in a more tentative, temporary manner, subject to scientists' fallibility. Lakatos (1978) and Kuhn (1962) both considered how anomalies and discrepancies could be catered for within a wider view of the growth and development of scientific knowledge. Even the more revolutionary views of Feyerabend (1993) still treat scientific knowledge as being based on theories that can be investigated and perhaps contradicted. Such views could perhaps make scientific method more appropriate for evaluating creativity at a given point in time and context. In that same Section, though, the point was made that Popper probably saw such temporary adoption of knowledge as being determined by what positive or negative evidence had been found up till that point, rather than to the changing nature of that evidence itself. As mentioned in discussions with other researchers on the representation of creativity in terms of formalisable scientific theories, 'creativity has another dimension to it that cannot be addressed in these terms. What concerns me is that applying the framework of knowledge as theory-based to creativity can misdirect you in fundamental ways' (Beynon, 2012, personal communications).

Creativity as a continuous rather than binary quality Chapter 3 (particularly Section 3.4.2) has shown how creativity is generally treated as something which can be comparatively measured and or treated as a continuous quantity rather than a binary state, i.e. creative or not creative. In other words, rather than treat something/someone as either creative or not creative, the interest is in the level of creativity demonstrated - the degree to which that person or thing is creative. When treating

the definition of creativity as a hypothesis for scientific method, such a treatment of creativity as continuous may jar somewhat with the traditional interpretation of a hypothesis in scientific method that is often exemplified in statements that are assigned a discrete truth value. A common assumption of such examples in scientific method is that hypotheses can be thought of as either valid or not valid, but not valid to a certain extent. The latter scenario does not seem to be ruled out as an option for scientific method hypotheses but it is not commonly exemplified in introductory literature to scientific method (e.g. Bird, 1998).

Similarly to the previous point, though, SPECS can accommodate this interpretation of creativity without introducing problem if Steps 1a, 1b and 2 are carried out in a suitable manner. The responsibility for using appropriate definitions lies with the evaluator, rather than with SPECS;[39] however SPECS requires the evaluator to clearly state definitions and derived evaluation standards being used, which makes them available for peer comment and review within the research community.

Difficulties of predicting empirical consequences expected As has been seen in Chapters 3 and 4, originality is an identified facet of creativity. Scientific method requires that a hypothesis is tested by the existence (or not) of evidence confirming predicted consequences of that hypothesis. There is therefore a question as to whether the predictability and expectability of consequences for a hypothesis is incompatible with the originality aspect of creativity. If so, this implies that problems can arise in identifying hypotheses for computational creativity evaluation. Scientific method, whether using methods of induction, falsification or alternatives, considers the impact of existing evidence for and against a hypothesis. As discussed in Chapter 2 Section 2.2, problems arise using scientific method where there is a lack of appropriate empirical evidence, positive or negative.

Similar problems also face SPECS. To cope with such difficulties, Chapter 2 Section 2.2 notes that where possible in our evaluations we should seek to work with empirically testable hypotheses, without anticipating desired instances of empirical evidence in too much detail. For example, the case studies in Chapters 6 and 7 look at aspects of the system and its workings rather than looking for particular output to be generated.

Taking a more holistic view of evaluation: Various issues were discussed in Chapter 2 Section 2.2 with the various methods for testing hypotheses. As solution, several types of such scientific method have been suggested over time, each with their benefits and disadvantages and it is possible to adjust some methods accommodate issues such as Hempel's paradox (Bird, 1998). Overall, though, scientific methods based on the hypothetico-deductive model and others like it are too simplistic, as Bird (1998) argued. Instead Bird suggested that rather than strictly following scientific method, we should look to more 'holistic' understandings of both the explanation of facts and confirming evidence (Bird,

[39]To assist the evaluator, guidance is provided in this Chapter on how to use SPECS and the requirements to identify both general aspects of creativity and domain-specific aspects, for a particular system. Case studies are also provided in Chapters 6 and 7, as examples of SPECS in application.

1998, p. 94), taking contextual and surrounding interrelated explanations into account (determining such explanations where necessary). Context and informed working allows you to take advantage of what you know.

Incorporating holistic qualities into a formal hypothesis is possible but is included more easily in the more informal approach to definition (rather than hypothesis) in SPECS. When implemented as recommended, i.e. with the creativity components from Chapter 4 as a basis for Step 1a, SPECS does indeed follow a holistic approach to evaluation, considering fourteen aspects of a system in the evaluation of its creativity that collectively build up a wide coverage of the nature of creativity. This type of approach is illustrated in the case studies in the following two Chapters.

Methodological guidance rather than prescription '*The only principle that does not inhibit progress is:* anything goes' (Feyerabend, 1993, p. 23).

As commented in Chapter 2 Section 2.2, the principle 'anything goes' of Feyerabend (1993) is consonant with the concept of creativity. This thesis takes the approach that some guidance in methods would be useful practically, to help computational creativity evaluators find appropriate methods available and share them with other researchers. However, Feyerabend (1993) rejected the idea of imposing specific methodological recommendations, arguing that selection of methods should be left open rather than being constrained. Feyerabend's views contrast with other proposals for scientific methods, which may be considered closer to a standard interpretation of scientific method, such as falsificationism or induction, where the methodological process is predetermined. SPECS, on the other hand, leaves the choice of methods up to the individual researcher(s), though it requires that the researcher clearly states (and is accountable for) their methodological choices.

In concluding this comparison between SPECS and scientific method, one can see that the SPECS methodology derived in this thesis is compatible with scientific method in some ways, mostly concerning the general structure and foundational principles of SPECS. There are however a number of differences between SPECS and scientific method, largely due to how SPECS handles the non-scientific and dynamic nature of computational creativity. These differences demonstrate that SPECS is differentiable from scientific method in several ways and that SPECS is more appropriate than scientific method for the task of evaluating computational creativity.

5.6 Applying the SPECS methodology: A look ahead

Section 5.3 outlined how the SPECS methodology can be implemented practically for use in evaluating creative systems. Evaluation case studies can demonstrate how this works in practice, test how applicable SPECS is and provide evaluative feedback for the systems concerned.

Chapters 6 and 7 will report how SPECS was applied in two case study scenarios. Following from observations in the survey of evaluative practice reported in Chapter 2, the two case studies explored

evaluation from two perspectives:

Case Study 1, Chapter 6: A detailed comparative evaluation of the creativity of musical improvisation systems. Evaluation was performed by domain experts and supported by information on creativity in musical improvisation.

Case Study 2, Chapter 7: Making a quick decision, in a short space of time and with limited information, on the creativity of systems in various domains. The systems being evaluated were presented at the International Computational Creativity Conference in 2011. Two judges were involved, whose domain expertise ranges from expert to novice, depending on the system.

Step 1a, a domain-general definition of creativity: the 14 components of creativity identified in Chapter 4 were used collectively as a definition.

Step 1b, identifying domain-specific requirements for creativity: investigations were conducted during each case study to see how to weight each of the 14 components most appropriately.

Step 2, stating what standards should be used to evaluate the system: the standards used for evaluation were the 14 components. Analysis of each component's contribution to overall creativity was weighted in terms of the priorities identified in Step 1b.

Step 3, testing the system and reporting results: systems were given numeric ratings out of 10 (10 being highest, 0 being lowest) for how well they satisfied or demonstrated each of the 14 components of creativity. Each rating was weighted according to the importance of that component in the domain, as well as being considered individually.

There were two main reasons behind the decision to use human assessment rather than empirical observations about system features:

1. Making observations about the system while it is running requires having the system available and being able to run it. This was not possible for the majority of case study systems evaluated, although in most cases there were examples of products generated by the systems.
2. By using human judges to give ratings of the components, useful feedback could be obtained from the judges on the efficacy of the components, on how easy they are to understand and to apply to computational creativity evaluation. Feedback was especially forthcoming in Case Study 1 (Chapter 6) where the judges were interviewed to obtain component ratings. During the interviews they were asked to voice their thought processes aloud whilst deciding on ratings.

Analysis in the case studies concentrated on what has been successful and where the system could be improved. Some comparison between systems were also performed:

- In the first case study, where all systems were from the domain of musical improvisation, systems were compared and contrasted to see which are more successful, both generally and for specific components.
- In the second case study, where the system domain varied, some limited comparison was made

across systems, but this was less relevant as the systems are designed to perform quite different tasks. In this case study, more emphasis was placed on evaluating individual systems.

5.7 Summary

A comparative, scientific evaluation of creativity is essential for progress in computational creativity, not least to justify how creative a computational creativity system actually is. Chapter 2 Section 2.3 showed that while creative systems are often evaluated with regard to the quality of the output, and described as creative by the authors, in all but a third of cases the creativity of these systems is not evaluated and claims of creativity are left unverified. Often a system is evaluated in isolation, with no reference to related systems even where such references may offer useful insights for current work.

Section 5.1 discussed these issues in the context of existing evaluation methodologies for computational creativity. These discussions contextualised the issues from a wider perspective of creativity research, both in computers and in humans. As described above, only some of these problems have been looked at. Many remain as troublesome issues for researchers who wish to evaluate their computational creativity. To address some of these issues, recommendations were made in Section 5.2.1: examining more of the Four Ps (Chapter 3 Section 3.4.2) than just the end product, using a combination of quantitative and qualitative methods, being clear about the definition of creativity when performing evaluation and also being clear about the methods used.

To resolve some of the above issues and progress towards the goal of this thesis, Section 5.2 outlined Evaluation Guidelines as s set of heuristics incorporating many of the recommendations arrived at in this thesis so far. Considering how these heuristics could be applied practically for evaluation (Section 5.3) has led to the devising of the SPECS methodology: the *Standardised Procedure for Evaluating Creative Systems*. SPECS requires the researcher to evaluate the creativity of system(s) using a more informed, clearly stated and justified understanding of what it means for the system(s) to be creative; the set of components derived in Chapter 4 is strongly recommended as a basis for this understanding. The SPECS methodology was presented, with discussion of various relevant application issues, in Section 5.3. Figures 5.6(a) - 5.6(g) provided further guidance on applying SPECS in various evaluation scenarios.

Section 5.5 addressed the question of SPECS' novelty compared to scientific method by acknowledging the parallels that can be drawn between SPECS and scientific method but clarifying how SPECS differs from scientific method. The differences mostly relate to how SPECS handles the non-scientific requirements of computational creativity research and how it deals with the dynamic and context-sensitive nature of creativity. Identifying these differences has illustrated how SPECS is more appropriate than scientific method for evaluation of computational creativity.

Chapters 6 and 7 will apply SPECS to evaluate the creativity of two sets of computational creativity systems, generate feedback on the systems' creativity and compare the systems against each other.

Chapter 6

Case Study 1: Detailed evaluation of musical improvisation systems

---

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012) and peer-reviewed conference papers (Jordanous, 2011a; Jordanous & Keller, 2011).

Figure 6.1: *Wordle* word cloud of this Chapter's content

Overview

This first case study illustrates how the SPECS methodology can be applied practically to use detailed domain expertise and experience in evaluating creative systems. Musical improvisation is the creative activity focused on; an introduction to creativity in this domain is given in Section 6.1. The creativity of three musical improvisation systems is evaluated: Voyager (Lewis, 2000), GenJam (Biles, 2007) and GAmprovising (Jordanous, 2010c). Section 6.2 gives details of each of these systems.

Section 6.3 describes the application of the SPECS methodology. Step 1, defining creativity in the relevant context, uses the 14 components from Chapter 4 as a general definition of creativity (Step 1a) and investigates how each component contributes to creativity in musical improvisation (Step 1b). Step 2, identifying appropriate standards with which to test the three systems, weights each component according to how much that component contributes to musical improvisation creativity. Step 3, testing the systems by the Step 2 standards, uses expert-provided ratings to evaluate the systems along each of the 14 components. These ratings are weighted according to component importance. Qualitative data is also collected from the judges' comments.

Information obtained through SPECS evaluation is analysed to produce evaluative feedback about the systems' creativity, both in comparing the systems to each other and for formative feedback about the systems' strengths and weaknesses in being creative musical improvisers. Results are presented and discussed in Section 6.4.

6.1 Musical improvisation as a creative domain

In his keynote talk at ICCC'11, George Lewis referred to improvisation as 'the ubiquitous practice of everyday life', communicating meaning and emotion such that while improvising, 'one hears something of oneself' (Lewis, 2011). Lewis (2011) reported how Evan Parker, an accomplished improviser on saxophone, describes mistakes in improvisation as the missing of chances. Parker himself says of improvisation: 'The activity is its own reward' (Parker, 2011).

Issues of choice and liberty were raised in Lewis (2011), having a choice of what expressive actions to perform and when to perform them. Neural evidence (Csikszentmihalyi, 2009; Friis-Olivarius, Wallentin, & Vuust, 2009; Berkowitz & Ansari, 2010) shows that brain activity during improvisation relates to brain activity when making choices. Lewis (2011) believes that this neural evidence demonstrates that one is never fully in control during improvisation.

Berliner describes how musical improvisers need to balance the known and unknown, working simultaneously with planned conscious thought processes and subconscious emergence of ideas (Berliner, 1994). Berliner examines how musical improvisers learn from studying those who precede them, then develop that knowledge to develop a unique style. The recent work of Louise Gibbs in musical improvisation education equates 'creative' with 'improvisational' musicianship. She highlights

invention and originality as two key components for creative improvisation (Gibbs, 2010).

Not all people accept creativity in musical improvisation can be defined. Bailey proposes that the creative process exists at a level beyond which can be definitively captured in words:

> 'a fundamental belief for some people ... [is that] musical creativity (all creativity?) is indivisible; it doesn't matter what you call it, it doesn't matter how you do it. The creation of music transcends method' (Bailey, 1993, p. 140)

Pressing (1987) however advocates making more explicit connections between improvisation and creativity. For evaluative purposes and a clearer understanding overall, it is most productive to follow the lead of those such as Berliner and Gibbs, who make the study of improvisational creativity more tangible by describing it in terms of subprocesses (Berliner, 1994) or components (Gibbs, 2010).

6.2 Introducing the musical improvisation systems being evaluated

The SPECS methodology was applied to compare and contrast the creativity of three musical improvisation systems:[1]

- My own improvisation system, named *GAmprovising* for this study (Jordanous, 2010c).
- *GenJam* (Biles, 2007).
- *Voyager* (Lewis, 2000).

This was a reasonable number of systems to evaluate and compare, given time constraints and the methods of evaluation used (described in Section 6.3). These three systems were chosen due to the variety of resources available for each system to inform evaluation. A key criterion for inclusion in the case study was that there were examples of system output available and at least one paper describing how the system works.[2] Below, details of all three systems are presented.

GAmprovising GAmprovising (Jordanous, 2010c) is a genetic algorithm-based system consisting of populations of several *Improvisers* which evolve over time towards becoming more creative. Each Improviser in the GAmprovising system improvises several different solos, by putting together randomly chosen notes into a MIDI melody. These random choices are directed through parameters on

[1] Originally the Impro-Visor system (Gillick, Tang, & Keller, 2010) was also included in Case Study 1 (Jordanous, 2012). In later analysis, post-evaluation, it was however discovered that the musical examples used to illustrate Impro-Visor had not been generated by the Impro-Visor system, as originally believed, but had been generated by a human musician with (occasional) assistance of Impro-Visor's advice functionality, as is consistent with the educational aims of the Impro-Visor system. As the examples used for evaluating Impro-Visor were not computer-generated, this system has been removed from the reports of Case Study 1 in this thesis. This part of the thesis acts as the correct report of Case Study 1 and as a retraction of claims made in Jordanous (2012) regarding Impro-Visor.

[2] One reason for the inclusion of GAmprovising was because as author of this system, it was useful personally to see how the system is evaluated in the context of other systems and obtain feedback on GAmprovising's strengths and weaknesses as well as a perception of its overall creativity. In the case study, authorship of this system was attributed not to myself but to 'Joanna Sondaur'. This is a pseudonym generated from re-arrangement of the letters in my name. This meant that judges were unaware that GAmprovising was my own system, avoiding any resulting bias.

note, rhythm and voice restrictions, with different Improvisers having different parameter settings. The generated improvisations generally tend to have a stylistically 'free' and avant-garde feel.

Two improvisations are selected from each Improviser's repertoire and played to a human judge.[3] The human judge rates each improvisation on two things: how typical it is as an improvisation, and how much they liked it. The judge's ratings are used to generate a measure of how creative the improviser is, using the criteria in Ritchie (2007) to convert these ratings into a measure of creativity.[4] After all Improvisers have been evaluated for fitness, the highest-scoring Improvisers are used to seed a new generation of Improvisers. The whole process is then repeated with future generations, until the user wishes to stop the program. The GAmprovising system is designed to develop increasingly more creative behaviour over time.

GenJam GenJam (Biles, 2007) (short for Genetic Jammer) is a real-time interactive improvisation agent. It improvises in a jazz style by constructing melodies composed of several different small tunes (*licks*) from a database, during live performance. In the original version of GenJam (Biles, 1994), a human mentor listens while GenJam is improvising. The mentor provides feedback by typing 'g' for good and 'b' for bad. GenJam learns from this feedback; the best licks are kept and used to create new licks the next time GenJam plays. A more recent version of GenJam (Biles, 2007) is autonomous, learning from pre-existing melodies that have been judged as sounding good rather than using user feedback and amending those phrases in real-time to produce new improvisations.

Before performances, GenJam is given information describing the tune and its arrangement, including instructions on performance and structure. GenJam performs improvisation with a human performer by listening to what the human improvises, comparing what it heard to its internal representation, and using intelligent methods to develop what the human plays into its response.

Biles has recorded and released a CD entitled GenJam (released 1996, *Drk records*), naming the group of performers (himself and GenJam) as the Al Biles Virtual Quintet. Al Biles regularly performs with GenJam under this group name. No other people have improvised with GenJam[5] as the source code for GenJam is not available for others to run, primarily because GenJam has been implemented on what is now legacy equipment and platforms (Mac Powerbook 180, using the Carnegie Mellon CMU MIDI Toolkit) (Biles, 2007).

Voyager Voyager (Lewis, 2000) consists of 64 individual MIDI *players*, all of which automatically improvise live music in real time in an avant-garde musical style. Several different players may be active and improvising at the same time. Every 5 to 7 seconds, some of the MIDI players are grouped together to form a new ensemble. Voyager may make this new ensemble the only group that

[3] This is like the evaluator being given a 'demo' of the Improviser rather than having to listen to all their solos.

[4] The system was intended as a test of Ritchie's criteria as a fitness function of creativity, and a proof-of-concept model testing if random music generation techniques constrained by threshold limits can act as heuristics to emulate creative musical behaviour. As Jordanous (2010c) reports, practical issues arose in implementing the criteria.

[5] This has been verified through personal communication with Al Biles (2012).

is playing, or add this group to those groups already playing, or replace one existing group with this new ensemble. Variable settings within Voyager determine how the new ensemble should sound, how it should improvise melodies (from a choice of 15 different methods), what notes it should use at what volume, and various other musical decisions.

Voyager can 'listen' to up to two human improvisers who are playing alongside the computer system, using incoming MIDI messages. Each new ensemble decides how the improvisations played by the human musicians influence what that ensemble plays. The ensemble might choose to imitate, directly oppose or completely ignore the information coming from the human musicians. If there are no human musicians playing with Voyager, the system improvises using internal processes, so it can play as a solo musician as well as with other musicians.

Voyager is recorded on two CDs released under George Lewis's name: *Voyager* (released 1993, *Avant*) and *Endless Shout* (released 2000, *Tzadik*). The saxophone player Evan Parker, who has performed with Voyager, comments that when the Voyager programming is performing well, Lewis leaves it running and does not interfere with its operation, until it crashes, 'which it often does!' (Parker, 2011).[6] Lewis's view on Voyager's robustness in performance is that Voyager 'can play for as long as I think it's good' (Lewis, 2011).

6.3 Applying the evaluation methodology in Case Study 1

Resources and data on all three systems were collated to assist the application of the SPECS methodology derived in Chapter 5 to evaluate and compare the three systems' creativity.

6.3.1 Step 1a: Domain-independent aspects of creativity

The work reported in Chapter 4 identified common components of creativity that we prioritise as most important for creativity in general, across all domains.[7] The components were used for Step 1a, as per the recommendations given in Chapter 5 Section 5.3.

6.3.2 Step 1b: Aspects of creativity in musical improvisation

To identify the relative importance of the 14 components in musical improvisational creativity, 34 participants with a range of musical experience were questioned.

Questionnaire procedure

Each participant was emailed a questionnaire document to fill in and return. The questionnaire asked participants to think about eleven words or sets of words in the context of musical improvisation, and briefly describe what these words meant to them in this context.

[6] This was not meant as a negative comment on computer music in general; when asked about the future of improvisation, Parker replied 'I think it's going towards computers, and interaction with computers' (Parker, 2011).

[7] This set of components is pictured in Figure 4.3 of Chapter 4.

'What do these words mean to you, in the context of musical improvisation?'

These words were taken from the top results of a precursor study to the work in Chapter 4 (Jordanous, 2010a), investigating how language use in academic writings about creativity differed from standard British English language usage as represented in the British National Corpus (BNC Consortium, 2007). Each word listed above was found to be used significantly more often in the creativity writings. The words were clustered manually into ten different concepts and presented to the participants in randomised order. The eleventh word given to the participants was always 'creativity'.

1. thinking / thought / cognitive.
2. process / processes.
3. innovation / originality / new / novel.
4. divergence / divergent.
5. openness.
6. ideas / discovery.
7. accomplishments / contributions / production.
8. intelligence / skills / ability / knowledge / talent.
9. problem / problem-solving.
10. personality / motivation.

    [the above ten were presented in random order]

11 creativity [always presented last]

The motivation was to get the participants used to the process of thinking about words related to creativity in the context of musical improvisation. Having 'creativity' as the last word to consider meant that participants had ten short practice trials before tackling the word this study was most interested in. Originally it was also hoped that some useful data may be given in the answers for these questions, as the words were all closely related to creativity (Jordanous, 2010a). In general, though, the answers given focused quite specifically on the relevant word or group of words. Unfortunately this meant that the data from the first ten questions was of limited usefulness for the purposes of this thesis except as practice trials for the eleventh question.

In choosing this strategy, two potential problems arose. Firstly, and most importantly, there was a possibility of priming participants with the ten preceding words and word groups. This was the main danger of adopting this strategy. It was felt, however, that the benefits of giving the participants these 'practice trials' outweighed any priming effects. The second potential problem was that the participants may have been fatigued from the experiment by the time they reached the eleventh word. As Figure 6.2 shows, though, there was however no noticeable drop off in the length of responses given by participants as they addressed the word 'creativity'. Additionally, several participants volunteered reports afterwards that they enjoyed doing the questionnaire and became fully absorbed in answering

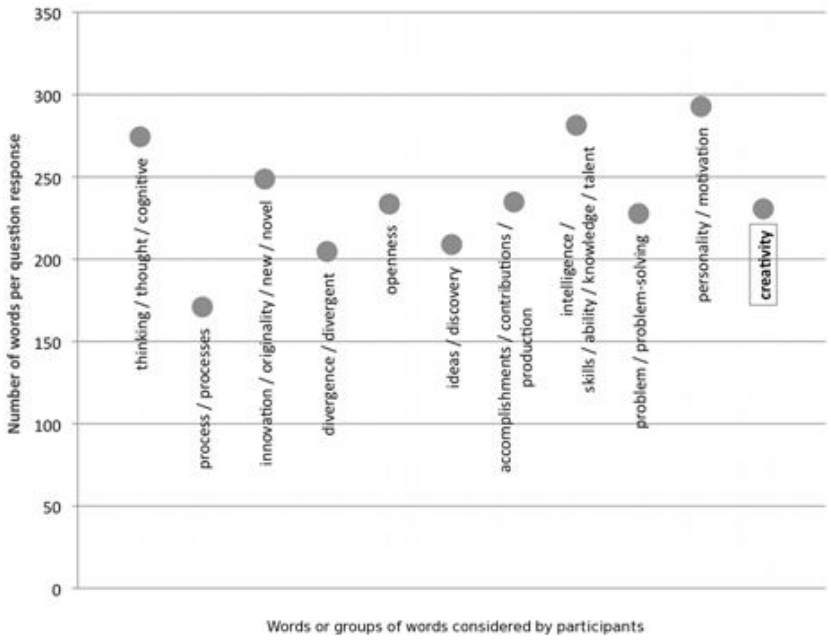


Figure 6.2: Responses for each question ranged from a mean length of 171 to 293 characters, with an overall mean of 237 characters. Responses for 'creativity' were a mean of 231 characters long.

the questions, rather than becoming bored or fatigued during the questionnaire process.

After finishing the questionnaire, participants were asked to read a debrief document, which briefly outlined the purposes of the questionnaire and introduced this research project. Having read this information, participants were asked this final question:

> *Are there any words which you feel are important for describing creativity in musical improvisation that have not been mentioned so far? If so, what are these words and why are they important?*

29 out of 34 participants added extra responses at this point, detailing words they identified as associated with musical improvisation creativity in this way.

Participants were asked to return both the completed questionnaire document and the debrief document to be analysed. They were encouraged to pass on any further comments or questions if they had any; this prompted further discussions with 6 participants, providing more data for analysis.

Questionnaire participants

Participants were asked about their musical experience and training and, if they were musicians,[8] what instruments and genres they played. A summary of participant demographics is given in Table

[8]Some non-musicians were included in the questionnaire as they had experience of listening to musical improvisation and were therefore able to give a slightly different perspective.

Table 6.1: Questionnaire participants: experience as musicians and as music improvisers. Musical experience: mean 20.2 years, s.d. 14.5. Improvising experience: 15.1 years, s.d. 14.3.

Level of experience	Musicians	Improvisers
Professional	15	10
Semi-professional	8	10
Amateur	8	9
None (Listeners)	3	5

6.1 and in Figure 6.3. Participants came from different improvisatory backgrounds and with different levels of expertise and experience of a variety of musical styles.[9] The participants were asked about what types of improvisation they did (including but not restricted to musical improvisation). All but three participants gave at least one example of their experience of improvising.

Analysis of questionnaire results

Responses to the question about creativity, the debrief question and any follow-on correspondence post-questionnaire were analysed using the 14 components from Chapter 4. This was done using response tagging; each point made by the participants was tagged according to which component it most closely illustrated. Negative as well as positive mentions were recorded.

SURVEY PARTICIPANTS	*non-improviser*	*amateur improviser*	*semi-pro improviser*	*pro improviser*	TOTAL
non-musician	2	1	0	0	**3**
amateur musician	3	4	1	0	**8**
semi-pro musician	0	3	4	1	**8**
pro musician	0	1	5	9	**15**
TOTAL	**5**	**9**	**10**	**10**	**34**

Figure 6.3: Questionnaire participants' musical and improvisational experience - how the population was distributed. The questionnaire was weighted towards professional musicians and improvisers.

After these responses were all tagged, each component was given a score that quantified the perceived importance of that component in the questionnaire data: the count of all positive mentions of that component minus the count of all negative mentions of that component.

$$I_c = count(pos_c) - count(neg_c) \tag{6.1}$$

[9]This was partly due to to the demographics of the participants, whose nationalities ranged from British to Brazilian, though the majority of participants were recruited from UK-based contacts.

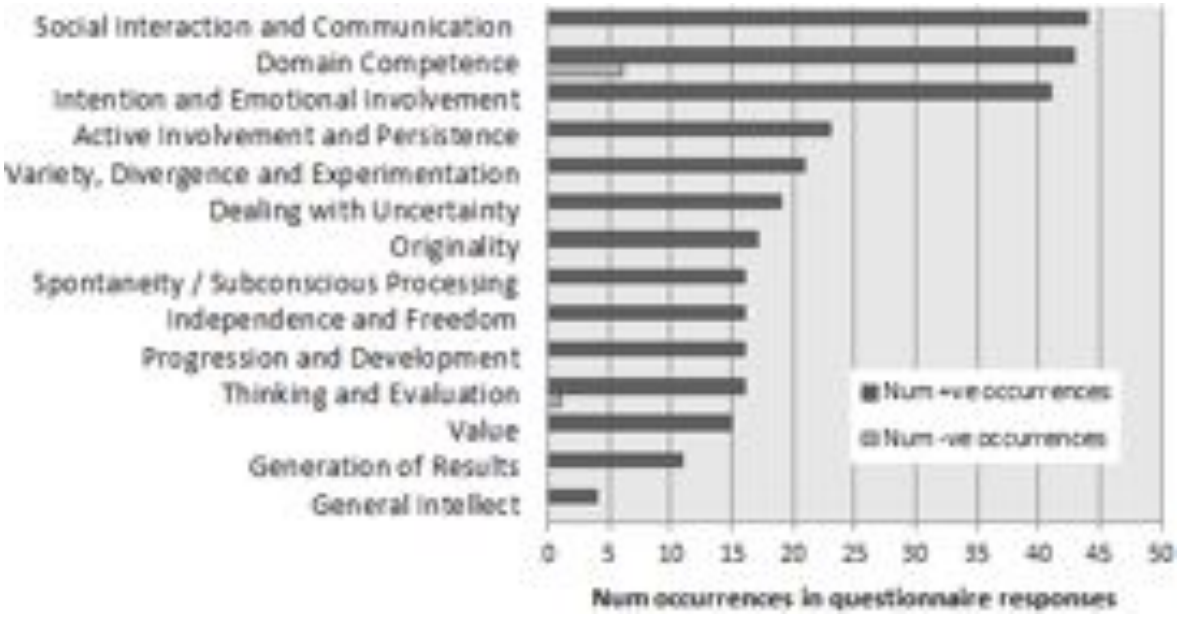


Figure 6.4: Importance and relevance of creativity components to improvisation.

where: I_c = Measured importance of creativity component c,
$count(pos_c)$ = number of positive mentions of that component in the questionnaire,
$count(neg_c)$ = number of negative mentions of that component in the questionnaire.

Figure 6.4 summarises the participants' responses. All components were mentioned by participants to some degree. Two components were occasionally identified as having a negative as well as positive influence. For example, over-reliance on domain competence was seen as detrimental to creativity, though domain competence was generally considered important. Of the 14 components from Chapter 4, those considered most relevant for improvisation were: *Social Interaction and Communication*, *Domain Competence* and *Intention and Emotional Involvement*. The importance counts were converted to weights by calculating the percentage of comments for each component in the sum total of all comments for all components (see Table 6.2).

6.3.3 Further findings in the questionnaire

Equating creativity and musical improvisation

Several respondents equated improvisation with creativity, when asked about creativity in the context of musical improvisation. This was not replicated in responses for the other ten words (or groups of words) given to participants in the questionnaire. Here are some quotes from responses:[10]

'improv the only way I feel that I can be truly creative during live performance'

[10]N.B. All quotes of questionnaire responses are unedited so occasionally contain grammatical/spelling inaccuracies.

Table 6.2: Converting the I_c values into weights representing the importance of each component.

Component	I_c	weight percentage
Social Interaction and Communication	44 - 0 = 44	14.9%
Domain Competence	43 - 6 = 37	12.5%
Intention and Emotional Involvement	41 - 0 = 41	13.9%
Active Involvement and Persistence	23 - 0 = 23	7.8%
Variety, Divergence and Experimentation	21 - 0 = 21	7.1%
Dealing with Uncertainty	19 - 0 = 19	6.4%
Originality	17 - 0 = 17	5.8%
Spontaneity / Subconscious Processing	16 - 0 = 16	5.4%
Independence and Freedom	16 - 0 = 16	5.4%
Progression and Development	16 - 0 = 16	5.4%
Thinking and Evaluation	16 - 1 = 15	5.1%
Value	15 - 0 = 15	5.1%
Generation of Results	11 - 0 = 11	3.7%
General Intellect	4 - 0 = 4	1.4%
	295	100.0%

> 'The word creativity in relation to improvisation is critical, and is the defining word I would use to describe improvisation'
>
> 'The very background impetus for an improvisation. This is what is expressed in every part of an improvisation'
>
> 'Improvisation is fundamentally about creativity'
>
> 'Improvisation is creative by it's very nature'

Creativity definitions

Several respondents offered their own definitions of creativity. Whilst no one standard definition emerged, these definitions show the variety of views on aspects of creativity in musical improvisation.

> 'Originality or doing something different with known elements - producing something new which hasn't been heard before'
>
> 'being yourself. Not conforming to the norm'
>
> 'doing whatever you feel like, following creative impulses'
>
> 'about our ability to organize our thoughts and go with the flow or thoughts in real time'
>
> 'give expression to and trust the heart'
>
> 'Improvising so (1) as to surprise, to be inventive, (2) to seem of worth (= a response like 'now that IS good'!), and (3) still to have a connection or link to the basic line, the tune on which the improvisation is being developed'

6.3.4 Step 2: Standards for evaluating the creativity of musical improvisation systems

Drawing upon the results from Step 1, the musical improvisation systems were evaluated along 14 standards, one for each of the 14 aspects in Chapter 4. Again this follows the recommendations given

in Chapter 5 Section 5.3.

6.3.5 Step 3: Evaluative tests for creativity in musical improvisation systems

There were six judges in total for Case Study 1.[11] Each judge was a musical improviser with knowledge and familiarity of this domain. The judges' improvisation experience collectively covered playing trumpet, saxophone, piano, bass, guitar, drums and laptop, in various genres including jazz, pop, electronica and contemporary music. Each judge had also studied computer programming up to degree level or worked as a programmer, and had studied (at least one degree level course or equivalent) computer music or computational creativity. None of the six judges were involved in the previous questionnaire in Step 1b. Their data was therefore obtained independently of the data collected in that questionnaire.

Each judge evaluated two systems each.[12] For each system, judges had 30 minutes to research and learn about the system. They were given audio (and where available, video) demos of the system in action, a representative paper describing the system and how it works, any available reviews of the system and/or interviews with the system programmers about the system. Judges could also conduct online searches if they wanted to and were given links to relevant websites.

After 30 minutes of research, the judge was interviewed by myself for 15-30 minutes (total time was dependent on how long it took to obtain the evaluation data). During interview, the 14 components were presented to the judge one at a time, in order of descending importance as identified through the questionnaire in Step 1b. The judge gave the systems a rating from 0 (lowest) to 10 (highest) for each component.[13] After evaluating both systems, a final question asked the judge which system overall they found most creative and why.

Judges were trained on the different components before beginning evaluations and were given an on-screen diagram, a print-out of the diagram and an information sheet with details of each component's meaning to refer to during the study.[14] They were also given example comments which collectively represented each component.[15] These comments were taken from quotes in the annotated questionnaire data for Step 1b. These statements were used to help analyse the three musical improvisation systems, for example:

- *How is the system perceived by an audience?* (Social Communication and Interaction).
- *What musical knowledge does the system have?* (Domain Competence).
- *Does the system get some reward from improvisation?* (Intention and Emotional Involvement).

[11] An additional two judges were used in pilot studies, but the data provided in these pilot studies is excluded from the evaluation results presented in this Chapter.

[12] Judges were restricted to evaluating two of the systems rather than all of them due to practical restrictions on time.

[13] Judges were allowed to use ratings of $x.5$ out of 10 if they specifically asked to. Hence the rating scale was effectively a 21-point numeric scale, with 5.0 as the midpoint between the two extremes of 0.0 and 10.0.

[14] The diagrams and information sheet gave the component descriptions from Chapter 4 Section 4.3 and Figure 4.3.

[15] All example comments are listed in Appendix D.

Each system was evaluated in a dedicated 50-60 minute session. Judges would evaluate one system, take a break of 5-10 minutes, then evaluate their second system. Each judge conducted the study individually rather than with other judges. Systems were presented to judges in the order listed below, to ensure that each system was evaluated by the same number of judges and also that each system was considered first by at least one judge and second by at least one judge. Originally this case study investigated four musical improvisation systems, rather than three (see Section 6.2); the fourth system is represented by an asterisk: *.

J1 1st: * , 2nd: GAmprovising.
J2 1st: Voyager, 2nd: * .
J3 1st: GAmprovising, 2nd: GenJam.
J4 1st: * , 2nd: GenJam.
J5 1st: Voyager, 2nd: GAmprovising.
J6 1st: GenJam, 2nd: Voyager.

6.4 Results and discussion

Both quantitative and qualitative data was collected on each system during SPECS evaluation. The quantitative data, in the form of judges' ratings for each component, are presented in Sections 6.4.1 and 6.4.2. The qualitative data gathered, in the form of comments made by judges during interview, are presented in Section 6.4.3.

6.4.1 Judges' evaluation ratings

The full set of evaluation data obtained from the judges is presented in Table 6.3. Some observations can be drawn from the raw data from the judges' ratings, as will be presented in this Section. It should be borne in mind, though, that some components make a more relevant contribution to musical improvisation than others. The ratings will therefore be weighted according to component importance, for Step 1b of SPECS and reported next, in Section 6.4.2.

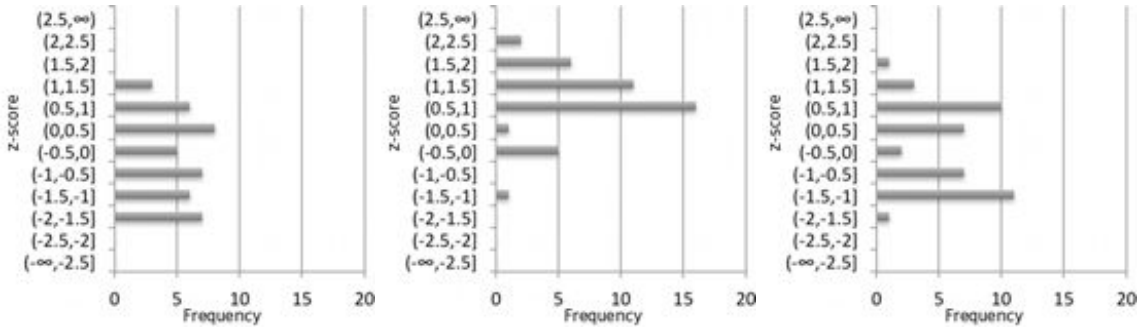
Trends in the judges' ratings

To more easily identify patterns in the ratings per system, the data in Table 6.3 is standardised by calculating z-scores: subtracting the mean (across all three systems) from each rating and dividing the result by the standard deviation (Butler, 1985). Transforming the data using standardisation removes overriding themes in the data, to enable more subtle differences to emerge (Erickson & Nosanchuk, 1992). Figure 6.5 shows the distribution of the z-scores, batched per system.

GAmprovising GAmprovising's z-scores (Figure 6.5(a)) are distributed over a range centred around the mean (a standardised rating of 0), with no standard distributional pattern or peaks.

Table 6.3: Judges' evaluation ratings out of 10 (unweighted) for the three systems in Case Study 1, with information on means and standard deviation (SD) within the data.

System	GAmprovising			GenJam			Voyager		
Judge	J2	J4	J5	J4	J6	J1	J3	J5	J1
Social Interaction & Communication	2	4	1	8	9	8	6	7	9
Intention & Emotional Involvement	3	5	0	6	7	5	6	3	2
Domain Competence	3	6	1	6	7	9	7	6	2
Active Involvement & Persistence	4	7.5	0	6	8	7	3	6	3
Variety, Divergence & Experimentation	4	6	2	4	7	9	3	3	8.5
Dealing with Uncertainty	5	6	0	7	4	10	2	5	5
Originality	4	8	5	6	6	9	2	7	1
Independence & Freedom	5	3	2	4	8	9	2	6	5
Progression & Development	5	7	3	6.5	7	9	1	2	0
Spontaneity/Subconscious Processing	4	8	0	6.5	8	2	5	7	5
Thinking & Evaluation	1	5	0	7	6	8	1	1	4
Value	3	5	0	8	8	4	8	6	5
Generation of Results	5	7	3	8	10	7	3	8	5
General Intellect	3	6	0	4	8	8.5	1	4	3
Mean rating (1dp)	3.6	6.0	1.2	6.2	7.4	7.5	3.6	5.1	4.1
Mean of all ratings (1dp)		3.6			7.0			4.3	
SD (1dp)	1.2	1.5	1.6	1.4	1.4	2.3	2.4	2.1	2.6
SD of all ratings (1dp)		2.4			1.8			2.4	



(a) GAmprovising. (b) GenJam. (c) Voyager.

Figure 6.5: The distribution of the z-scores, batched per system. Ratings are standardised by subtracting the mean and then dividing by the standard deviation of the whole set of ratings. The x-axis represents the number of z-scores in a particular interval and the y-axis gives details of each interval. N.B. $(a,b]$ in interval notation denotes, for a number x in that interval, that $a < x \leq b$.

GenJam It can be seen that the z-scores of GenJam's ratings (Figure 6.5(b)) are mostly about 0. The majority of GenJam's z-scores are only slightly higher than 0, with the distribution of ratings tailing off as ratings become higher and higher.

Voyager Figure 6.5(c) shows the distribution of Voyager's z-scores to be bimodal, with a peak at the interval (0.5, 1] and a similar peak at (-1.5, -1] and a trough at (-0.5, 0], which is roughly the mean

point. This means that in general, judges tended to give Voyager marks some way above or below the mean, but were less likely to give Voyager average marks.[16]

Overall distribution of the judges' ratings

Figure 6.6 shows the judges use of the rating scale, via the frequency with which each rating was assigned by judges. In total, the three systems were rated on 14 components, each by three judges, for a total of 126 ratings. The frequency distribution shown in Figure 6.6 is not particularly skewed in either a positive or negative direction, showing that overall the judges were not biased towards lower or higher ratings out of 10, or to a particular subset of ratings within the rating scale.

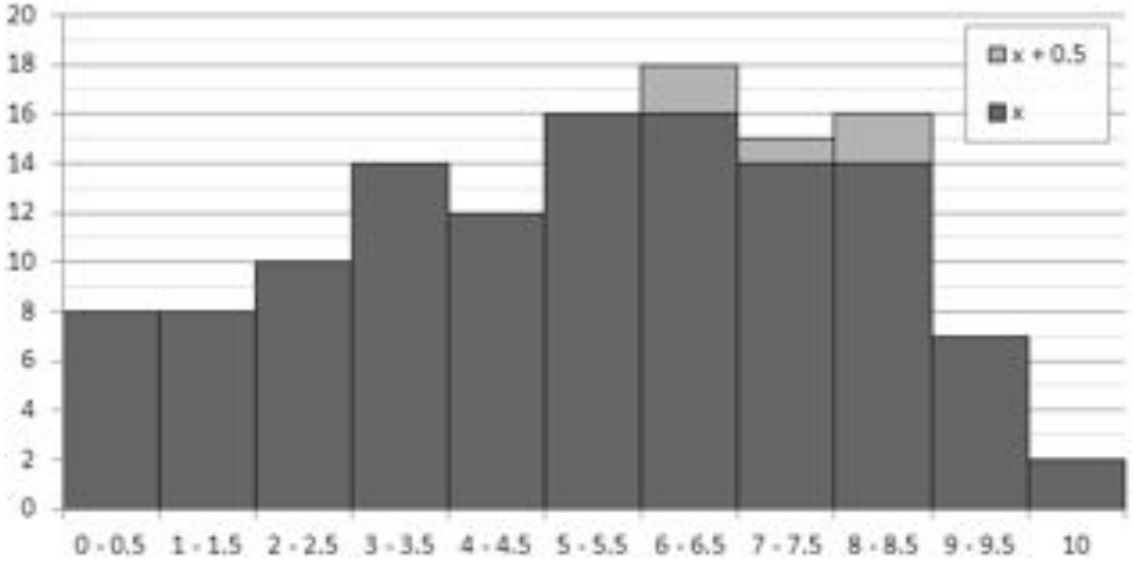


Figure 6.6: Graph showing the frequency distribution of judges' ratings. Most ratings were given in whole numbers. Ratings of the format x.5 are marked separately in this stacked bar chart.

Overall means and standard deviations in the ratings data

This Section focuses on the last four rows of Table 6.3. All observations about means are preliminary, pending further investigation after component ratings have been weighted.

If taking a single evaluative score for each of the three systems,[17] the mean of the ratings across all components and all judges for each system could be used. These means show that GenJam has been rated considerably higher than the other systems, with a mean rating (7.0) almost double the other mean ratings. The standard deviation for all ratings of a system combined shows that judges

[16] One could speculate that the bimodality of the ratings reflected a split in judges' opinions about the avant-garde style of free improvisation (Collins, 2011, personal communication).

[17] A single summative evaluation score would however provide no formative feedback on strengths and weaknesses of individual systems for future development and learning from evaluation.

agreed more about GenJam (1.8) than they did for the other systems. The spread of mean ratings also reflects this. GAmprovising showed the largest difference in overall opinions, with a difference of 4.8 between highest and lowest ratings received. GenJam showed the most consistency in opinions over all components, with a difference in judges' highest and lowest mean ratings of only 1.3.

Looking at the mean ratings across all components for a judge and for each system, GenJam received the highest two mean ratings. GAmprovising received the next highest rating, although it also received the lowest overall rating.

Overall, individual judges' ratings for each component of a system varied the most for Voyager, where the standard deviation across each judge's component ratings was greater than 2.0 in each of the three cases. GAmprovising showed the least variance, with no standard deviation per judge's ratings greater than 1.6. This may be due to the lower ratings overall given to GAmprovising.

Inspection of the judges' given ratings per component

If the evaluative information in Table 6.3 is reduced to a set of summary values that does not distinguish between components, valuable information available from these ratings will be overlooked.

To more easily see trends within individual component ratings for Table 6.3, component ratings' averages (mean and median) and standard deviation are presented in Table 6.4 and Table 6.5 respectively. Figure 6.8(a) gives a graphical comparison of the mean ratings across systems.

Table 6.4: Mean and median ratings (unweighted) for each system across all judges' ratings. Mode averages are not as useful here as there are only three ratings per component. Abbreviations: All = All systems, GA = GAmprovising, GJ = GenJam, V = Voyager.

Type of average	Mean				Median			
System	GA	GJ	V	*All ratings*	GA	GJ	V	*All ratings*
Social Interaction & Communication	2.3	8.3	7.3	*6.0*	2	8	7	*5.7*
Intention & Emotional Involvement	2.7	6.0	3.7	*4.1*	3	6	3	*4*
Domain Competence	3.3	7.3	5.0	*5.2*	3	7	6	*5.3*
Active Involvement & Persistence	3.8	7.0	4.0	*4.9*	4	7	3	*4.7*
Variety, Divergence & Experimentation	4.0	6.7	4.8	*5.2*	4	7	3	*4.7*
Dealing with Uncertainty	3.7	7.0	4.0	*4.9*	5	7	5	*5.7*
Originality	5.7	7.0	3.3	*5.3*	5	6	2	*4.3*
Independence & Freedom	3.3	7.0	4.3	*4.9*	3	8	5	*5.3*
Progression & Development	5.0	7.5	1.0	*4.5*	5	7	1	*4.3*
Spontaneity/Subconscious Processing	4.0	5.5	5.7	*5.1*	4	6.5	5	*5.2*
Thinking & Evaluation	2.0	7.0	2.0	*3.7*	1	7	1	*3*
Value	2.7	6.7	6.3	*5.2*	3	8	6	*5.7*
Generation of Results	5.0	8.3	5.3	*6.2*	5	8	5	*6*
General Intellect	3.0	6.8	2.7	*4.2*	3	8	3	*4.7*

Mean ratings of components

Overall, Figure 6.8(a) shows that GenJam performs best in 13 out of 14 components. For *Spontaneity and Subconscious Processing*, Voyager is rated 0.2 higher on average than GenJam. This difference is minimal though and GenJam still attracts ratings for this component that average above the midpoint of 5 out of 10. Between the other two systems, there is some crossover and neither system consistently outperforms the other, though Voyager is more often the superior system per component. Statistical tests are in general not applicable due to the limited amount of data.

Some individual feedback on performance can be obtained from looking at the unweighted ratings. Table 6.4 shows each system's relative strengths and weaknesses via component performance. It should be remembered that comments on individual components become more or less relevant to creativity in musical improvisation after the components have been weighted. Also, the ratings for systems were often affected by two contrasting opinions, for example Judge 5 ranked GAmprovising nearly 5 points lower on average than Judge 4. Median ratings for each component should therefore also be checked, where extremes of opinion impact less on overall averages per component.[18]

GAmprovising GAmprovising's highest mean rating is for *Originality*, for which it scores a mean of 5.7. Most of GAmprovising's mean ratings fell on or below 4.0.

1. Strengths:[19] *Originality* (5.7), *Generation of Results*, *Progression and Development* (both 5.0).
2. Weaknesses: *Thinking and Evaluation* (2.0), *Social Interaction and Communication* (2.3).

GenJam GenJam's mean ratings were considerably higher than the other systems. All component ratings but one for GenJam were above the maximum mean rating achieved by GAmprovising.

1. Strengths: *Social Interaction and Communication* (8.3), *Generation of Results* (8.3).
2. Weaknesses: *Spontaneity and Subconscious Processing* (5.5), *Intention and Emotional Involvement* (6.0).

Voyager Voyager achieved the highest rating of the three systems for *Spontaneity and Subconscious Processing* (5.7). Detracting from this, Voyager also received the lowest mean rating achieved by any system: 1.0 for *Progression and Development*. The low mean rating for *Intention and Emotional Involvement* (3.7) is notable because Lewis (2000) devotes a healthy section of the paper to the emotional aspects of the system and would presumably treat it as a priority.

1. Strengths: *Social Interaction and Communication* (7.3), *Value* (6.3).
2. Weaknesses: *Progression and Development* (1.0), *Thinking and Evaluation* (2.0).

[18] Take for example the median rating for *Dealing with Uncertainty* for GAmprovising. This was higher than the mean, at 5 rather than 3.7. This was due to the removal of influence of a rating of 0 / 10 for this component from Judge 5, who gave GAmprovising seven ratings of 0 / 10 in total.

[19] Strengths and weaknesses in this system are picked out relative to the component performance for a particular system. Mean ratings are given in brackets alongside the selected components, for comparison between systems.

Commonalities across all systems All systems received ratings above 5.0 / 10.0 for *Generation of Results* (6.2 overall mean). Means across all ratings for a component also exceeded 5.0 for *Generation of Results* (6.2), *Social Interaction and Communication* (6.0), *Originality* (5.3), *Domain Competence* (5.2), *Variety, Divergence and Experimentation* (5.2), *Value* (5.2) and *Spontaneity/Subconscious Processing* (5.1). There were however no components for which all systems performed very well, with no means across all ratings over 6.2.

Components which all or most systems performed relatively badly at were: *Thinking and Evaluation*, with a mean of 3.7 across all systems, where the poor scores (2.0) of two systems were boosted somewhat by GenJam's 7.0 mean for this component; *Intention and Emotional Involvement* (4.1 mean across all systems) and *General Intellect* (4.2 mean across all systems), that were raised by GenJam's mean rating of 6.0 and 6.8 respectively; and *Progression and Development* at 4.5 mean, again boosted by the higher mean rating of GenJam (7.5) and pulled lower by Voyager's mean (1.0).

All systems attracted differences in opinions between judges. Across three systems and fourteen components each (a total of 42 sets of component ratings), there were only 6 cases where judges' individual ratings per component for a system were within 2 points of each other, 1 case where judges' ratings were within 1 point of each other, and no cases where all three judges gave exactly the same ratings for one component on a particular system. Given that there were 21 possible values to choose from in the rating scale from 0 to 10 (including .5 values), the last point is unsurprising. The frequency of ratings differing by larger intervals, however, suggests that the median average would also be worth examining.

Notable differences between systems Looking at Figure 6.8(a), there is a clear gap between GenJam and all other systems in almost all components. The systems are most spread out for *Progression and Development* and *Social Interaction and Communication*, with a range of 6.5 and 6 points respectively between highest (GenJam in both cases) and lowest (Voyager / GAmprovising respectively) mean ratings. Other discrepancies between mean ratings are found for *Thinking and Evaluation* (where mean ratings per system disagree by a maximum of 5.0), *Intention and Emotional Involvement* (4.3), *Value* (4.0), *General Intellect* (4.1) and *Domain Competence* (4.0). In all these cases GenJam has the highest mean score and the lowest scoring system varies: GAmprovising/Voyager (=), GAmprovising, GAmprovising, Voyager and GAmprovising, respectively.

Median ratings of components

As has been noted above, ratings for each system varied considerably across the three judges. Taking a median rating in the case of three judges is the same as discarding the highest and lowest rating and using the middle rating. This view of the data is less affected by outlying data, so occasionally differs from the view obtained from taking means of ratings and is worth investigating.

Figure 6.8(c) tells much the same story as 6.8(a), with an even greater distinction between GenJam and the other systems. The other two systems are again mixed in performance. Again the ratings for GAmprovising average out at 5 or below, with the best performing components from before all scoring a median of 5. The only component with any noticeable change in performance for GAmprovising after using medians rather than means to average the ratings is *Dealing with Uncertainty*, for which GAmprovising can now be said to demonstrate an middling level of performance (5). The median ratings for GenJam give an even more flattering view of the system, with generally higher medians than means.There are only minor changes in ratings for GenJam. Its median rating for *Spontaneity and Subconscious Processing* (6.5) is now the highest rating for this component of the three systems, replacing Voyager as the system performing best on this component. The highest and lowest rated components for Voyager remain the same whether using means or medians. Generally, Voyager's lowest rated components (using means) became lower rated on average when using medians.

Commonalities across all systems If all the ratings for a component are averaged using medians, then again *Generation of Results* is the only component that all three systems perform at least adequately well on, with a median across systems of 6. This is the highest median across all ratings. The components which received the lowest overall median ratings across all judges differ little from the observations for the mean data: *Thinking and Evaluation* (3), *Intention and Emotional Involvement* (4), *Originality* (4.3) and *Progression and Development* (4.3).

Notable differences between systems The median ratings show how the systems are ordered per component, as can be seen in Figure 6.8(c). GenJam outperformed the other systems on several components, including *Intention and Emotional Involvement* (3 points higher), *Active Involvement and Persistence* (3 or 4 points higher), *Variety, Divergence and Experimentation* (3 or 4 points higher), *Dealing with Uncertainty* (2 points higher), *Independence and Freedom* (3 or 5 points higher), *Progression and Development* (2 or 6 points higher, *Thinking and Evaluation* (6 points higher), *Value* (2 or 5 points higher), *Generation of Results* (3 points higher) and *General Intellect* (5 points higher). GenJam and Voyager achieved considerably higher median ratings for *Social Interaction and Communication* than GAmprovising (8 and 7 respectively, compared to 2). GenJam and Voyager also outperformed GAmprovising for *Domain Competence* (7 and 6 respectively, compared to 3). For *Originality* GenJam and GAmprovising outperformed Voyager (6 and 5 respectively, compared to 2).

Standard deviation in the ratings data

Table 6.5 show how standard deviation ratings varied across systems and across components. The larger the standard deviation, the more variety and disagreements there were in the judges' ratings for that system and that component. The largest standard deviations were found for:

1. GAmprovising's *Spontaneity and Subconscious Processing* (4.0), *Active Involvement and Per-*

sistence (3.8), *Dealing with Uncertainty* and *General Intellect* (both 3.0).

2. GenJam's *Spontaneity and Subconscious Processing* (3.1) and *Dealing with Uncertainty* (3.0).
3. Voyager's *Variety, Divergence and Experimentation* (3.2) and *Originality* (3.2).

Spontaneity and Subconscious Processing and *Dealing with Uncertainty* appear more than once in the above list, indicating a greater propensity for judges to disagree about a system on this component.

Table 6.5: Standard deviation (to 1dp) for each system across all judges' (unweighted) ratings.

System	GAmprovising	GenJam	Voyager	All
Social Interaction and Communication	1.5	0.6	1.5	3.0
Intention and Emotional Involvement	2.5	1.0	2.1	2.3
Domain Competence	2.5	1.5	2.8	2.6
Active Involvement and Persistence	3.8	1.0	2.0	2.6
Variety, Divergence and Experimentation	2.0	2.5	2.1	2.5
Dealing with Uncertainty	3.2	3.0	2.0	2.8
Originality	2.1	1.7	2.6	2.6
Independence and Freedom	1.5	2.6	2.6	2.5
Progression and Development	2.0	1.3	1.7	3.1
Spontaneity/Subconscious Processing	4.0	3.1	1.5	2.7
Thinking and Evaluation	2.6	1.0	2.3	3.0
Value	2.5	2.3	1.0	2.7
Generation of Results	2.0	1.5	3.1	2.4
General Intellect	3.0	2.5	3.0	2.9

For GAmprovising, standard deviation across components ratings was relatively large. This shows that judges tended to disagree to quite an extent over component ratings for GAmprovising. Generally, smaller standard deviations were recorded for GenJam than for the other systems.

If standard deviations are collectively considered per component, as in the last column of Table 6.5, the largest standard deviations recorded overall and therefore the largest differences between judge opinions were for *Progression and Development* (3.1), *Social Interaction and Communication* (3.0) and *Thinking and Evaluation* (3.0). Overall, the lowest standard deviations are for *Intention and Emotional Involvement* (2.3), *Generation of Results* (2.4) and *Variety, Divergence and Experimentation* (2.5), meaning that in general judges' opinions were more in agreement for these components.[20]

6.4.2 Weighting the judges' ratings according to component importance

As remarked above, the data from the judges' ratings can only be interpreted to some extent; Section 6.3.2, performing Step 1b, shows that some components make a much greater contribution to musical creativity than others. Component ratings were therefore weighted using the numeric weights

[20]This may also indicate that these components were more easily understood than the others, however this speculation cannot be confirmed or denied using solely the standard deviation information as evidence. Chapter 8 looks at judges' comments during the evaluation process, and also notes the information on standard deviation to some extent, to investigate how well the components were understood by judges.

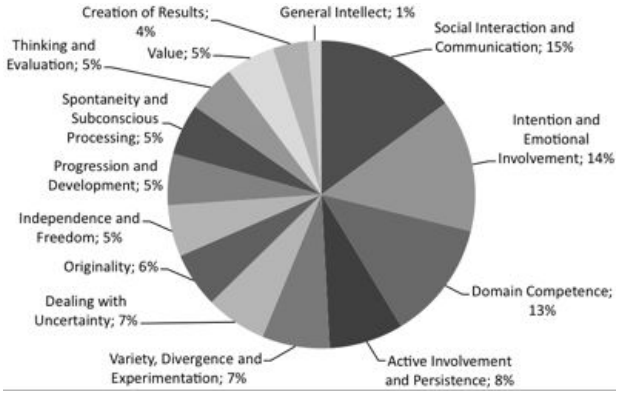


Figure 6.7: The contribution of each of the 14 components towards creativity in musical improvisation, illustrating the percentages calculated in Step 1b of this case study.

obtained in Step 1b (Table 6.2, illustrated in Figure 6.7). Differences in more important components became magnified, with differences in less important components reduced. The updated (weighted) rating results and rating averages (mean and median) are reproduced in Tables 6.6 and 6.7 respectively. A graphical representation of the weighted ratings is given in Figure 6.8(b).

Table 6.6: Judges' evaluation ratings (weighted and rounded to 1dp) for the three systems in Case Study 1: an update of Table 6.3. N.B. The weighted ratings may now exceed 10 despite the original data being ratings out of a maximum of 10. The maximum possible value for a rating is now equivalent to its weight, given in brackets after the component name (e.g. 5.1 for *Value*).

System	GAmprovising			GenJam			Voyager		
Judge	J2	J4	J5	J4	J6	J1	J3	J5	J1
Social Interaction and Communication (14.9)	3.0	6.0	1.5	11.9	13.4	11.9	8.9	10.4	13.4
Intention and Emotional Involvement (13.9)	4.2	7.0	0.0	8.3	9.7	7.0	8.3	4.2	2.8
Domain Competence (12.5)	3.8	7.5	1.3	7.5	8.8	11.3	8.8	7.5	2.5
Active Involvement and Persistence (7.8)	3.1	5.9	0.0	4.7	6.2	5.5	2.3	4.7	2.3
Variety, Divergence and Experimentation (7.1)	2.8	4.3	1.4	2.8	5.0	6.4	2.1	2.1	6.0
Dealing with Uncertainty (6.4)	3.2	3.8	0.0	4.5	2.6	6.4	1.3	3.2	3.2
Originality (5.8)	2.3	4.6	2.9	3.5	3.5	5.2	1.2	4.1	0.6
Independence and Freedom (5.4)	2.7	1.6	1.1	2.2	4.3	4.9	1.1	3.2	2.7
Progression and Development (5.4)	2.7	3.8	1.6	3.5	3.8	4.9	0.5	1.1	0.0
Spontaneity & Subconscious Processing (5.4)	2.2	4.3	0.0	3.5	4.3	1.1	2.7	3.8	2.7
Thinking and Evaluation (5.1)	0.5	2.6	0.0	3.6	3.1	4.1	0.5	0.5	2.0
Value (5.1)	1.5	2.6	0.0	4.1	4.1	2	4.1	3.1	2.6
Generation of Results (3.7)	1.9	2.6	1.1	3.0	3.7	2.6	1.1	3 .0	1.9
General Intellect (1.4)	0.4	0.8	0.0	0.6	1.1	1.2	0.1	0.6	0.4
Mean rating (1dp)	2.4	4.1	0.8	4.5	5.3	5.3	3.1	3.7	3.1
Mean across judges (1dp)	2.4			5.0			3.3		

Standard deviation information is not revisited here; components are no longer measured on a common scale (out of 10) but are adjusted according to individual weights. Standard deviation measurements therefore become meaningless for comparison as they will be affected by this adjustment.

To compare the effects of weighting the component ratings by importance, Figure 6.8 shows the mean and median ratings for each system before and after weighting. Observations made about the unweighted ratings can be considered post-weighting to see their relevance to creativity in musical improvisation. Other observations also become more apparent post-weighting, for components of high importance in this domain.

Overall means in the ratings data

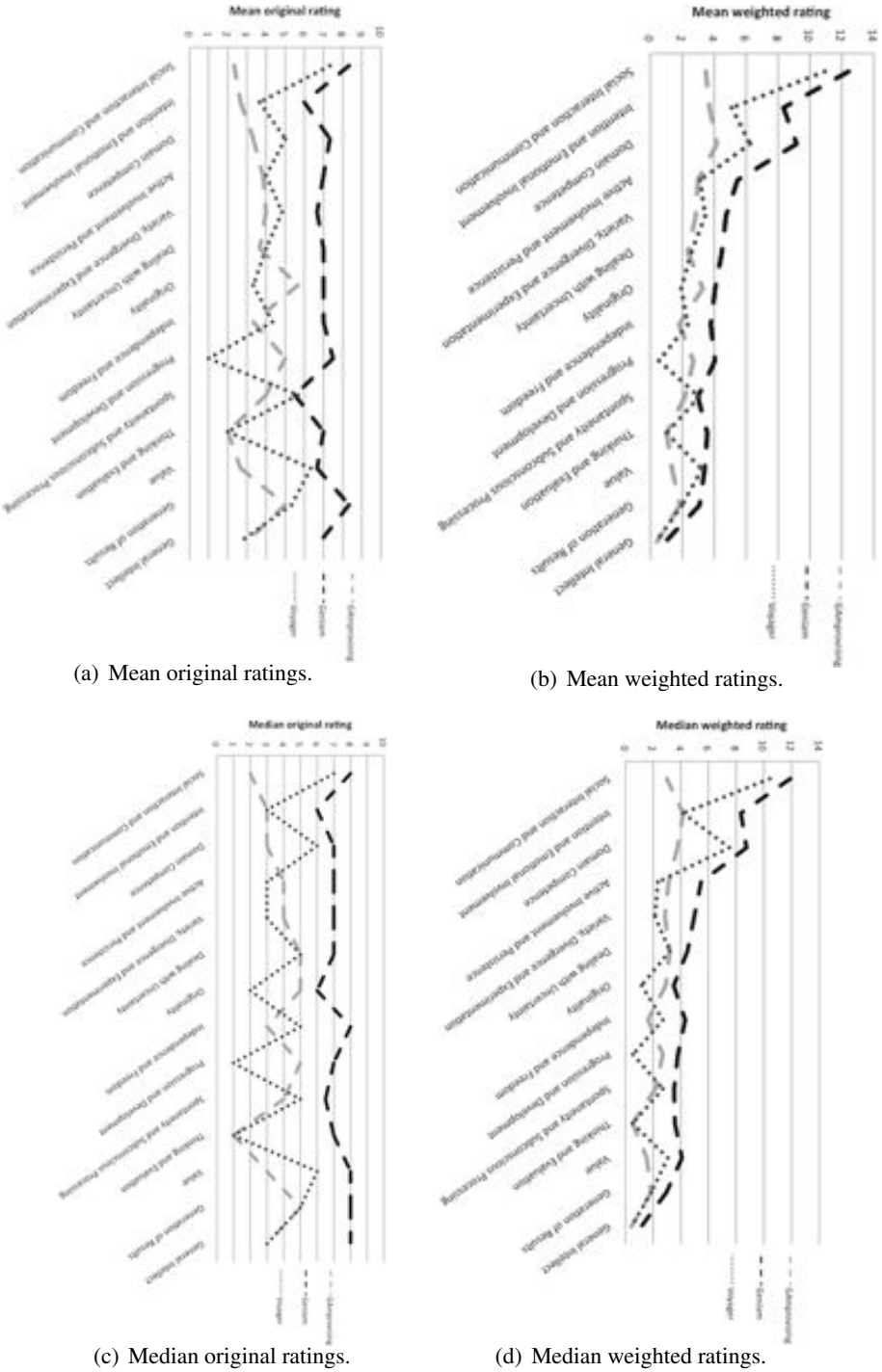
Table 6.7: Mean and median ratings (weighted) for each system across all judges' ratings: an update of Table 6.4. Abbreviations: All = All systems, GA = GAmprovising, GJ = GenJam, V = Voyager.

Type of average	Mean				Median			
System	GA	GJ	V	*All*	GA	GJ	V	*All*
Social Interaction and Communication (14.9)	3.5	12.4	10.9	*7.8*	3.0	11.9	10.4	*6.7*
Intention and Emotional Involvement (13.9)	3.7	8.3	5.1	*4.7*	4.2	8.3	4.2	*4.2*
Domain Competence (12.5)	4.2	9.2	6.3	*6.6*	3.8	8.8	7.5	*6.9*
Active Involvement and Persistence (7.8)	3.0	5.5	3.1	*3.7*	3.1	5.5	2.3	*3.1*
Variety, Divergence & Experimentation (7.1)	2.8	4.7	3.4	*3.4*	2.8	5.0	2.1	*2.5*
Dealing with Uncertainty (6.4)	2.3	4.5	2.6	*2.8*	3.2	4.5	3.2	*3.2*
Originality (5.8)	3.3	4.1	1.9	*2.9*	2.9	3.5	1.2	*2.8*
Independence and Freedom (5.4)	1.8	3.8	2.3	*2.5*	1.6	4.3	2.7	*2.2*
Progression and Development (5.4)	2.7	4.1	0.5	*2.1*	2.7	3.8	0.5	*1.6*
Spontaneity/Subconscious Processing (5.4)	2.2	3.0	3.1	*2.4*	2.2	3.5	2.7	*2.4*
Thinking and Evaluation (5.1)	1.0	3.6	1.0	*1.7*	0.5	3.6	0.5	*0.5*
Value (5.1)	1.4	3.4	3.2	*2.9*	1.5	4.1	3.1	*3.3*
Generation of Results (3.7)	1.9	3.1	2.0	*2.3*	1.9	3.0	1.9	*2.2*
General Intellect (1.4)	0.4	1.0	0.4	*0.6*	0.4	1.1	0.4	*0.6*

Previous observations about means can be investigated further, now that component ratings have been weighted. Again, GenJam is rated higher overall (5.0), if looking at the mean across judges in the final row of Table 6.6. The systems are in the same descending order according to this mean with a similar spread, with a gap between GenJam and the next highest rated system, Voyager (3.3), which is then followed by GAmprovising (2.4).

GAmprovising mean ratings per judge are still the most spread out, ranging from 0.8 to 2.5 for a spread of 1.7. Voyager and GenJam are still second and third most spread out, but swop places, with a range of 0.8 for GenJam (4.5 - 5.3) and 0.6 for Voyager (3.1 - 3.7). Varying opinions between judges on less important components now has less impact on the overall average per judge per system, whereas differences in opinions for more important components have more effect on the overall range of judge opinions. What this indicates is that the discrepancies in opinions for GAmprovising and Voyager occur more in components which are relatively unimportant for musical improvisational creativity, whereas discrepancies for GenJam occur on more important components.

Previously GenJam had received the highest two mean ratings from judges, with GAmprovising



(a) Mean original ratings.

(b) Mean weighted ratings.

(c) Median original ratings.

(d) Median weighted ratings.

Figure 6.8: Mean and median averages for all judges' evaluation ratings (before and after weighting by component importance) for the three systems.

receiving the next highest mean rating. Now, GenJam receives all three of the highest mean ratings (5.3, 5.3, 4.5), with the next highest mean rating belonging to GAmprovising (4.1). Again, GAmprovising also received the lowest overall mean rating (0.8 from Judge 5). Voyager's lower marks have been de-emphasised, so it now receives higher mean ratings from each judge compared to the lower GAmprovising mean ratings.

Inspection of the judges' weighted ratings and the weights per component

The magnitude of ratings is now closely linked to how important that particular component is for creativity. For example, in *Social Interaction and Communication*, *Interaction and Emotional Involvement* and *Domain Competence* we see ratings of double figures, particularly for GenJam. On the other hand, the ratings for *General Intellect* is weighted to be a maximum of only 1.4, rather than 10. Figure 6.7 shows how components are weighted relative to each other.

Mean ratings of components

Using Figure 6.8(b) to illustrate trends within the data in Table 6.6, the higher performance of GenJam becomes more apparent now the data is weighted. The component where GenJam is outperformed by other systems (*Spontaneity and Subconscious Processing*) is relatively less important for creativity in this domain, so the difference between GenJam and Voyager becomes less marked. As before, ratings for Voyager and GAmprovising often overlap each other.

Looking at individual systems, the weighted rating now represents a combination of two aspects: the performance of the system on that component and the importance of that component. Feedback from evaluation can be extracted accordingly. To improve a system's musical improvisational creativity, more important components to address first are those with higher weightings. For example, improving the social abilities of GAmprovising could raise its mean rating for this component from 3.5 to a maximum of 14.9, for a potential gain of 11.4 points. In contrast, even if the system's general intelligence was raised to the maximum extent possible, this improvement would only see a change in mean rating from 0.4 to 1.4, for a potential gain of 1.0 point, more than ten times smaller than the potential gain from improving the system's interaction and communication.

To this end, the following feedback makes recommendations for improvements which could see significant gains in evaluative ratings, de-prioritising less highly-weighted components. Suggestions for how to improve creativity can be also inspired in some cases by the judges' comments on the systems, which are reported per component in Section 6.4.3.

GAmprovising GAmprovising's highest mean rating was originally for *Originality*; however after weighting, the largest contributor to its musical improvisation creativity is its *Domain Competence*. None of its ratings are particularly high and it is notably weak on the most important components, with a mean rating of between 3.5 - 4.2 for each of the top three components rather than the double

figure ratings seen in some of the other systems. Improving performance on these three components could drastically increase GAmprovising's creativity as perceived by the judges.

GenJam GenJam's weighted ratings were highest on average for all but two components: *Spontaneity and Subconscious Processing* and *Value*, two relatively unimportant components in this domain. It scored particularly well for *Social Interaction and Communication* (12.4). The areas where it could make most gain in terms of mean ratings are *Intention and Emotional Involvement* (potential gain of 5.6), *Domain Competence* (potential gain of 3.3) and to some extent, *Social Interaction and Communication* (potential gain of 2.5). As for GAmprovising, devoting attention to improving the top three components would be most effective, though it would not see the drastic improvement that GAmprovising would (a total potential gain of 11.4 as opposed to 29.9 for GAmprovising).

Voyager Voyager's highest mean rating after weighting was for *Social Interaction and Communication* (10.9). Although previously Voyager achieved the highest rating of the three systems for *Spontaneity and Subconscious Processing*, the difference between Voyager and GenJam on this component after weighting was reduced to only 0.1. An emphasis on the three most important components would give Voyager a potential points gain of 19, particularly through improvements in *Intention and Emotional Involvement* (for a gain of up to 8.8 points) and *Domain Competence* (6.2 points maximum to be gained). Given Lewis's attention to emotion within Voyager (Lewis, 2000), further development on Voyager's emotional involvement and intention could take an interesting direction, though it is unlikely Voyager will undergo any significant future work (Lewis, 2011).

Commonalities across all systems One recommendation for improvement was common to all systems: improvements in *Social Interaction and Communication*, *Intention and Emotional Involvement* and *Domain Competence* will reap greatest rewards in improving creativity. On the other hand, work in improving aspects such as *General Intellect* of the system and the ability of the system for *Generation of Results* is less important for creativity in musical improvisation. Improvements in such areas are likely to be minimal compared to improvements in the more important components.

Notable differences between systems GenJam stood out as superior to the other systems, overall. This is partly due to its high mean ratings for *Social Interaction and Communication* and *Domain Competence*, although it should be remembered that it was rated higher on average than the other two systems for 13 of the 14 components. The other two systems seem to follow different patterns of ratings. GAmprovising has less peaks and troughs in the mean ratings trend shown in Figure 6.8(b), instead following a general descent in mean rating as the components become less important (with one or two exceptions, such as *Originality*, a peak, and *Thinking and Evaluation*, a trough).

Median ratings of components

Weighting the components according to importance has meant that analysis becomes focused on general trends in the data as components change in importance. As a consequence, individual variances in components become less notable except for the most important components. Taking median rather than mean averages of the data therefore reveals fewer observations of note. One thing that can be noted, from comparing Figures 6.8(c) and 6.8(d) (median ratings before and after weighting) is that the significant gap between median ratings of GenJam and those of other systems is reduced somewhat in significance after weighting. GenJam is still clearly the highest performing system on all 14 components when using medians to average ratings, however the margin between it and the other two systems becomes less noticeable after weighting.

6.4.3 Qualitative feedback on the systems

The judges were encouraged to voice their thoughts as they decided on ratings for each system. As a result, qualitative feedback could be gathered as formative feedback for each system, with most of the components attracting some comments. Judges' comments are summarised below:

GAmprovising (Judges 2, 4 and 5)

- *Social Interaction and Communication* was described as 'pretty minimal' (Judge 2), without much of its own *Intention and Emotional Involvement*.
- *Domain Competence* was limited to knowledge of the blues scale. 'Apart from that its a "blank slate", taught knowledge by the user' (Judge 2). Judge 4 felt more positively than the other judges about the system's ability to demonstrate musical abilities.
- *Active Involvement and Persistence* was generally praised due to the genetic algorithm approach, though Judge 5 criticised the lack of temporal knowledge in the performance.
- The system was seen as quite experimental (*Variety, Divergence and Experimentation*) but in a trivial and user-controlled way.
- As a system it was seen to be good at *Dealing with Uncertainty*, although this was not extended to the individual Improvisers.
- Opinions were divided on *Originality* with Judge 4 seeing it as being very good at being surprising and unique, but Judges 2 and 5 only seeing trivial originality, again controlled by the user's tastes.
- The system was seen to have little *Independence and Freedom*, being constrained by the genetic algorithmic method programmed in and the user's preferences.
- For *Progression and Development* distinctions were made by Judge 5 between the individual improviser, which cannot progress, and the system, which can. The other judges took a more

system-centric view, seeing some building of knowledge between improvisations.

- *Spontaneity/Subconscious Processing* was either seen as being able to wait for moments of inspiration (Judges 2 and 4) or to be entirely irrelevant for the system (judge 5).
- *Thinking and Evaluation* was noted as absent; this was seen as a flaw in GAmprovising needing addressing.
- Opinions were again divided as to GAmprovising's *Value*, either seeing something to appreciate (Judge 4) or dismissing it as worthless randomness (Judge 5).
- *Generation of Results* was praised by judges 2 and 4, though Judge 5 criticised its ability to recognise its own products as complete.
- *General Intellect* was mostly rated without comment, except for Judge 5's mentioning that they saw no attempt by GAmprovising to be intelligent.

GenJam (Judges 1, 4 and 6)

- *Social Interaction and Communication* in GenJam was praised highly, particularly in how it responds to what it hears.
- While attracting reasonable ratings for this component, GenJam was felt to reflect the *Intention and Emotional Involvement* of the human player rather than those inherent to the system.
- *Domain Competence* was also highly praised, with GenJam seeing as possessing a lot of relevant musical knowledge.
- *Active Involvement and Persistence* was seen as good though there was doubt as to whether it would become aware of problems occurring.
- GenJam was seen to be able to diverge quite a lot (*Variety, Divergence and Experimentation*) but Judges 4 and 6 commented that its variation was limited by its programming.
- Judges reported difficulty with rating *Dealing with Uncertainty* due to lack of examples in the information they had been given. Ratings varied because of this.
- The level of *Originality* was considered to be fairly high by Judge 1, but Judges 4 and 6 raised questions as to the extent to which GenJam could be original.
- *Independence and Freedom* in GenJam was seen as high in the autonomous version (Biles, 2007) although it needs training input beforehand.
- *Progression and Development* was noted by all three judges in the context of the solo and overall, due to the use of genetic algorithm techniques.
- GenJam was seen to be fairly spontaneous within its programmed limits.
- *Thinking and Evaluation* was seen as being the user's responsibility, not the systems, and that the system could perform better for this, though it was able to constantly monitor what it was doing and behave rationally.

- *Value* was generally perceived as high, though Judge 1 quickly found the solos to become boring. Judge 6 was interested in playing with the system to practice improvising.
- The ability to generate end products was praised for *Generation of Results*.
- To some degree GenJam demonstrated *General Intellect*, through awareness of taste and alternative modes of thought. Judge 1 commented: 'it could be nice if it decides what version to use, out of the different versions of the system by involving all different algorithms'. Judge 4 was unconvinced by GenJam's intelligence, unlike the other two judges.

Voyager (Judges 1, 3 and 5)

- *Social Interaction and Communication* was seen as a major achievement of the system, especially its ability to communicate with other performers.
- Voyager was seen to follow the *Intention and Emotional Involvement* of the other performers by Judge 5. Judge 3 thought it had its own style but Judge 1 held the opposite opinion.
- *Domain Competence* was seen as limited but Voyager was judged to possess the right type of skills for this style of improvisation.
- Voyager was sometimes interpreted by judges as being passive in interaction (*Active Involvement and Persistence*), though Judge 5 saw that Voyager could improvise on its own and make appropriate use of silence, rather than pausing because it was no longer involved.
- *Variety, Divergence and Experimentation* was rated highly by Judge 1 but perceived as limited by Judges 3 and 5.
- *Dealing with Uncertainty* was exhibited to some extent by Voyager but judges felt this was more to do with the style of music rather than any particular programming within the system.
- The *Originality* within Voyager divided judges' opinion and seemed to be heavily linked to style. The freedom of Voyager prompted this comment from Judge 1: 'all it can do is anything'.
- *Independence and Freedom* was demonstrated to a limited degree within the free style of the improvisation but the system was seen as being unable to break its constraints independently.
- *Progression and Development* was viewed poorly by all judges, due to the 5-7 second time frame controlling Voyager's improvising and the lack of knowledge of higher-level structure.
- Voyager was seen as demonstrating high *Spontaneity* but at a level of conscious decisions rather than *Subconscious Processing*.
- *Thinking and Evaluation* was given relatively low ratings. There was 'something going on' (Judge 1) but Voyager was seen to lack higher-level strategies, reacting rather than thinking.
- *Value* judgements were mixed, depending on judges personal tastes.
- *Generation of Results* was praised but Voyager's ability to detect whether it had produced an end product was questioned.

- Generally Voyager was not seen as demonstrating any real *General Intellect*. Judge 1 remarked that the paper on Voyager (Lewis, 2000) was more sophisticated than the system itself.

6.4.4 Main conclusions from SPECS component evaluation

Overall, GenJam was found to be the most creative of the three systems. Voyager performs better than GenJam in a less important components (*Spontaneity and Subconscious Processing*); however GenJam's higher average ratings for nearly all components, and its particular strengths in components of importance such as *Social Interaction and Communication*, demonstrate its greater musical improvisational creativity. To improve on musical improvisational creativity, greatest gains can be made in all three systems by concentrating efforts first on more important components such as *Social Interaction and Communication*, *Intention and Emotional Involvement* and *Domain Competence* and by taking the judges' qualitative feedback into account. Work on other components will be of some benefit but this is limited, relative to these three components.

6.4.5 Judges' overall preferences

At the end of the study, judges were asked which of the two systems they had evaluated was, in their opinion, the more creative, and why. Table 6.8 reports this data. Results from collating judges' preferences are inconclusive; none of the systems was always seen as more creative than others.

If a system is given 1 point for being ranked first and 0 points for being ranked second (last) then GenJam and Voyager score 2 points each and GAmprovising scores 1 point.

Looking at the head-to-head rankings between pairs, we see an ordering emerge:

1. GenJam
2. Voyager
3. GAmprovising

This ranking is not dissimilar to the findings from analysis of the component rankings, being consistent with the general consensus of opinion for five judges, but is however at odds with Judge 4's opinion.[21] This discrepancy is due to this judge's higher ratings and opinions for the GAmprovising system, compared to the other two judges rating this system. The dissenting voice of Judge 4 against the general consensus of the other five judges illustrates the difficulty in identifying which systems are definitively more creative than others, and whether there can always be a 'right answer' in terms of relative creativity. Differing opinions on the concept of creativity are to be expected, given the multi-faceted and subjective nature of creativity and the lack of universal agreement on the meaning of this concept (Chapter 3 discusses this in detail).

[21] As is noted elsewhere (for example on occasions in Section 6.4, Judge 4's evaluation differed from other judges, giving GAmprovising ratings averaging 5 points higher than Judge 5 for example (a large margin for ratings given out of 10).

Table 6.8: Judges' feedback ordering the 2 systems they evaluated in terms of their overall creativity. System 1 is the system identified as more creative, System 2 is the system identified as less creative. *This case study originally included a fourth system, evaluated by Judges 2, 3 and 6. This system is represented in these comparisons as * . Though this fourth system has now been removed from Case Study 1 (see Section 6.2), it is interesting to see if the judges who evaluated this system found the other system they evaluated to be the more or the less creative of the two they evaluated.*

Judge	System 1	System 2
1	GenJam	Voyager
2	*	GAmprovising
3	Voyager	*
4	GAmprovising	GenJam
5	Voyager	GAmprovising
6	GenJam	*
Reasons		
1	Limitations [constraints] in GenJam allow it to be more thorough and develop itself	
2	System * (now removed) is task-specific, learns from previous solos	
3	Voyager is reactive, produced different responses, more musically interesting to hear	
4	GAmprovising has lots of freedom and still seemed to read off itself	
5	Voyager more sophisticated; interaction, progress, doesn't just produce notes at random	
6	GenJam's interactive element means it responds to other things around it	

The judges' reasoning for their opinions is perhaps more useful than the summative feedback gained by comparisons, as formative feedback for developing artificial musical creativity. Having appraised all 14 components for two systems each, judges prioritised the ability to be varied and free whilst still operating within constraints and using domain competence, react to what is happening around it, use sophisticated and considered methods, learn from experience and self-evaluate and produce musically interesting results. This covers the components *Variety Divergence and Experimentation*, *Domain Competence*, *Social Interaction and Communication*, *Spontaneity and Subconscious Processing*, *General Intellect*, *Thinking and Evaluation*, *Value* and *Generation of Results*.

Whilst there may be some priming effect from having looked at the components prior to offering an overall opinion, this shows the importance of having a varied set of components to meet multiple requirements for this type of creativity. The judges' feedback here strongly reflected components such as *Social Interaction and Communication* and *Domain Competence*, highlighted as important in the improvisation questionnaire in Section 6.3.2. *Intention and Emotional Involvement* was not mentioned by judges overall, but this could be due to the unfamiliarity of applying this human-creativity-centric component to computational systems, even by people with some familiarity of computational

creativity or computer music.[22] The feedback also showed it was necessary to retain the diversity of the set of components and consider less important components (according to the questionnaire results) such as *Thinking and Evaluation* and *Value* as well, rather than minimising the set of components used to those found to be most important.

6.5 Summary

Section 6.1 considered musical improvisation as a creative domain. In Case Study 1, the SPECS methodology was applied to evaluate the creativity of three systems that improvise musically: GAmprovising (Jordanous, 2010c), GenJam (Biles, 2007) and Voyager (Lewis, 2000) (Section 6.2). The purpose of this evaluation was to evaluate these three systems using detailed analysis and domain expertise to inform the evaluation.

Section 6.3 reported how SPECS was implemented by customising the set of 14 components identified in Chapter 4 such that components are weighted according to that component's overall contribution to a system's musical improvisation creativity. Expert judges were asked to research two systems each, spending 30 minutes on each system to research about its functionality and performance. The judges then provided quantitative feedback, rating out of 10 how the system performed on each of the components, complemented by qualitative feedback on each component. The quantitative ratings were later weighted according to relative component importance, as identified from investigations in Section 6.3.2. Analysis of the weighted ratings was supported by the qualitative data. Finally the judges were asked to rank the two systems they had evaluated in terms of their relative creativity, giving reasons for their rankings.

The application of SPECS generated much information from evaluating the three musical improvisation systems. Results of Case Study 1 were reported and discussed in Section 6.4. While GenJam emerged from the evaluation process as the most creative system, according to general consensus, the findings from SPECS that seem more useful are the feedback details and comments for each system, to inform future development of the systems. For example GenJam's creator, Biles, could take inspiration from how Voyager exhibits *Spontaneity and Subconscious Processing*. For each system, strengths and weaknesses relating to that system's creativity were identified:

- GAmprovising's identified strengths were its ability to generate results, develop those results and progress as a system, though it is poor at using reasoned self-evaluation and at being interactive.
- In contrast, GenJam's interactive abilities were praised alongside its ability to create results, whilst it could improve on its spontaneity and originality.

[22] This will be discussed further in Chapter 9 Section 9.1.

- Voyager was considered to be highly interactive and communicative, with fairly high associated value attached to the system and its products. Voyager's ability to develop what it does and progress over time was criticised though, as was its ability to think and self-evaluate.

Section 6.3.2 investigated which components make the greatest contribution to musical improvisation creativity. The SPECS results showed that for all three systems, to make them more creative, it is most profitable to concentrate on improving the systems' abilities at the three most important components: *Social Interaction and Communication*, *Domain Competence* and *Intention and Emotional Involvement*. Key aspects of creativity in musical improvisation were therefore identified: the ability to communicate and interact socially; the possession of relevant musical and improvisational skills and knowledge; and the emotional engagement and intention to be creative. Conversely, the actual musical results produced during improvisation were found to be relatively less important for creativity when compared to the process of improvising. Also, general intelligence was considered less important than having specific expertise and knowledge.

When building musical improvisation systems and intending to make them as creative as possible, such formative feedback contributes towards this goal. This is not to say that other components should be neglected; all components were shown to have some contribution to creativity (with all being mentioned in the questionnaire in Section 6.3.2 and several being highlighted in judges' qualitative feedback, presented in Section 6.4.3). The top three components, however, collectively accounted for 41.3% of musical improvisation creativity, according to the weights identified in Section 6.3.2, illustrated in Figure 6.7.

As Section 6.4.5 reported, the SPECS results overall reflected the majority of the judges' overall opinions when comparing systems' creativity. This is with the exception of one judge, Judge 4, whose feedback was shown to differ to that of other judges to some extent. That there is often not a single 'right answer' in creativity evaluation indicates the relative importance and need for formative evaluation compared to summative evaluation. Formative evaluation identifies constructive criticism that can be made about the system, as well as feedback which highlights the strengths of the system for the developers to be aware of and for other system developers to learn from. Both qualitative and quantitative feedback have proven useful for this purpose during the evaluation reported in this Chapter.

Chapter 7

Case Study 2: Snapshot evaluation of creative systems

Figure 7.1: *Wordle* word cloud of this Chapter's content

Overview

Case Study 2 explores to what extent SPECS can be applied across creative systems demonstrating different types of creativity, in contrast to Case Study 1's focus on systems operating specifically within the domain of musical improvisation creativity (Chapter 6). Case Study 2 also explores the scenarios where we do not have the full information desired for evaluation, or where we may have only limited time to complete evaluation, or be limited as to who can perform evaluation. Using short (seven minute) presentations on creative systems at the 2011 International Computational Creativity Conference (ICCC'11), two judges use SPECS to evaluate the creativity of five systems performing five different creative tasks: the collage generation module for the artistic system *The Painting Fool* (Colton, 2008b) by Cook and Colton (2011); a poetry generator (Rahman & Manurung, 2011), the *DARCI* system by Norton, Heath, and Ventura (2011) for generating images to illustrate given adjectives; a reconstruction of the *MINSTREL* story-telling system (Turner, 1994) (Tearse et al., 2011); and a musical soundtrack generator matching emotions in a narrative to the music generated (Monteith, Francisco, Martinez, Gervás, & Ventura, 2011). These systems are introduced in Section 7.2.

The intentions behind this case study are to apply SPECS to creative systems in different types of creative domain and to explore how the SPECS methodology can be used for a quickly-obtained brief evaluation, under pressures of time and limited information. Section 7.3 reports how the two judges rate each system along the Chapter 4 components given the information in the short presentations. Case Study 2 results in formative evaluative feedback for the systems to help researchers develop the creativity of their system. Section 7.4 reports and discusses the generated feedback, which is fairly detailed even given the time and information pressures. An unexpected but beneficial extra finding of the evaluation is that it highlights which ICCC'11 presentations had contained adequate information for judging the creativity of their systems. Hence feedback is also offered in Section 7.4 about what to include in future presentations to maximise information about how creative the systems are.

When the creative domain varies across systems, comparisons between systems prove less relevant as the systems are designed to perform different tasks, requiring different interpretations of creativity.[1] Instead the focus moves to evaluating individual systems, though some comparisons are made between systems where there are commonalities in domains or creative priorities of that domain.

7.1 The impact of digital resource availability for comparative research evaluation

It is more straightforward to evaluate systems for which we have full access to view and run the source code, with as much time available as we need, all necessary data and a line of communication with the system developers. This ideal evaluation scenario, however, is often not possible. Taking time

[1] As was discussed in Chapter 3 Section 3.6.4.

constraints first, the amount of time we can spend on evaluation is partly dictated by factors[2] such as the allocation of researchers' time (particularly when conducting multiple projects or when time allocations are dictated by funding), deadlines for conferences etc., time demands within a project and the scheduling of other tasks to be completed within the project. Further demands on researchers' time include teaching, administration and other research work. There are often also constraints on the time and availability at appropriate times of other people involved in performing the evaluation.[3] This last point is especially pertinent for evaluation study participants; the use of evaluators who are not involved in the system's development has previously been recommended in Chapters 2 and 5 (as it is likely to lead to more independent, unbiased evaluative feedback than if evaluation was performed solely by the researcher(s) responsible for the research), but Chapter 5 Section 5.1.4 highlighted several practical issues in using participants for evaluation studies and experiments. Another important issue impacting upon evaluation is if there are problems with availability of relevant software, data or more detailed information for a creative system(s) that we are interested in.

We could choose not to use systems for comparative evaluation if we do not have the full access and data that we would like; while this reduces the evaluation workload it sacrifices the opportunity to learn from this system. Alternatively, we can include systems in comparative evaluation even if we only have partial information for that system, keeping aware of the constraints on what we can learn from such evaluation but taking advantage of what is available, for formative feedback into the development of our own existing and future systems. As Bentkowska-Kafel reflects, however, on the lack of availability of computer artworks and their related research resources, 'lessons from the past are difficult to learn' (Bentkowska-Kafel, 2009, p. 149).

When would we wish to learn from other existing systems? Systems of historical interest would have intrinsic value, even if the system can no longer be obtained. For example, James Meehan's TALESPIN system (Meehan, 1976) has proven to be a seminal work in the field of story generation, even though the code has been lost and only a 'micro-TALESPIN' version exists today (Meehan, 1981), which was itself published over 30 years ago. Similarly, Christopher Longuet-Higgins produced software for expressive musical performance which was widely praised by those who heard it (Darwin, 2004, and personal communications with: Darwin, 2012; Dienes, 2012; Torrance, 2012; Thornton, 2012). Unfortunately, the system was not made available as code or in a published report before Longuet-Higgins' death, and the code was archived but now cannot be restored due to the use of obsolete data storage formats.[4]

[2] Chapter 9 Section 9.1.3 will reflect upon time constraints in the case studies in greater detail.

[3] For example, an evaluation may include involvement from participants in evaluation studies; any other researcher(s) involved in the evaluation or in any associated wider projects; staff providing administration, support services or access to technical equipment; assistants and/or students; project advisors and/or team leaders/principal investigators.

[4] According to personal communications with Jeremy Maris and other IT support staff at the University of Sussex, where

Interest in other systems should not, however, be limited to those systems which have proven significant over time. We can learn from what our peers are doing in closely related research areas, and also by cross-applying work from less related areas to our own interests. Creative systems are by their nature likely to be different to every other system and it is useful to see how a creative domain has been approached in different ways. As described in Chapter 5 Section 5.4, there may be systems that are related in some way to systems that we are developing, where it would be of interest to learn more about the research behind that system(s); in particular it would be useful to gain knowledge from seeing the system in operation, as well as reading published reports. As example, in evaluating the GAmprovising system against GenJam and Voyager, several useful insights arose for the development of GAmprovising from learning from how GenJam and Voyager operated (as reported in Chapter 6).

As Robey (2011) has remarked, research that produces computer programs is surrounded by issues of software sustainability. Unfortunately, even for more recent systems, it can be difficult to retrieve all information necessary for full evaluation of a system. Bentkowska-Kafel (2009) and Robey (2011) have highlighted the speed at which current or cutting-edge digital resources can quickly become obsolete or lost, sometimes in a matter of only a few years.

Digital resources such as source code may not have been made available publicly. If the researchers are still active in academia then the code may be obtainable, however they may have left academia (or in some cases, as for Christopher Longuet-Higgins, the researcher may have passed away). As example, the improvising systems in Hodgson (2006c) could not be obtained in time for Case Study 1, despite Hodgson's thesis being dated just prior to the start of this current doctoral project.

Code may be unavailable or obsolete even if obtained, due to having been written for what is now legacy equipment. (This was the case for GenJam, in Case Study 1.) If code has been made available publicly, its continued availability may not be guaranteed, for example if funding runs out for online hosting costs. Additionally, the Preserving Virtual Worlds project final report (quoted in Hong, Crouch, Hettrick, Parkinson, & Shreeve, 2010, pp. 34-35) identifies various reasons why available code may become unusable, including hardware or software obsolescence, third party dependencies, proprietary or poorly documented code as well as concerns about protecting intellectual property rights (especially in more competitive scenarios).

> 'digital information lasts forever - or five years, whichever comes first.' (Rothenburg, 1999, p. 2)

The impact of digital obsolescence and resource loss for researchers has long been recognised, in a scope which extends broadly, beyond computational creativity research to include any research which makes use of digital resources. In 1967 the UK Data Archive was established and supported

Longuet-Higgins' computer files were archived, and with a digital archive specialist, Gareth Knight.

by the Social Science Research Council as a large-scale digital archive and curation service for social science and humanities research data, with the provision of long term funding that is still ongoing today (now mainly ESRC/JISC-funded). Similar services have also been set up, such as the Arts and Humanities Data Service (Dunning, 2006) which received large-scale investment for 12 years from the AHRC and JISC.[5]

The UK Data Archive show evidence that "[m]any research funders and publishers are committed to a long-term strategy for data resource provision and encourage researchers to share data" and that "[r]esearch data are viewed by many funding bodies as a public good which should be openly available to the academic community and often beyond." Some journals, including the high impact series of *Nature* journals, require authors to make their data and materials available, as a condition of publication. Such models of sharing digital research contrast with the 'digital silo' approach to research (Nichols, 2009), where research data and outputs are not shared but are inaccessible to other researchers.

Van den Eynden, Corti, Woollard, Bishop, and Horton (2011) ask the question 'Why Share Research Data?'. Their various answers are reproduced here in full as each point has relevance as to why computational creativity researchers would benefit from investigating other researchers' data, and digital outputs such as creative systems,[6] in computational creativity research (as well as being relevant more broadly in research):

'Sharing research data:

- encourages scientific enquiry and debate
- promotes innovation and potential new data uses
- leads to new collaborations between data users and data creators
- maximises transparency and accountability
- enables scrutiny of research findings
- encourages the improvement and validation of research methods
- reduces the cost of duplicating data collection
- increases the impact and visibility of research
- promotes the research that created the data and its outcomes
- can provide a direct credit to the researcher as a research output in its own right
- provides important resources for education and training'

(Van den Eynden et al., 2011, p. 3)

[5]ESRC, AHRC and JISC are UK-based research funding bodies. ESRC: Economic and Social Research Council; AHRC: Arts and Humanities Research Council; JISC: Joint Information Systems Committee.

[6]Though the UK Data archive focuses on sharing digital data, their points are equally applicable to other digital outputs of research such as the software produced.

These points have been echoed elsewhere (for example Hong et al., 2010) and are perhaps best summarised in these quotes:

> 'without cultural artifacts, civilization has no memory and no mechanism to learn from its successes and failures.'[7]
>
> 'The emergence of modern technologies is opening new possibilities to the way scientists perform and disseminate their research. However, current publication infrastructures still focus on processing single monolithic resources within isolated data silos.' (Boulal, Iordanidis, & Quast, 2012, p. 162)

It could be argued that many legacy (and some contemporary) computational creativity systems would attract this criticism, should those systems not be available for some reason. If we can only access partial information about a system, can we still learn from it? Also, returning to the questions surrounding time availability and access to evaluators with appropriate expertise, if we only have limited time to perform evaluation, or have insufficient access to judges with appropriate expertise, would such evaluation still be useful? Initial impressions of the creativity of the programs may differ from their long term impressions, but may still be important if their first impressions are strong enough to make evaluative judgements on.[8] In response to the motivations outlined in this Section, Case Study 2 explores the feasibility and usefulness of performing evaluation in a scenario where only limited information and time is available, to see what can be learnt.

7.2 Creative systems at the 2011 International Computational Creativity Conference

The International Computational Creativity Conference (ICCC) is an annual international conference series dedicated to computational creativity research. Since its inception in 2010 it has been the main presentation venue for the latest findings in computational creativity research, taking over this role from the previous International Joint Workshops in Computational Creativity (IJWCC), from which the conference series evolved. ICCC'11 was held in Mexico in April 2011. Many creative systems were presented, demonstrating various types of creativity in different domains.

At ICCC'11, papers were presented to the conference audience in talks lasting seven minutes (a particularly brief amount of time for talks). There is a limit to what can be presented in this time and it is unlikely that all desired information can be provided, but the ICCC'11 organisers posited that enough information could be delivered for the audience to become acquainted with the paper content. During the seven-minute talks at ICCC'11, presenters aimed to convey enough information about their paper so that people could discuss issues raised, in a group of related talks.

[7] Original source unattributed, quote taken from *http://archive.org/about* (last accessed November 2012).

[8] From personal communications (2012) with Gareth White from *Vertical Slice*, a company specialising in video game usability testing.

To test how standardisable SPECS is across different types of system, Case Study 2 evaluates five different creative systems. This second case study also replicates the formation of initial 'snapshot' judgements of how creative something is based on our first impression, using incomplete information and without necessarily having expertise in that type of creative activity. Five of the creative systems presented at ICCC'11 were selected for Case Study 2, representing a variety of creative domains. These five systems were evaluated on their creativity, based on the information presented in the seven-minute talks. The evaluation also generated qualitative feedback for the presenters of the evaluated systems, from two perspectives: the perceived creativeness of their system and the quality of information that they presented about their system in the brief time frame permitted. A further purpose of this second case study was to apply the SPECS methodology more quickly than the previous case study. While the level of detail in Case Study 1 (Chapter 6) reinforces the quality and quantity of information gathered from the evaluation process, sometimes it is desirable to make a quick evaluation of a system at a particular time. This may be for instant feedback during development, or there simply may not be the time or expertise available to make a more detailed and informed evaluation. This case study illustrates how the methodological proposals in this thesis have been thus adapted.

One more point to note is that the systems evaluated in this case study were from a variety of domains, rather than just one domain. Some comparisons can be made between systems from different domains, and some interesting insight can be gained from doing so.[9] On the whole, though, such comparisons are less useful than comparisons made between systems from similar domains, as there are fewer areas of crossover so therefore less relevant information is available from the comparison. Some non-obvious conclusions may still however be reached this way, through viewing the systems from different perspectives. Comparing systems across different creative domains can also give a general (if limited) impression of relative progress in each domain.

The systems to be evaluated were taken from presentations in the first day of the conference. From the three main paper sessions that day, two judges evaluated systems presented in the first and third session, using the second session as an opportunity to take a break from the evaluation work. Evaluations were performed on one day only, partly to allow the judges to focus on taking part in the conference itself and partly because the process and previous ratings given were kept fresh in the judges' minds through evaluation as the evaluation process was completed in one day.

Of the five presentations in the first session (entitled 'The Applied'), the judges decided that three presented details of a computational creativity system that could be evaluated. Two further systems from the third session ('The Narrative') were also evaluated, for a total of five systems evaluated in this case study.[10] The talks at ICCC'11 provided information on the five systems highlighted for

[9] As shall be seen in Section 7.4.

[10] Other computational systems were also presented in this third session, however only two were selected for evaluation,

evaluation. These five systems, along with the papers they were presented in, the authors and the creative domain which the system operates in, are listed in Table 7.1.

Table 7.1: The five systems from ICCC'11 that were evaluated for Case Study 2.

Paper	System (if named)	Domain	Purpose
Cook and Colton (2011)	Module of *The Painting Fool*	Art	Collage generation
Rahman and Manurung (2011)	Adapted from an earlier system: McGonagall (Manurung, 2003)	Poetry	Poetry generation
Norton et al. (2011)	*DARCI*	Art	Image selection
Tearse et al. (2011)	Reconstruction of *MINSTREL*	Narrative	Story generation
Monteith et al. (2011)	-	Music	Soundtrack generation

7.3 Applying the SPECS methodology in Case Study 2

7.3.1 Step 1a: Domain-independent aspects of creativity

As for Case Study 1 (Chapter 6), the 14 components of creativity identified in Chapter 4 were used as a definition of creativity in a general context.

7.3.2 Step 1b: Domain-specific aspects of creativity in the ICCC'11 systems

Chapter 3 Section 3.6.4 discussed how creativity varies across domains. This case study evaluated systems that operate in the creative domains of art, poetry, narrative or music. It cannot be assumed that each of the 14 components are equally important across these domains. It is likely that some components vary in importance across domains, compared to other components.[11]

Case Study 2 investigates the forming of first impressions about creativity rather than detailed and highly informed analyses of creativity. Given this, and also to give contextual information for the judges' evaluations, the judges were asked to provide information during evaluation about how they perceived the relative importance of each component in the systems' domains.[12]

7.3.3 Step 2: Standards for evaluating the creativity of the ICCC'11 systems

Each system was evaluated according to a combination of two sets of information: how well the system performs on each of the 14 components and also how well it performed on the specific subset

to ensure that the evaluated systems covered a spread of different creative domains. The third session focused on narrative generation creativity. Including further systems from this session could have skewed the distribution across creative domains.

[11] Case Study 1 (Chapter 6) investigated this variance specifically with regards to creativity in musical improvisation.

[12] See Section 7.3.4. Even if judges were not experts in a given domain, it is still useful to see how they perceive creativity in that domain, partly as insight on a 'layperson's' view of that type of creativity, and partly as such perceptions are influential in guiding their evaluation.

of components highlighted by the judges as important contributors to creativity in that domain.

7.3.4 Step 3: Evaluative tests for creativity in the ICCC'11 systems

Evaluation was performed by two judges, each with varying competencies in the different domains. The judges, myself and another conference attendee, evaluated each system according to the information presented about the system in the conference talk. Judges were familiarised with the components and their meanings prior to evaluation[13] and copies of the evaluation form were given to judges in advance to allow the judges to familiarise themselves with the form.

Each system was evaluated during the 10 minute period comprising the seven-minute talk and the three-minute changeover period between talks.[14] Systems were evaluated independently by each judge. Each judge filled in one evaluation form (see Appendix E for the form) for each system. This evaluation form listed the 14 components of creativity derived in Chapter 4, with the component explanations from Chapter 4 (Section 4.3) as a reminder of what each component represented.[15]

The judges recorded what general creative domain each system was designed to operate in (e.g. art, narrative generation). They also assessed their level of expertise and competence in that domain as either *Basic*, *Reasonable* or *Expert*. For each component, judges rated how successfully the system performed on that component requirements, giving a score between 0 (lowest) and 10 (highest). The judge rated the system based on the information given in the conference talk; if they felt that not enough information was given about a particular component to provide a rating, then this rating was left blank. Each component was categorised according to how important the judge felt that component was for creativity in the domain which that system operated in. The contribution of that component to creativity in the system's domain was categorised as one of:

- Crucial for creativity.
- Quite important for creativity.
- A little important for creativity.
- Not at all important for creativity.[16]

Originally it was intended that judges should evaluate the contribution of each component at the same time as rating the system's performance on that component. It was soon found, however, that

[13] One of the judges for this Case Study was myself, as shall be explained in Section 7.3.5. Having derived the components and the evaluation method, I had more knowledge for this task than the second judge, but gave the second judge training on the components and their meanings beforehand. Using myself as one of the judges proved useful in gathering information and personal experience on what it was like to use SPECS for evaluation.

[14] At ICCC'11 there were no question periods after each talk in a session; instead all the talks were given back-to-back, then all presenters in a session would come back up to the front of the room and act as panellists for a discussion session.

[15] These explanations were written in a lighter font that could be written over if the judges wanted to make notes there.

[16] This option was not selected by either judge for any of the systems. In other words, the two judges thought that each component was at least a little important for creativity in each of the domains covered in Case Study 2.

judges did not have enough time to perform both tasks during the 10 minute period of evaluation for each system. Instead they prioritised rating of performance on each component at the expense of rating each component's importance for creativity in that domain. Additionally, judges were influenced by the system presenters as to their perceived importance of each component in their creative domain.

Instead, after the main evaluation had been done, a second stage of evaluation took place to establish how important each component was in different creative domains. For each main creative domain covered by the evaluated systems, the judges independently categorised the contribution of each of the 14 components to creativity in that domain, using the scale described above. Obtaining this data at this post-evaluation stage had the benefit that for the domain covered by more than one evaluated system (art), the component contribution judgements could be duplicated for each of the two systems, keeping this data consistent across different systems in the same domain.

After the conference the resulting evaluations were collated and analysed as individual evaluations and in combination. As well as identifying areas where both judges felt the system performed well or badly, discrepancies greater than 2 out of 10 between the two judges' ratings were highlighted for further inspection, as were ratings which were at an extreme (≥ 8 or ≤ 3) from one judge and left unrated by the other judge. Similarly, as well as identifying which components were deemed most important in a particular creative domain, differences in opinion were also noted. When considering both types of differences of opinion, the judges' self-rated expertise was taken into account. Judges may have interpreted the information imparted by the speakers's talks in differing ways according to:

- The background knowledge and opinions of the judge.
- Prior knowledge of a particular system.
- Language difficulties (as one of the judges was not a native speaker of English).
- Poor/inefficient reporting by the presenter.
- Mis-hearing or missing information because of listener fatigue or less interest in the talk.

7.3.5 Choosing the judges for Case Study 2

Two judges were used for this evaluation: myself and another computational creativity researcher. More than one judge was used so that differences in individual opinions could be identified and so the opinions of each judge could be moderated by the opinions of the other.

Ideally, more than two judges would have been preferable, to capture more of a general consensus. Recruiting more than two judges was complicated, however, by the limited availability of judges in the experimental situation for this case study. Conference attendees were interested in hearing the session talks for their own benefit rather than for performing evaluation tasks, or were involved in presenting their own talks or chairing sessions. For the two people that acted as judges for this case study, restrictions were placed on the number of systems evaluated, so as to allow the two judges to

participate more fully in the majority of the conference once their evaluation duties were complete.[17]

7.4 Results and Discussion

Unless stated otherwise, the judges had no knowledge of the five systems being evaluated, prior to the ICCC'11 conference. In Tables 7.2 through to 7.7, the following conventions are adopted:

- Data is reported in the format: <Judge 1 data> / <Judge 2 data>.
- Notable discrepancies between judges' ratings are highlighted using bold type.
- The character '-' for a judge's rating indicates there was insufficient information given for the judge to be able to rate that particular component.

7.4.1 Creativity evaluation of Cook & Colton's collage generator

Evaluation ratings for the collage generation module of *The Painting Fool* in Cook and Colton (2011) are reported in Table 7.2.

General observations

- Both judges rated the collage generator very highly for its ability for *Generation of Results* (10 and 9), considered crucial for creativity by Judge 1 and quite important by Judge 2.
- The collage generator received maximum ratings for how it demonstrated *Intention and Emotional Involvement* (10 and 10). This component was emphasised as a key attribute for creativity during the ICCC'11 presentation but was rated slightly lower in importance by the two judges (Judge 1 saw this as only a little important, whilst Judge 2 thought it was quite important).
- The system performed very well for *Social Interaction and Communication* (9 and 7), although the judges were divided as to the importance of this component for artistic creativity.
- The system received either high or mid-range ratings, with no ratings below 5 out of 10.

Domain-specific observations (artistic creativity)

Judge 1 reported expert knowledge of the domain of art, whilst Judge 2 reported basic knowledge.

The components identified as being **crucial** for artistic creativity by at least one judge were:[18]

[17] One option which was considered, to enable a greater number of judges to be used, was to video the talks (getting prior permission from the speakers) and then use the videos for evaluation sessions at a later date. This option was not taken up, however, for two reasons. Firstly, the location of the conference, Mexico City, was not somewhere where I felt comfortable carrying around video equipment, for personal safety reasons. Secondly, the atmosphere at conferences is distinct from an atmosphere recreated in experimental conditions outside of the conference, not least because of the difference between watching a live presentation physically in front of you as part of a conference audience and watching a taped recording of that presentation in an experimental set up, away from the conference, with the speakers not present. The conference environment, with its focus on computational creativity research, was felt to be the most appropriate for carrying out evaluations for this case study.

[18] Excluded from this list is *Social Interaction and Communication*; although Judge 2 identified this as crucial for artistic creativity, Judge 1 deemed it only a little important in this domain and was more familiar with the domain.

Table 7.2: Judges' data for Cook & Colton's collage generator. System domain: Art. Domain knowledge: Judge 1 = expert, Judge 2 = basic.

Component	Rating/10	Contribution for creativity
Active Involvement and Persistence	- / 6	Quite important / Quite important
Generation of Results	10 / 9	Crucial / Quite important
Dealing with Uncertainty	- / -	A little important / Quite important
Domain Competence	- / -	Quite important / Quite important
General Intellect	**- / 8**	A little important / A little important
Independence and Freedom	**- / 8**	Quite important / Quite important
Intention and Emotional Involvement	10 / 10	A little important / Quite important
Originality	- / -	Crucial / Quite important
Progression and Development	- / 7	Quite important / Quite important
Social Interaction and Communication	9 / 7	**A little important / Crucial**
Spontaneity and Subconscious Processing	- / 5	Crucial / Quite important
Thinking and Evaluation	- / -	A little important / Quite important
Value	- / -	Crucial / Quite important
Variety, Divergence and Experimentation	5 / -	A little important / Quite important

- *Generation of Results* (**Crucial /** Quite important).
- *Originality* (**Crucial /** Quite important).
- *Spontaneity and Subconscious Processing* (**Crucial /** Quite important).
- *Value* (**Crucial /** Quite important).

This list demonstrates that in artistic creativity, there is an emphasis on *Originality* and *Value*, *Generation of Results* and being spontaneous rather than over-analytical (*Spontaneity and Subconscious Processing*). Hence these aspects should receive greater attention if the objective of the work is to demonstrate artistic creativity. The other components should not be neglected as they were all considered important to some degree for this type of creativity, though the *General Intellect* component was rated as only a little important by both judges so can be de-prioritised. For three of the four components highlighted as most significant, there was insufficient information for both judges to evaluate the collage generator in the 7-minute presentation.

Conclusions for Cook & Colton's collage generator

This system performed well in some ways, particularly in producing results, demonstrating an intention to be creative and a social ability. There was less data from which to draw general conclusions about this system's creativity, however, particularly for aspects which are important for systems working in artistic creativity. At present only a shallow but generally positive impression can be drawn of the system's creativity. For a more informed evaluation, it would be useful to concentrate further

investigations on how original and valuable the system or the system's results can be, with reflection on its ability to be spontaneous.

7.4.2 Creativity evaluation of Rahman & Manurung's poetry generator

Evaluation ratings for the system in Rahman and Manurung (2011) are reported in Table 7.3.

Table 7.3: Judges' data for Rahman & Manurung's poetry generator. System domain: Poetry. Domain knowledge: Judge 1 = basic, Judge 2 = basic.

Component	Rating/10	Contribution for creativity
Active Involvement and Persistence	- / -	Quite important / Quite important
Generation of Results	10 / 9	Crucial / Crucial
Dealing with Uncertainty	5 / 4	A little important / Quite important
Domain Competence	9 / 8	Quite important / Crucial
General Intellect	- / -	A little important / Quite important
Independence and Freedom	- / -	Quite important / Quite important
Intention and Emotional Involvement	**- / 2**	A little important / Quite important
Originality	- / -	Crucial / Quite important
Progression and Development	9 / 7	Quite important / Quite important
Social Interaction and Communication	**8 / 2**	Quite important / Quite important
Spontaneity and Subconscious Processing	- / -	Crucial / Quite important
Thinking and Evaluation	9 / 7	Quite important / Quite important
Value	- / -	Crucial / Crucial
Variety, Divergence and Experimentation	5 / -	Crucial / Quite important

General observations

- The poetry generator was rated highly (10 and 9) for its capability for *Generation of Results*, with both judges considering this crucial for creativity in poem-writing.
- *Domain Competence* was seen as important and also received high ratings (9 and 8).
- Judges also gave fairly high ratings to the system for how well it exhibited *Progression and Development* (9 and 7) and *Thinking and Evaluation* (9 and 7), which were both considered quite important components of creativity.
- Quite a few components were left unrated by both judges (6 components) or by one of the two judges (2 components).
- Where ratings were given by both judges, Judge 1 gave higher ratings than Judge 2, but on the whole the judges' ratings were very similar here. A notable exception is in *Social Interaction and Communication*, deemed quite important by both judges, but which was given a rating of 8 by Judge 1 and 2 by Judge 2. Upon investigating this further, Judge 1 reported that they gave

a high rating as they considered the poetry to make a good social contribution, but Judge 2 felt that the system made little use of feedback and two-way communication so gave a low rating.[19]

- The poetry generator scored very poorly for its demonstration of *Intention and Emotional Involvement* by Judge 2 (2). Judge 1 did not feel able to rate the system on this component.

Domain-specific observations (poetic creativity)

Both judges felt that they only had basic competence in the creative domain of poetry generation.

The components identified as being **crucial** for poetic creativity by at least one judge were:

- *Generation of Results* (**Crucial / Crucial**).
- *Value* (**Crucial / Crucial**).
- *Domain competence* (Quite important **/ Crucial**).
- *Originality* (**Crucial /** Quite important).
- *Spontaneity and Subconscious Processing* (**Crucial /** Quite important).
- *Variety, Divergence and Experimentation* (**Crucial /** Quite important).

Of these components, the poetry generator in Rahman and Manurung (2011) scored highly for *Generation of Results* and *Domain Competence* but received an average rating or no rating for *Variety, Divergence and Experimentation*. From the information in the seven-minute talk, neither judge felt able to judge the system on its *Value*, *Originality* or the ability to demonstrate *Spontaneity and Subconscious Processing*. Therefore, although the poetry generator was considered to perform well in some important aspects of poetic creativity, information was missing from the talk presentation, hindering the judges in forming opinions on other important aspects.

Conclusions for Rahman & Manurung's poetry generator

Overall the system received mostly high or medium-high ratings, with little negative comment. Six components were not rated by either judge and a further two components were only rated by one judge. While acknowledging the limit to the amount of information that can be imparted in seven minutes, there was a lack of information about some components which make an important contribution to creativity in this domain. To better simulate creativity in poetry generation, Rahman & Manurung could work more on developing the system's ability to experiment and diverge, and for more general creativity, on the ability of the system to demonstrate the intent and emotional involvement with being creative. When presenting the system, more information on the originality and value in the system would be useful in making judgements on the creativity of the system, as well as comments on how spontaneous the system is and any ability it has to conduct parts of the creative process

[19]This discrepancy reflects partly on the breadth of coverage of this component. As was discussed in Chapter 4, the components are constructed from clusters of several related aspects rather than representing individual aspects, to reduce dimensionality in the results. Chapter 9 Section 9.1 will reflect upon this point further.

at some subconscious level.

7.4.3 Creativity evaluation of Norton et al.'s image generator *DARCI*

Evaluation ratings for the system in Norton et al. (2011) are reported in Table 7.4.

Table 7.4: Judges' data for Norton et al.'s image generator *DARCI*. System domain: Art. Domain knowledge: Judge 1 = expert, Judge 2 = basic.

Component	Rating/10	Contribution for creativity
Active Involvement and Persistence	- / -	Quite important / Quite important
Generation of Results	9 / 9	Crucial / Quite important
Dealing with Uncertainty	**- / 8**	A little important / Quite important
Domain Competence	- / 7	Quite important / Quite important
General Intellect	- / 6	A little important / A little important
Independence and Freedom	- / 4	Quite important / Quite important
Intention and Emotional Involvement	**- / 2**	A little important / Quite important
Originality	- / -	Crucial / Quite important
Progression and Development	- / 7	Quite important / Quite important
Social Interaction and Communication	10 / 8	**A little important / Crucial**
Spontaneity and Subconscious Processing	5 / 7	Crucial / Quite important
Thinking and Evaluation	5 / 7	A little important / Quite important
Value	- / 7	Crucial / Quite important
Variety, Divergence and Experimentation	**10 / 5**	A little important / Quite important

General observations

- *DARCI* was rated highly by both judges for its ability for *Generation of Results* (9 and 9) and for *Social interaction and Communication* (10 and 8).
- Middling to high ratings were given by both judges for *DARCI*'s *Spontaneity and Subconscious Processing* (5 and 7) and its ability for *Thinking and evaluation* (5 and 7).
- In general *DARCI*'s ratings were between 5 and 10. Judge 2 considered *DARCI* to be poor at demonstrating *Intention and Emotional Involvement* (2) and *Independence and Freedom* (4), though Judge 1 did not provide ratings for these components.
- Judges opinions were less consistent for this system than for other systems. Judge 1 scored *DARCI* 5 points higher for *Variety, Divergence and Experimentation* than Judge 2 (10 and 5). Additionally, Judge 2 provided more ratings overall than Judge 1 did.
  - Judge 2 had previous acquaintance with the *DARCI* system so probably incorporated that background knowledge in their evaluations, given that Judge 2 provided many more ratings for this system than Judge 1.

– Judge 1 has a considerably higher level of expertise in this domain than Judge 2, having expert knowledge rather than just basic knowledge.

Domain-specific observations (artistic creativity)

Judge 1 reported expert knowledge of the domain of art, whilst Judge 2 reported basic knowledge.

The components identified as being **crucial** for artistic creativity by at least one judge were:

- *Generation of Results* (**Crucial /** Quite important).
- *Originality* (**Crucial /** Quite important).
- *Spontaneity and Subconscious Processing* (**Crucial /** Quite important).
- *Value* (**Crucial /** Quite important).

DARCI was considered to be good at *Generation of Results* (9 and 9) and was given middle-to-high ratings for its *Spontaneity and Subconscious Processing* (5 and 7). Judge 2 rated the *Value* in *DARCI* as fairly high (7) although Judge 1 did not rate this. Neither judge felt able to comment on the *Originality* of *DARCI*. As for *The Painting Fool*'s collage generation module in Cook and Colton (2011), *General Intellect* was considered by both judges to make only a little contribution to artistic creativity, so this aspect of the system can be given less attention than others.

Again, judges' opinions were fairly inconsistent about the relative importance of several components, only agreeing on the level of importance of contributions for 5 of the 14 components. For *Social Interaction and Communication*, as for Cook & Colton's artistic collage generator, Judge 1 considered this component to be only a little important for artistic creativity as opposed to Judge 2's opinion that it is crucial for creativity.

Conclusions for DARCI

DARCI was considered by the judges to be good at creating results and interacting and communicating socially. It also showed an appreciable level of spontaneity and subconscious processing, as well as the ability to reason and evaluate. From the perspective of artistic creativity, *DARCI* performed well in some key aspects but it would be useful to have more information on *DARCI*'s ability to demonstrate originality and value, to evaluate *DARCI* better as a creative artist. On a more general note, this evaluation of *DARCI* also showed that differing levels of expertise and prior knowledge of a system between judges can have noticeable effects on evaluation results.

7.4.4 Creativity evaluation of Tearse et al.'s narrative generator

Evaluation ratings for the system in Tearse et al. (2011) are reported in Table 7.5.

Table 7.5: Judges' data for Tearse et al.'s reconstruction of the story generator *MINSTREL*. System domain: Narrative. Domain knowledge: Judge 1 = basic, Judge 2 = reasonable.

Component	Rating/10	Contribution for creativity
Active Involvement and Persistence	- / 7	Quite important / Quite important
Generation of Results	10 / 9	Crucial / Crucial
Dealing with Uncertainty	- / -	Quite important / Quite important
Domain Competence	**8 / -**	Quite important / Quite important
General Intellect	- / 6	Quite important / Quite important
Independence and Freedom	- / 4	Crucial / Quite important
Intention and Emotional Involvement	- / -	A little important / Quite important
Originality	8 / 7	Crucial / Quite important
Progression and Development	5 / 5	Quite important / Crucial
Social Interaction and Communication	- / -	Quite important / Crucial
Spontaneity and Subconscious Processing	- / -	Crucial / Quite important
Thinking and Evaluation	- / 7	A little important / Quite important
Value	4 / 5	Crucial / Quite important
Variety, Divergence and Experimentation	5 / 6	Quite important / Quite important

General observations

- The *MINSTREL* reconstruction was considered to be good at *Generation of Results* (10 and 9) and demonstrating *Originality* (8 and 7).
- This system was considered to be only average for *Progression and Development* (5 and 5) and *Variety, Divergence and Experimentation* (5 and 6), reflecting a lack of any noticeable ability to develop and experiment (but also no particular weakness in these areas).
- The *Value* associated with the *MINSTREL* reconstruction in Tearse et al. was given medium to low ratings (5 and 4).
- Judge 1 felt the system showed good *Domain Competence* within the domain of narrative generation (8), though Judge 2 did not feel they had enough information to comment on this.
- Judge 2 felt that the system was reasonably good for *Active Involvement and Persistence* (7), employing *Thinking and Evaluation* (7) and demonstrating *General Intellect* (6). This judge gave a slightly low rating to the system's ability to work with *Independence and Freedom* (4), though. Judge 1 could not comment on any of these components. This may be partly due to Judge 2's greater domain knowledge (reasonable as opposed to Judge 1's basic knowledge).
- In general the *MINSTREL* reconstruction received more ratings in the middle of the scale as opposed to high ratings. It received no ratings lower than 4 out of 10.
- From the information in the ICCC'11 talk, judges were able to provide at least one rating between them for all but four of the components.

Domain-specific observations (narrative creativity)

Judge 2 considered themselves to have reasonable knowledge of creativity in narrative generation. Judge 1's domain knowledge was at a basic level.

The components identified as being **crucial** for narrative creativity by at least one judge were:

- *Generation of Results* (**Crucial / Crucial**).
- *Independence and Freedom* (**Crucial /** Quite important).
- *Originality* (**Crucial /** Quite important).
- *Progression and Development* (Quite important **/ Crucial**).
- *Social Interaction and Communication* (Quite important **/ Crucial**).
- *Spontaneity and Subconscious Processing* (**Crucial /** Quite important).
- *Value* (**Crucial /** Quite important).

In terms of components considered important for narrative creativity, The *MINSTREL* reconstruction performed well for *Generation of Results*, and almost as well for *Originality* according to both judges. It was however thought to demonstrate only an average capacity for *Progression and Development* and its *Value*. Judge 2 gave the system quite a low rating for its *Independence and Freedom* (4), whilst Judge 1 did not have enough information about the system's autonomy. Both judges felt unable to evaluate Tearse et al.'s system for its abilities in *Social Interaction and Communication* and *Spontaneity and Subconscious Processing*.

Conclusions for Tearse et al.'s reconstruction of MINSTREL

The MINSTREL reconstruction was generally considered average to good for most of the four components of creativity. In evaluation its strengths were identified as the ability to create results and be original, which are both important contributors to creativity in narrative generation as well as in general. In some areas, the *MINSTREL* reconstruction could perform better, particularly in components that contribute highly to narrative creativity such as the value in the system, its ability to progress and develop and to be independent and work freely. To consider this system's creativity in more depth, particularly in the domain of narrative creativity, more emphasis should be placed on discussing the system's independence, and its social ability and spontaneity should both be considered.

7.4.5 Creativity evaluation of Monteith et al.'s musical soundtrack generator

Evaluation ratings for the system in Monteith et al. (2011) are reported in Table 7.6.

General observations

- Six components were left unevaluated by both judges. For a further three components, only Judge 2 supplied evaluation data.

Table 7.6: Judges' data for Monteith et al.'s soundtrack generator. System domain: Music. Domain knowledge: Judge 1 = basic, Judge 2 = expert.

Component	Rating/10	Contribution for creativity
Active Involvement and Persistence	- / -	Crucial / Quite important
Generation of Results	**- / 9**	Crucial / Quite important
Dealing with Uncertainty	- / -	Quite important / Quite important
Domain Competence	5 / 7	Crucial / Quite important
General Intellect	- / -	A little important / A little important
Independence and Freedom	**7 / 4**	Quite important / Quite important
Intention and Emotional Involvement	**10 / 2**	A little important / Quite important
Originality	- / 7	Quite important / Quite important
Progression and Development	- / -	Crucial / Quite important
Social Interaction and Communication	5 / 5	**A little important / Crucial**
Spontaneity and Subconscious Processing	- / -	Crucial / Quite important
Thinking and Evaluation	- / -	A little important / Quite important
Value	8 / 7	Quite important / Quite important
Variety, Divergence and Experimentation	- / 5	Crucial / Quite important

- There was some inconsistency between judges when evaluating this system. Whilst Judge 2 felt the system in Monteith et al. was very good at *Generation of Results* (9) and reasonable for *Originality* (7), Judge 1 did not receive enough information to judge the system on these components. There was an 8-point difference in opinion between judges for *Intention and Emotional Involvement* (10 and 2) and a 3-point difference in the judges for *Independence and Freedom* (7 and 4). In both these cases Judge 1 was more positive than Judge 2.
- In terms of the system's ability to show *Intention and Emotional Involvement*, for which there was the largest discrepancy recorded in this case study, Judge 1 considered how the system matched music to target emotions, whereas Judge 2 interpreted this component as the system's demonstration of its own emotions and intention.
- In general, the soundtrack generator received ratings of 5 or higher for its performance on the different components. The only ratings lower than 5 were from Judge 2, for *Independence and Freedom* and *Intention and Emotional Involvement* (2). As discussed above, these lower ratings were not supported by Judge 1.

Domain-specific observations (musical creativity)

Judge 2 had expert knowledge in this domain whereas Judge 1 had basic knowledge.

The components identified as being **crucial** for musical creativity by at least one judge were:

- *Active Involvement and Persistence* (**Crucial /** Quite important).

- *Generation of Results* (**Crucial /** Quite important).
- *Domain Competence* (**Crucial /** Quite important).
- *Progression and Development* (**Crucial /** Quite important).
- *Spontaneity and Subconscious Processing* (**Crucial /** Quite important).
- *Variety, Divergence and Experimentation* (**Crucial /** Quite important).

Although Judge 1 considered all the above to be crucial for creativity, the information in the ICCC'11 talk was only sufficient for Judge 1 to rate one of these six components, *Domain Competence*, for which they gave a mid-range rating (5). Judge 2 rated the system's *Domain Competence* slightly higher (7) and rated the system highly on *Generation of Results* (9) though they rated the system as only mid-range for *Variety, Divergence and Experimentation*. From the contents of the ICCC'11 talk on this system, neither judge felt equipped to evaluate the system on its *Active Involvement and Persistence*, *Progression and Development* or *Spontaneity and Subconscious Processing*.[20]

A similarity between musical creativity and artistic creativity is that judges found *General Intellect* to be less important than other aspects of creativity.

There was disagreement over the contribution of the *Social Interaction and Communication* component to musical creativity. Judge 2 considered this as the only crucial component for this type of creativity, rating all other components as quite important but not crucial. Conversely, Judge 1 felt this component was only a little important for musical creativity, but had only basic domain knowledge compared to Judge 2's expert domain knowledge. As the system received average ratings from both judges (5 and 5), this component is unlikely to make a large contribution (positive or negative) to the perceived creativity of this system.

Conclusions for Monteith et al.'s soundtrack generator

The strengths of the soundtrack generator in Monteith et al. are in the value it demonstrates, which was considered 'quite important' for musical creativity. It received mid to high ratings for its domain-specific competency, but this aspect could be improved on for greater demonstration of musical creativity. To appear more musically creative, Monteith et al. could also improve the system's capacity for experimentation and divergence. In considering the system as a musically creative entity, a little more information is required on the results created by the system. It would also be helpful to see discussion of its ability to be actively and persistently involved in the creative process, to progress and develop, and to be spontaneous and to simulate subconscious processing. The two judges often disagreed in their evaluation of this system, as was the case in previous systems in this case study where there was a sizeable difference in judges' domain expertise (artistic systems).

[20] It is noted that the presenter of this talk, Pablo Gervás, acknowledged during his presentation that of the five authors of this paper, he felt he was the least knowledgeable about the system's exact details.

7.4.6 Comparisons across different systems

The judges' evaluation ratings for all five systems are summarised in Table 7.7 and Figure 7.2.

Table 7.7: Judges' ratings for all five systems evaluated in Case Study 2. Discrepancies are in bold type. '-' indicates a lack of information to rate this component.

Component	Cook & Colton	Rahman & Manurung	Norton et al.	Tearse et al.	Monteith et al.
Active Involvement and Persistence	- / 6	- / -	- / -	- / 7	- / -
Generation of Results	10 / 9	10 / 9	9 / 9	10 / 9	- / 9
Dealing with Uncertainty	- / -	5 / 4	**- / 8**	- / -	- / -
Domain Competence	- / -	9 / 8	- / 7	**8 / -**	5 / 7
General Intellect	**- / 8**	- / -	- / 6	- / 6	- / -
Independence and Freedom	**- / 8**	- / -	- / 4	- / 4	7 / 4
Intention and Emotional Involvement	10 / 10	**- / 2**	**- / 2**	- / -	**10 / 2**
Originality	- / -	- / -	- / -	8 / 7	- / 7
Progression and Development	- / 7	9 / 7	- / 7	5 / 5	- / -
Social Interaction and Communication	9 / 7	**8 / 2**	10 / 8	- / -	5 / 5
Spontaneity and Subconscious Processing	- / 5	- / -	5 / 7	- / -	- / -
Thinking and Evaluation	- / -	9 / 7	5 / 7	- / 7	- / -
Value	- / -	- / -	- / 7	4 / 5	8 / 7
Variety, Divergence and Experimentation	5 / -	5 / -	**10 / 5**	5 / 6	- / 5

Individual components:

Active Involvement and Persistence This component was considered to be quite important for all the different types of creativity in this case study, by both judges, and one judge considered it crucial for musical creativity. In evaluation, however, Judge 1 did not feel able to rate a single system for this component, whilst Judge 2 only rated two systems (Cook & Colton, 2011; Tearse et al., 2011). From this it can be concluded that presenters at ICCC'11 did not usually feel the need to give details about their system's ability to affect the creative process and continually work to be creative even when there were problems. This has however been judged to be an important component of creativity for all systems evaluated, so should receive more attention.

Generation of Results The evaluated systems all performed extremely well at producing results, with ratings of 9 or 10 out of 10 for all systems (except for Monteith et al. (2011) where Judge 1 needed more information in order to rate the system for this component). This component was seen as crucial for creativity by at least one judge in all forms of creativity evaluated in Case Study 2. This is therefore a vital component of creativity in this case study and one that all the case study systems were considered to perform well in.

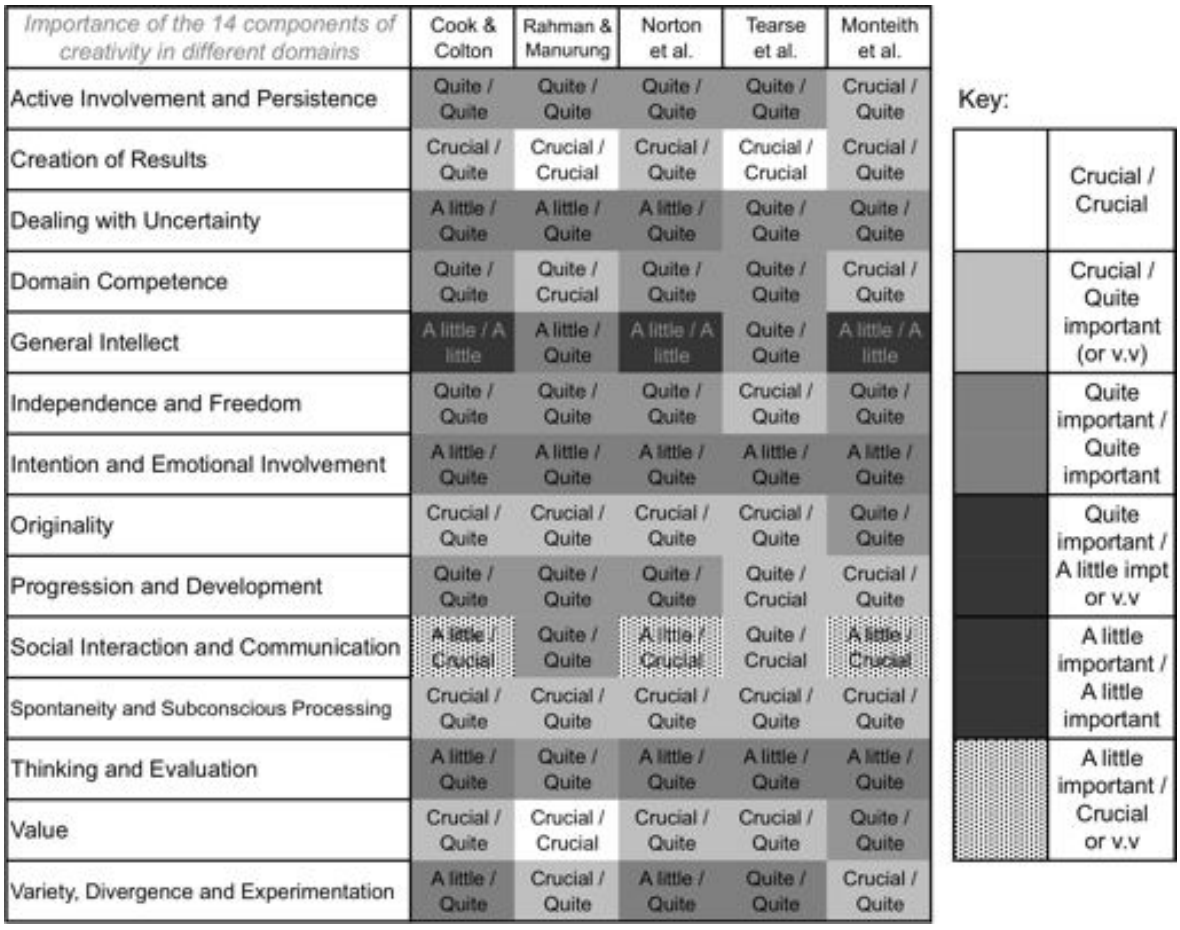


Figure 7.2: Graphical display of the relative importance of the 14 components for creativity in the creative domains of each of the five systems evaluated in Case Study 2. The lighter the background colour, the more important that component is in that particular domain.

Dealing with Uncertainty The ability to work in scenarios where not all information was specified fully was considered quite important in musical creativity and narrative creativity. It was deemphasised slightly in importance for artistic creativity and poetic creativity. This shows a difference between artistic and musical creativity on one hand, and poetry and narrative creativity on the other, contrasting with parallels drawn between these two pairings of creativity domains in other components. Perhaps matching the component's subject matter, judges generally felt uncertain about rating this component. Only Rahman & Manurung's poetry generator received ratings, being given a mid-range rating by both judges (5 and 4) and with *DARCI* (Norton et al., 2011) being given a high rating (8) by Judge 2. Especially for systems working with music or narrative, creativity evaluations would be more informed if given more information about the system's ability to deal with uncertainty.

Domain Competence Across all types of creativity in this case study, domain knowledge and expertise were judged to make at least quite an important contribution, with one of the two judges considering it a crucial component for poetic creativity and musical creativity. Four of the five systems were rated by at least one judge on this component (the exception was Cook & Colton's collage generator). Ratings were between 7 and 9 in all but one case (Monteith et al.'s soundtrack generator received 5 and 7 out of 10 from the judges). The poetry generator by Rahman & Manurung received the most universally positive ratings for this component (9 and 8).

General Intellect Whilst Judge 2 highlighted Cook & Colton's collage generator as the most intelligent of the three systems they rated for general intellect, Judge 1 did not feel able to comment on this component for a single system in evaluations. From the information on component importance, Judge 1 felt that this component was only a little important in all domains except for narrative creativity, so concentrated their attention on other components. Judge 2 considered this component to be quite important for poetic creativity, as opposed to only a little important, although they were not able to rate the poetry generator in Rahman and Manurung (2011) for its intellect. Otherwise the two judges were in agreement about the relative lack of contribution of *General Intellect* to creativity, echoing a similar distinction drawn between creativity and intelligence by several authors, (e.g. Guilford, 1950; Cope, 2005; Robinson, 2006; de Barros, Primi, Miguel, Almeida, & Oliveira, 2010).

Independence and Freedom Identified as at least quite important for all domains of creativity investigated in Case Study 2, and crucial for creativity in narrative generation by Judge 1, there was a lack of agreement on whether speakers at ICCC'11 provided enough information to judge systems on this component. Only one of the five systems was rated by Judge 1 (the musical system by Monteith et al.). Judge 2 rated four of the five systems on their freedom and independence, including Monteith et al.'s soundtrack generator, but the two judges disagreed on their rating (7 from Judge 1, but only 4 from Judge 2). Overall the collage generator by Cook & Colton appears the most independent and unrestricted, but there was a lack of information on this component to make more concrete evaluation statements, despite its perceived importance.

Intention and Emotional Involvement For all systems in all domains evaluated, Judge 2 thought this was quite important for creativity, whereas Judge 1 considered this only a little important. This is despite one ICCC'11 talk (Cook & Colton, 2011) promoting system intentionality as key for modelling creativity. Systems were left unrated or rated either 2 or 10 out of 10; no other scores were given. The judges agreed that Cook & Colton's collage generator demonstrated this component to a maximum degree (both scoring it 10 out of 10), in keeping with the amount of attention devoted to this component in the associated talk. They disagreed on all four other systems, either by not rating a system the other judge had rated or by giving opposite ratings. The rather human-like nature of this

component and a discrete interpretation (a system either demonstrates some intention and emotion or does not) may be an explanation for the more extreme ratings given for this component and the disagreement as to this component's importance.

Originality Judge 1's emphasis on the crucial contribution of originality in all but one domain (music, where it still made quite an important contribution) and Judge 2's opinion of this component as 'quite important' meant that this component should play a large part in evaluations. This was also discussed in Chapter 3 Section 3.4.1. Often, though, the two judges were unable to comment on *Originality* in a system. A system can itself be original, use original processes or produce original work. The ICCC'11 talks mostly reported details of how the system operates, with some information on the system itself and its products. Generally though, the talks provided a lack of examples to illustrate the system's work and there were no demonstrations of the systems at work.[21] This is at odds with the focus placed on the end product in previous creativity evaluation methods (Ritchie, 2007). Of the systems that were rated, the *MINSTREL* reconstruction by Tearse et al. was rated quite highly by both judges (8 and 7) and the soundtrack generator by Monteith et al. was rated the same by Judge 2 (but left unrated by Judge 1). Where a system's originality is commented on in presentation, the comments left a favourable impression. This information was however missing surprisingly often.

Progression and Development While commonly considered 'quite important' for creativity across domains, this component was seen as crucial for narrative creativity by Judge 2 and crucial for musical creativity by Judge 1. Where ratings were given for this component, all ratings were 5 or above. Rahman & Manurung's poetry generator received the highest rating (9 from Judge 1, supported by a 7 from Judge 2). Tearse et al.'s narrative generator was given a middling rating of 5 from both judges. The two artistic systems (Cook & Colton, 2011; Norton et al., 2011) were only rated by Judge 2 (each receiving a reasonably good rating of 7); Judge 1 did not feel there was enough information about this to offer their expert opinion in this component, despite considering it quite important. Despite its importance for musical creativity, neither judge could rate the soundtrack generator by Monteith et al. for this component.

Social Interaction and Communication This provoked the greatest overall differences in opinion between the two judges. Judge 2 considered this a crucial component for creativity in all case study domains. Judge 1 saw this component either as quite important (for poetic creativity and narrative creativity, both of which require linguistic creativity) or only a little important (for artistic creativity and musical creativity where the main medium of creative expression is not usually based around

[21] This evaluation case study could have been conducted during the 'Show and Tell' session at ICCC'11, where systems were demonstrated in action and questions could be asked about the system, but both judges felt that the associated practical difficulties with recording evaluation data and extracting information at these more informal sessions made the talk sessions a more viable option for evaluating systems.

words). Four of the five systems were rated by both judges. Ironically the narrative generator system[22] was the only system left unrated; both judges felt they had insufficient information to evaluate it on its communication and interaction. *DARCI* (Norton et al., 2011) received the highest ratings for this component (10 and 8), followed by Cook & Colton's collage generator (9 and 7). That the artistic systems performed best here suggests that communicative interaction may be simpler to achieve in image form than in words or music. The soundtrack generator (Monteith et al., 2011) was rated lower than these two artistic systems (5 and 5). Opinion was split on Rahman & Manurung's poetry generator (a high rating of 8 from Judge 1 against the low rating of 2 from Judge 2).

Spontaneity and Subconscious Processing Only one system was rated by both judges (Norton et al., 2011), receiving middling to high ratings (5 and 7). Judge 2 also gave a middling rating to Cook & Colton's collage generator (5). For other systems, the judges did not have enough information to feel confident making initial judgements about this component. The two attributes in this components are both highly related to human creativity; again this may have indirectly contributed to the lack of ratings provided for the computational systems, even though Judge 1 considered this crucial, and Judge 2 quite important, for all types of creativity examined. Perhaps this is due to some extra thought being required[23] to attribute spontaneity to a computer system, or to conceive of a computer doing some processing at a subconscious level.[24]

Thinking and Evaluation Although evaluation can be seen as a key part of the evaluation process (Chapter 3 Section 3.4.1), this was given slightly lower priority than other components by the judges in this evaluation. This is influenced by Judge 1 seeing this component as only a little important for all domains in the case study except poetic creativity, although Judge 2 considered this component quite important for all types of creativity considered in the case study. Judge 1's opinions on the importance of the *Thinking and Evaluation* and *Spontaneity and Subconscious Processing* components suggest that this judge prefers to interpret creativity in general as being spontaneous rather than the result of reasoning, unlike Judge 2 who sees these as equally important. Given the higher emphasis of Judge 1 for this component in poetic creativity, it is notable that Judge 1's rating of the poetry generator system (Rahman & Manurung, 2011) was high (9). Judge 2 also gave this system a reasonably high rating (7), marking it as the best performing system on this component overall. This reflects the importances attached to this component in poetic creativity as opposed to other domains considered. Of the other systems, Judge 2 considered the systems by Norton et al. and Tearse et al. to be equal in performance on this component to the poetry generator. Judge 1's rating of *DARCI* (Norton et al., 2011) (5) lowers the overall evaluation of *DARCI* on this component, though. Judge 1 did not have

[22] Arguably this could be said to be the system most aimed at communicating content from the narrator to the audience.
[23] And possibly subconscious reticence, despite the judges' experience with computational creativity research.
[24] See Chapter 1 Section 1.4.

enough information to rate this component for the *MINSTREL* reconstruction (Tearse et al., 2011).

Value Much of the comment for *Originality* applies here. An aspect of creativity that is highly prioritised by the judges in this case study and in wider research (Chapter 3 Section 3.4.1) was reported to a surprisingly low degree in the ICCC'11 talks. Unlike the originality of systems though, judges were more critical of the purported value of systems when they felt able to rate the system on this. Monteith et al.'s system was given quite high ratings for value (8 and 7) and *DARCI* (Norton et al.'s system) was given a score of 7 for *Value* by Judge 2 (but left unrated by Judge 1). In contrast, the value demonstrated in the *MINSTREL* reconstruction by Tearse et al. was given only middling ratings (4 and 5). Judge 1 felt unable to rate the value in both artistic systems despite expert knowledge of artistic creativity. Again this may be down to the lack of output examples given or system demos, or it may be due to not enough information about the inherent value in the system.

Variety, Divergence and Experimentation The importance of this component for creativity varied according to the creative domain for Judge 1, though Judge 2 thought it was quite important in all domains covered. Whilst considering this component crucial for creativity in generation of poetry and music, and quite important for narrative creativity, Judge 1 found this component only a little important for artistic creativity, the domain where they had expert knowledge. All ratings provided for this component were around 5 or 6 with one exception, where Judge 1 found *DARCI* (Norton et al., 2011) to score a maximum rating of 10 for this component. Judge 2 disagreed with this high rating, giving *DARCI* a rating of 5. Both judges agreed that Tearse et al.'s narrative generator deserved a middling rating (5 and 6) and all other systems were rated 5 by one judge and left unrated by the other . In other words, apart from the 10 rating given to *DARCI* by one judge, no system stood out as performing particularly well for this component.

Using the data on component importance

From inspection of the ratings, no one system stands out as being rated higher overall than the other systems. With two judges and a lack of ratings in several components, statistical analysis of the results would not be appropriate here. From the qualitative information gained through this evaluation process, however, more useful feedback can be identified:

- The collage generation module of *The Painting Fool* system (Cook & Colton, 2011) performed well on 1 of the 4 most important components for artistic creativity. There was insufficient information to evaluate this system's performance on the other 3 components.
- The poetry generator (Rahman & Manurung, 2011) performed well on 2 of the 6 most important components for artistic creativity but was given middling ratings for 1 of these 6 components, with insufficient information preventing evaluation of the other 3 components.
- The image generator *DARCI* (Norton et al., 2011) performed well on 2 of the 4 most important

components for artistic creativity but was given middling ratings for 1 of these 4 components, with insufficient information preventing evaluation of the other 1 component.

- The narrative generator (Tearse et al., 2011) performed well on 2 of the 7 most important components for artistic creativity but was given middling ratings for 2 of these 7 components and low ratings for 1 of these 7. There was insufficient information to evaluate this system's performance on the other 2 components.
- The soundtrack generator (Monteith et al., 2011) performed well on 2 of the 6 most important components for musical creativity but was given middling ratings for 1 of these 6 components, with insufficient information to evaluate this system's performance on the other 3 components.

This information is summarised in Table 7.8.

Table 7.8: The performance of Case Study 2 systems along the components identified as most important for that particular system's domain.

System	Cook & Colton	Rahman & Manurung	Norton et al.	Tearse et al.	Monteith et al.
# Key components	4	6	4	7	6
# High ratings	1 (25%)	2 (33%)	2 (50%)	2 (29%)	2 (33%)
# Middling ratings	0 (0%)	1 (17%)	1 (25%)	2 (29%)	1 (17%)
# Low ratings	0 (0%)	0 (0%)	0 (0%)	1 (14%)	0 (0%)
# Left unrated	3 (75%)	3 (50%)	1 (25%)	2 (29%)	3 (50%)

Therefore, though all presenters could have provided more information on important components for their system, some systems performed better in evaluation than others, notably *DARCI*:

- *DARCI* (Norton et al., 2011) was rated highly on 50% of the components key to creativity in its domain, with the remaining systems scoring between 25% (Cook & Colton, 2011) and 33% (Rahman & Manurung, 2011; Monteith et al., 2011).
- Accounting for middling ratings as well, again *DARCI* is ahead of the other systems, with 75% of its key components receiving a high or middling rating. Three systems had 50-54% of its key components receiving high or middling ratings (Rahman & Manurung, 2011; Tearse et al., 2011; Monteith et al., 2011). *The Painting Fool*'s collage generator (Cook & Colton, 2011) only received high or middling ratings for 25% of its key components.
- Tearse et al.'s reconstruction of *MINSTREL* was the only system to receive a low rating for one of its key components, though it did have the largest number of key components to address.
- The data in Table 7.8 can be quantified such that high ratings score 2 points, middling ratings score 1 point and low ratings or unrated components score 0 points, with the total divided by

the number of key components.[25] With this scoring system, overall rankings can be generated:

1. Norton et al.: 5/4 = 1.25 points.
2. Tearse et al.: 6/7 = 0.857 points (to 3 s.f.).
3. Rahman & Manurung: 5/6 = 0.833 points (to 3 s.f.) and Monteith et al.: 5/6 = 0.833 points (to 3 s.f.).
4. Cook & Colton: 2/4 = 0.5 points.

- Of the ICCC'11 talks, the two talks on *DARCI* (Norton et al., 2011) and by Tearse et al. (2011) supplied most creativity evaluation information, given the percentage of components which the judges could evaluate from the talk. Cook & Colton's talk was least informative to the judges for this purpose, probably due to its focus on system intentionality.

Overall reflections

Figure 7.2 shows that the most important contributions to creativity across all domains came from:

- *Generation of Results*.
- *Originality*.
- *Spontaneity and Subconscious Processing*.
- *Value*.

There was some variance between creative domains and between judges but on the whole these four components were usually deemed most important for creativity, showing the judges' emphasis on being original, useful and productive without doing this too mechanically or laboriously. To this list of four components may also be added *Social Interaction and Communication*, but the judges disagreed several times over the importance of this component.

There were variances between domains:

- In artistic creativity, the four (or five) components identified above were highlighted as most important for creativity.
- In poetic creativity, as well as the components identified above, *Domain Competence* and *Variety, Divergence and Experimentation* were thought important. This shows a need for informed, skilful exploration and divergent experimentation if constructing a narrative creatively.
- In narrative creativity, *Independence and Freedom*, *Progression and Development* can be added to the four components listed above and there was more agreement about the particular importance of *Social Interaction and Communication*. This type of creativity focuses more on the ability of the system to develop and progress without external constraints, learning from its environment while not being restricted by it.

[25] It is acknowledged that this is one of several ways to quantify this information.

- In musical creativity, *Active Involvement and Persistence*, *Domain Competence*, *Progression and Development* and *Variety, Divergence and Experimentation* were all prioritised in addition to three of the four components above. *Value* was less important for musical creativity than other types, though it was still considered quite important by both judges. This reflects the judges' interpretation of musical creativity as skilfully and competently trying out new things in order to develop on what is being done, without being put off by problems encountered, or by always needing to demonstrate value and usefulness.

Overall, all systems were judged to be excellent at *Generation of Results*, with almost agreement between the judges. *Originality* was rated highly when discussed in enough detail by the presenters (Tearse et al., 2011; Monteith et al., 2011). In contrast *Value* varied between the systems, with Monteith et al. performing the best, matched in the opinion of Judge 2 (but not Judge 1) by *DARCI* (Norton et al., 2011). Whilst the importance of *Spontaneity and Subconscious Processing* was emphasised by judges, generally the ICCC'11 talks contained insufficient information to rate this concept.

Judges tended to agree about components such as *Generation of Results*, *Domain Competence*, *Progression and Development*, *Thinking and Evaluation* and *Value*. These are all components which we are familiar with evaluating, for example, in education tests, quality testing or skills testing. Components such as *Intention and Emotional Involvement* and *Social Interaction and Communication* are less familiar in this context; these components showed the least agreement between the two judges.

Individual observations that stand out from the evaluation data include:

- The maximum ratings for *Intention and Emotional Involvement* for Cook & Colton's system.
- The highest ratings in *Domain Competence*, *Progression and Development* and *Thinking and Evaluation* being awarded to Rahman & Manurung's poetry generator.
- The two artistic systems (Norton et al., 2011; Cook & Colton, 2011) standing out as best performing for *Social Interaction and Communication*. Of these two systems, Norton et al.'s *DARCI* received the highest ratings from both judges; this is appropriate given that it works by interacting with an online audience and learning from human-supplied feedback.

Judges reported how the talks at ICCC'11 often did not deliver sufficient information on the components considered most important for creativity in that domain. In many cases, though, the talks did deliver information about components such as *General Intellect*, which was considered only a little important for creativity in all but one domain (though still identified as contributing to creativity in general, as Chapter 4 shows). With such time limitations, for maximum reflection of the creativity of their system, presenters should focus their talks based on the evaluative feedback identified throughout the analysis in this Section. Then our initial evaluations of a system's creativity can be based on the most important criteria, allowing more accurate first impressions to be formed.

7.5 Summary

Case Study 1 (Chapter 6) performed detailed expert evaluation focusing on a single creative domain. In contrast, Case Study 2 explored an alternative application of SPECS, showing how SPECS could be applied to evaluate the creativity of various types of creative systems from different creative domains. This Case Study aimed to capture first impressions and initial evaluations of how creative systems were, with limited information and resources, and under time pressures (Section 7.1).

At the International Computational Creativity Conference in April 2011 (ICCC'11), a variety of creative systems were presented. Section 7.2 reported how the format at this conference allowed the speakers seven minutes in which to present their systems, giving presenters a significant challenge on what information to include in this limited time. Section 7.3 described how five of the systems presented at ICCC'11 were evaluated through first impressions of their creativity. Demonstrating how SPECS can be applied to systems in various creative domains, the systems in Case Study 2 came from various domains, from artistic creativity to poetry construction. Two judges evaluated each system by rating system performance out of 10 on each of the 14 components based on the information imparted in the respective ICCC'11 presentation. If judges felt insufficient information was given on a particular component in the seven minutes, it was left unrated. Judges also rated how important they felt each component was in the system's creative domain, as well as their own expertise in that domain.

Analysis of these evaluations has provided two types of information that could be useful to researchers, even given the constraints on the time and resources available for evaluation:

1. Information about how creative the systems were perceived to be and what information contributed to this, relative to the creative domain.
2. Whether the system presentations at ICCC'11 provided all the information considered to be most important for creativity in that system's domain.

Although the latter point was not an intended aim of Case Study 2, such information is useful to the authors as feedback about the information they communicated in this short time frame.

Seven minutes is admittedly not a long time to learn about a system. Both judges, however, felt well enough informed about each system to rate its abilities and performance on at least some of the components. Section 7.4 presented the results of Case Study 2, with detailed discussion of each individual system's results (Sections 7.4.1 to 7.4.5) and comparisons across all results obtained in Case Study 2 (Section 7.4.6). More evaluation ratings were collected for components that can be objectively measured such as *Generation of Results*, or verified by inspection, such as seeing the amount of interaction and communication for *Social Interaction and Communication* or the amount of variety for *Variety, Divergence and Experimentation*. Often components were left unrated by both

judges, showing that the presentation focused on some aspects of creativity at the expense of others (probably due to the time limit and the intentions of the speakers for their talks).

For components that are more amenable to being tested for in existing types of tests, such as *Value* or *Progression and Development* in educational testing, the judges gave more similar ratings than for components such as *Intention and Emotional Involvement* or *Social Interaction and Communication*, where there were notable discrepancies; differences of 8 marks and 6 marks respectively could be found in these component's evaluation data.

Case Study 2 also looked at how judges perceived disciplinary priorities for different types of creativity. Judges tended to favour some components over others across all creative domains, possibly revealing slight differences in their theoretical perspectives on creativity.[26] For example, Judge 1 consistently emphasised the importance of *Spontaneity and Subconscious Processing* for creativity, favouring a view of creativity incorporating 'aha' moments of inspiration rather than considered reasoning. In contrast, Judge 2 emphasised a view of creativity as situated within an environment of *Social Interaction and Communication*. As discussed in Chapter 3 Section 3.4.2, both these views are compatible but have different foci, either on the creative individual's reactions to the creative environment (Judge 1) or the overall view of interactions and influence between the creative person and their environment (Judge 2). Again the issue of differing views of creativity arose, echoing part of Case Study 1's findings (Chapter 6 Sections 6.4.5 and 6.5). This issue will be investigated alongside several others in Chapters 8 and 9, as the SPECS evaluation methodology is itself evaluated.

[26]Both judges are computational creativity researchers.

Chapter 8

Evaluation: Comparing alternative creativity evaluations of the case study systems

---

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012).

Figure 8.1: *Wordle* word cloud of this Chapter's content

Overview

The SPECS methodology for evaluating the creativity of computational systems has been used for the evaluation of several creative systems, using the set of components identified in Chapter 4 as a working and customisable definition of creativity. The components have been applied in SPECS evaluation both for detailed expert evaluation (Chapter 6) and in forming first impressions of the creativity of a creative system using limited information (Chapter 7). The SPECS approach has given detailed information on each system's strengths and weaknesses through constructive formative feedback. It has also afforded a comparative evaluation of the creativity of different systems, providing a needed solution to a complex methodological issue (Jordanous, 2011a, also Chapter 2 of this thesis).

As Chapter 2 has described, other creativity evaluation methodologies and strategies have been proposed in the past. This current Chapter describes how alternative creativity evaluations to SPECS are also conducted on the Case Study 1 systems (Section 8.1), using: a survey of human opinion (Section 8.1.1); Ritchie's empirical criteria framework (Ritchie, 2007, Section 8.2.1); and Colton's creative tripod framework (Colton, 2008b, Section 8.2.2). The evaluative process and results are compared and contrasted in Section 8.3.1. Sections 8.3.3, 8.3.4 and 8.3.5 consider how usable and accurate the evaluation data was when obtained through human opinion surveys, Ritchie's criteria and Colton's creative tripod, respectively.

External evaluation of the evaluation methodologies is carried out with the researchers behind the systems from Case Study 1 (Section 8.4). Five meta-evaluation criteria are identified in Section 8.4.2 from related literature and other sources and are used to evaluate and compare the success of the different computational creativity evaluation methodologies implemented in this Chapter, supplemented by feedback obtained via the application of the FACE framework recently proposed by Colton et al. (2011). The meta-evaluation results and feedback, reported in Section 8.4.4, highlight the relative strengths and weaknesses of each creativity evaluation methodology in the context of Case Study 1. To complement this external evaluation, the Section 8.4.2 criteria are used to collate further considerations of the detailed findings in this Chapter, for both case studies. Results of these considerations are reported in Section 8.5.

8.1 Using alternative evaluation methods: Human opinions of system creativity

The Chapter 2 Section 2.3 survey on evaluation methods reported that computational creativity researchers often rely on human judges' opinions to evaluate how creative their system is. Comments on the subject of this thesis have commonly taken the form 'why don't you just ask people how creative the system is?', under the assumption that people's working definition of creativity was intuitively sound and the definition of creativity did not need further clarification. This assumption has

percolated through to research literature:

> 'A major difficulty in studying creativity is the lack of an objective definition of creativity. Because creative writing is highly subjective ("I don't know what is creativity, but I recognise it when I see one"), we circumvent this problem by using human judgement as the ground truth.' (Zhu, Xu, & Khot, 2009, p. 87)

There are issues with assuming that a 'ground truth' exists with creativity judgements, due to variability in opinions, and time or context-dependent factors that may influence this judgement.[1] But is this a suitable way to evaluate systems' creativity? This method of creativity evaluation was therefore applied to each system in the case studies so that evaluation information obtained this way could be compared with the SPECS evaluation results and feedback.

8.1.1 Human opinions of creativity in Case Study 1: Musical improvisation systems

To examine if human judgement can be relied upon as 'the ground truth', an online survey was carried out collecting people's opinions about the creativity of the three systems in Case Study 1 (Chapter 6). If people are purely asked how creative these three systems are[2] and if they are asked to compare the three systems' creativity, what evaluative feedback will be obtained?

Survey methodology

On starting the survey, participants were informed that the survey was for a research project on computational creativity systems. Computational creativity was described to participants as computers performing in a way which, if a human were to demonstrate that behaviour, would be called creative. Throughout the survey it was made clear that music was being improvised by computer systems designed for this task, rather than humans. Participants were also given a very brief introduction to the research project before starting the survey.

The survey was promoted on social networking sites and relevant mailing lists, with the intent of recruiting participants with a range of levels of expertise in improvisation and computer programming. During the survey, participants were asked about their expertise and experience in improvisation, music and computer programming so their answers could be put into this context. Participants were encouraged to take part, however knowledgeable they felt about improvisation and computer music.

The three Case Study 1 systems were presented to participants in randomised order, to avoid any biases arising from ordering. For each system, participants were presented with a short description of the system and how it works, and three short (30 seconds) recordings as examples of improvisations by the system. To give more details of the 9 extracts:[3]

[1] Variances in opinions on creativity will be examined further in Chapter 9 Section 9.1.4 and Chapter 10 Section 10.4.2.
[2] Rather than performing more detailed reductionist evaluations of creativity.
[3] All extracts are available at `http://www.informatics.sussex.ac.uk/users/akj20/creativitySurvey/`.

GAmprovising During and after development, GAmprovising had been run several times. By default, GAmprovising saves improvisations of the population of 'Improvisers' after the final cycle of evolution in the genetic algorithm that is used; hence a large number of sample tracks had been saved during development. A prior decision had been made to use extracts of 30 seconds length for all systems, which was long enough to give an idea of the tracks but short enough so that 9 tracks[4] could be listened to without making the online surveys too long. Of the sample GAmprovising tracks, therefore, those that were longer than 30 seconds were collected together and the system programmer (myself) selected three tracks or 30-second excerpts of tracks that best represented the variety and quality of the system's products. This method of choosing three example recordings was preferred to random selection as it more closely matched that of the other systems, where the system programmers had selected their choice of tracks to make available as examples of that system.

GenJam The first track for GenJam was a solo chorus by GenJam taken from the track 'Analog Blues' on the 1996 CD entitled 'GenJam' by the 'Al Biles Virtual Quintet'. This is an original composition by Biles, described as 'an island-flavored, syncopated ramble' (Spivak, 1996). There was also a track with two solo choruses by GenJam from the track 'Change Tranes' over the chord sequence to the jazz standard 'Giant Steps'. This track was included on a CD accompanying Miranda and Biles (2007) which has a Chapter on GenJam (Biles, 2007). The third track saw GenJam trading solos with Al Biles on trumpet (taking it in turns to solo) for 8 bars each and then for 4 bars, on a latin track entitled 'The Rake' (Miranda & Biles, 2007).

Voyager The extracts used to demonstrate Voyager were 30-seconds taken from three tracks on the 1993 album 'Voyager' (Avant): 'Duo 1', 'Duo 2' and 'Duo 3' (all recorded as interactions between Voyager and George Lewis on trombone)

Given these resources, they were asked how creative they thought that system was, giving comments if they wanted to. Responses were via a Likert scale ranging from 'Not at all creative' to 'Completely creative' (Figure 8.2). Participants were asked not to spend too long thinking about their answers, but to give answers as they came to mind. They were also asked how confident they felt about their answers, responding using a second Likert scale (Figure 8.3).

It was expected that few, if any, participants would have heard of the case study systems before. To check this, participants were asked for each system if they had heard of it before and if so, were asked to describe in what context they had encountered the system.

After all three systems had been presented, participants were given a list of the three systems and a sample track from each system as a reminder of what it sounded like. They were asked to rate the

[4]Originally 12 tracks, given that a fourth system was originally also included in Case Study 1. As explained in Chapter 6 Section 6.2, this system has since had to be removed from the reports of Case Study 1 due to problems with the origin of the musical examples used.

Figure 8.2: Responses available to participants when asked how creative a system is.

Figure 8.3: Responses available to participants when asked how confident they felt about a given answer.

three systems in terms of creativity (from most creative to least creative) and to indicate how confident they felt about their answer, with room to comment if desired.

Participants were then asked if they felt that in general they had been given sufficient information to evaluate the systems, using a Likert scale (Figure 8.4) to answer. If participants would have liked other information when judging the systems, they were asked what information they would have liked. After this, participants were asked to summarise how confident they felt in general about their answers throughout the survey, using the Likert scale in Figure 8.3 and commenting if they wanted.

Figure 8.4: Responses available to participants when asked if they agreed with a given statement.

A potential source of bias, given that participants knew the music was being improvised by computer systems, was in people's reactions to computational creativity (Chapter 1); as Moffat and Kelly (2006) found, people's perceptions of creativity may be influenced by conscious or subconscious biases about computational creativity. Although it was not practical to try to capture subconscious biases via the online survey format, the final question in this survey looked at participants' views and attitudes towards computational creativity in general. Participants were presented with eight statements (in randomised order) covering a range of opinions on computational creativity:

- I like the idea of computers being creative.
- Talk of computational creativity shocks and disturbs me.
- Computer systems cannot be creative yet but could be in the future.
- Computer systems can produce creative behaviour/actions.

- The thought of computational creativity is exciting.
- There are already examples of computer systems being creative.
- It is possible for computer systems to be creative.
- Computer systems cannot be creative at all and never will be.

Participants were asked to what extent they agreed or disagreed with each statement. The responses, reported below in Section 8.1.1, gave some idea as to participants' general attitudes to computational creativity. This helped to indicate any possible biases affecting their answers in the overall evaluation survey (at least, as far as they were aware of their attitudes and biases).

After being given a final opportunity to comment on any aspects of the survey, participants were thanked and debriefed with some more information on the specifics of this research.

Participant demographics

Participants were recruited from social networking sites, some mailing lists and from musical contacts, with participants encouraged to publicise the survey to others if they felt inclined. Participants were asked about basic demographics:

- Gender [male / female].
- Age group [18-24 / 25-34 / 35-54 / 55+].
- Highest educational level reached [see Figure 8.5(c) for options].

Of 111 respondents, 73 were male and 38 female (Figure 8.5(a)). This male/female imbalance was expected to some extent; in common areas of musical improvisation such as jazz and free improvisation, men are generally more highly represented than women.[5]

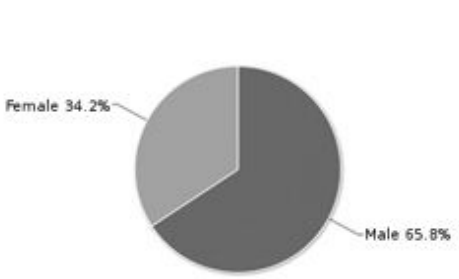
All age groups were represented (see Figure 8.5(b)), with the most-represented age group being 25-34 years.[6] The median age of participants was in the 25-34 age group, at approximately 29.9 years. All but 13 participants were educated to at least degree level (Figure 8.5(c)), with 60 participants having achieved or studying at post-graduate level. This perhaps reflects the use of academic mailing lists to publicise the study, though there may be several other reasons such as a large proportion of immediate recipients of social networking adverts being university students or graduates, an increased uptake in university education and other reasons.

Status as a computer programmer

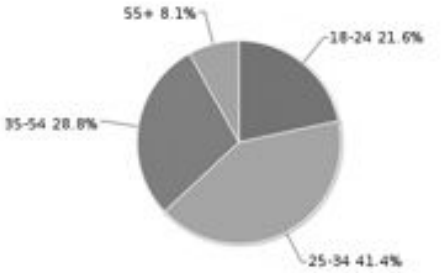
- Would they describe themselves as a computer programmer?

[5] An example from my own personal experience is in big bands, who play music with a significant improvisational component, it is not atypical to see only a couple of women (or fewer) in a band numbering 16-18 (or more) people.

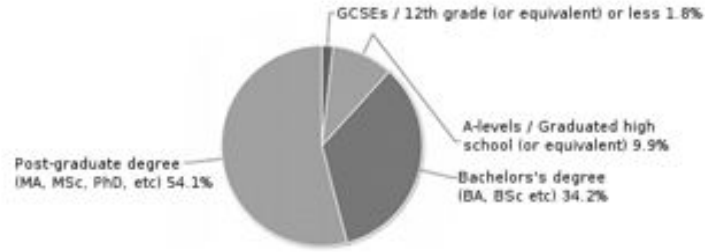
[6] Possibly the reason for this peak in the age groups was a consequence of using social networking seeded with my own contacts as one of the recruitment methods for the survey and also perhaps due to basing the survey online rather than conducting it using other means.



(a) Gender.



(b) Age groups.



(c) Highest level of education reached (completed or current study).

Figure 8.5: Demographics for participants in the survey in Section 8.1.1.

  - No. Not a computer programmer.
  - Yes. Amateur computer programmer (can write programs but no money earned from it).
  - Yes. Semi-professional computer programmer (earn some money through programming but it is not the main source of income).
  - Yes. Professional computer programmer (earn a living primarily through programming).
- Expertise in computer programming. [see Figure 8.7(a) for options]
- Number of years of computer programming experience and training.

Out of 111 respondents, nearly 50% did not consider themselves to be a computer programmer (54 / 111). Of the remaining 57 participants, 19 were professional computer programmers, 14 were semi-professional and 24 were amateurs. These proportions were evened out to some extent when participants were asked about their computer programming expertise. Whilst 34.2% reported no knowledge of computer programming at all (38/111), 22.5% felt they had basic knowledge (25/111), 25.2% reasonable knowledge (28/111) and 18% expert knowledge of computer programming (20/111).

Figure 8.6 shows that a significant proportion of the participants (42/111 or 37.8%) had no years of programming experience or training. 32 participants had between 1-5 years' programming experience and 17 people had 6-10 years programming experience. This accounts for all but 20 participants.

Figure 8.6 shows a decrease in frequencies as the number of years' experience in the groups increases.

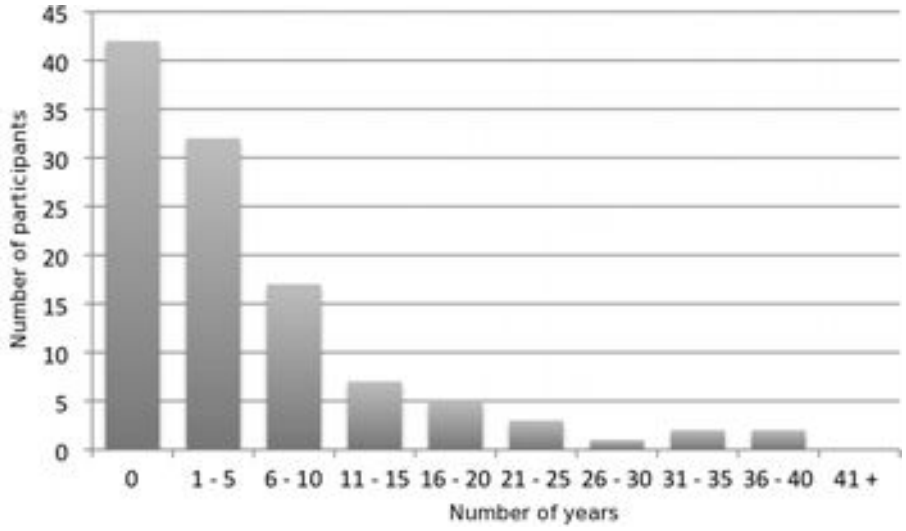


Figure 8.6: Responses to the question: 'How many years of computer programming experience/training do you have?' in the survey in Section 8.1.1.

Status as a musician

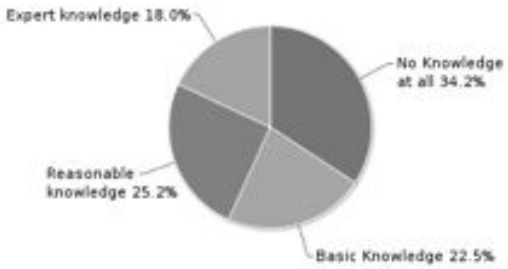
- Would they describe themselves as a musician?
  - No. Not a musician.
  - Yes. Amateur musician (can play music but no money earned from it).
  - Yes. Semi-professional musician (earn some money through music but it is not the main source of income).
  - Yes. Professional musician (earn a living primarily through music).
- Expertise in musical improvisation (performing and/or listening) [see Figure 8.7(b) for options].
- Expertise in music (not just improvisation) (performing and/or listening) [see Figure 8.7(c) for options].
- Number of years of musical experience and training.

All but 19 out of 111 respondents (17.1%) described themselves as musicians to some extent.[7] Of the 92 musicians, 47 considered themselves amateurs, 24 semi-professional and 21 professional musicians. 17 out of 111 people claimed no knowledge at all of musical improvisation (15.3%). Most people reported reasonable knowledge of musical improvisation (44.1% or 49/111). 31 people (27.9

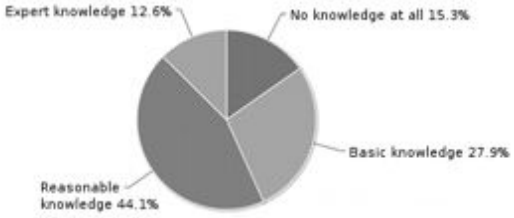
[7]This was unsurprising as many musicians passed on the survey to their fellow musicians.

%) had basic knowledge and 14 people had expert knowledge of musical improvisation (12.6%).

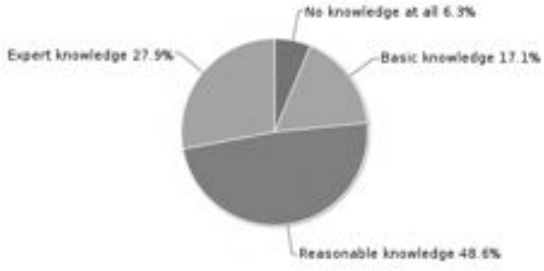
Levels of knowledge of music in general were higher than for knowledge specifically about musical improvisation, as Figures 8.7(b) and 8.7(c) show. People generally had some musical knowledge, even if they did not necessarily have as much knowledge about improvisation in music. Only 7 of the 111 people (6.3%) said they had no musical knowledge at all, with 19 people (17.1%) having basic knowledge, 54 people (48.6%) having reasonable knowledge and 31 people self-classifying themselves as musical experts (27.9%).



(a) Computer programming.



(b) Musical improvisation.



(c) Music in general.

Figure 8.7: Relevant expertise and knowledge of participants in the survey in Section 8.1.1.

The range of musical experience was more variable than for the range of computer programming experience, as can be seen from comparing Figures 8.8 and 8.6. Whereas all but 20 participants had 10 years or less in experience of computer programming, and 42 participants had no experience of computer programming, 64 participants (57.7%) had more than 10 years musical experience and only 12 participants had no years' musical experience. The mode average was in the category of 11-15 years experience, which represented 20 participants.

The categories representing greater experience with music were better represented than in computer programming. The maximum value given for musical experience was 67 years, as opposed to 40 years as maximum for computer programming, probably due to the limited length of time computers

have been available for common use.

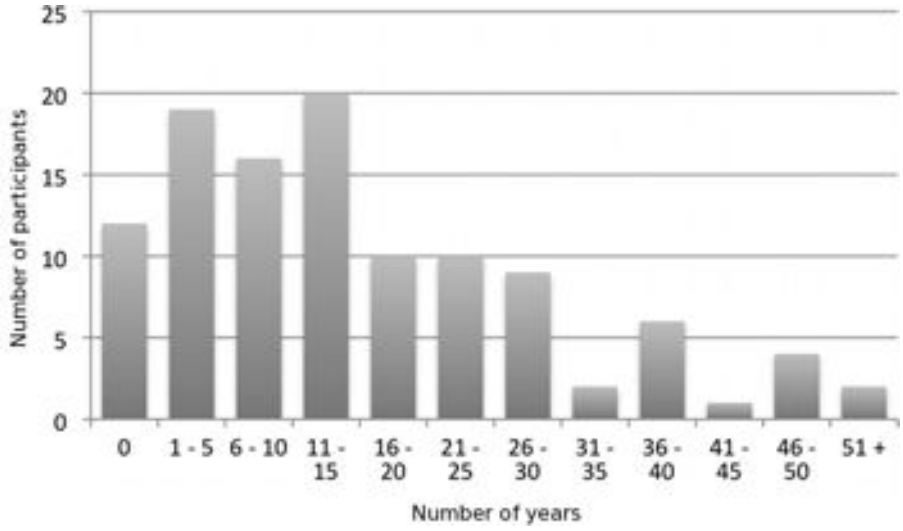


Figure 8.8: Responses to the question: 'How many years of musical experience/training do you have' in the survey in Section 8.1.1.

Main findings from the survey: Case study systems' creativity

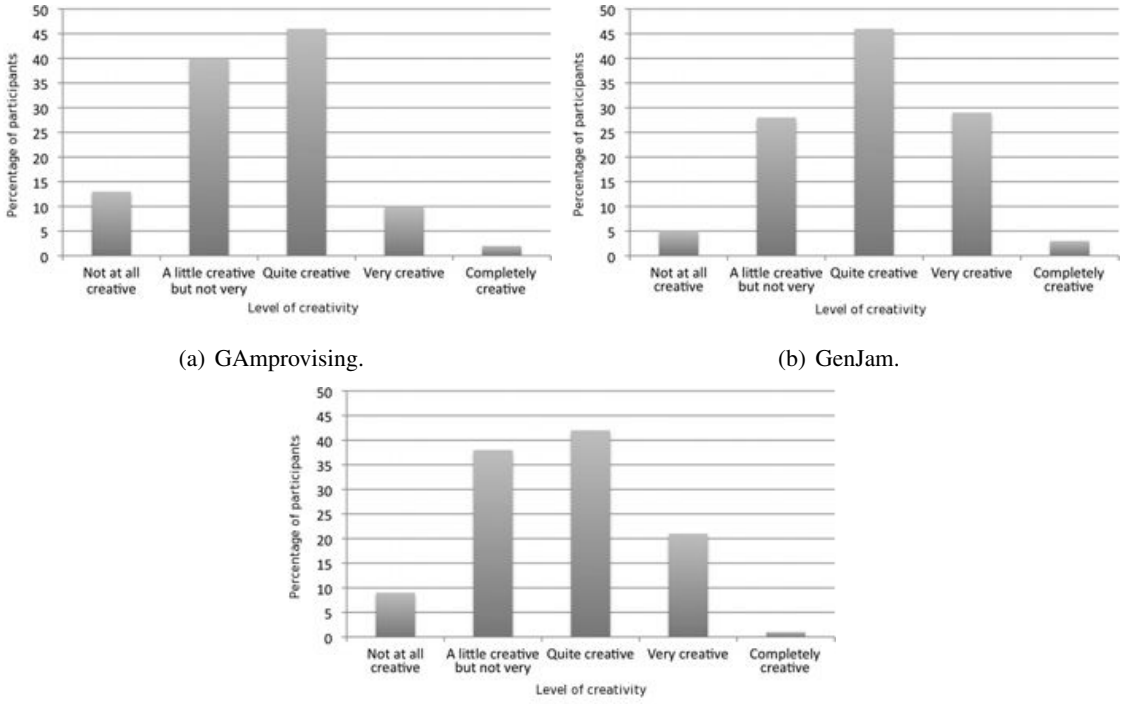
Table 8.1 summarises the data on ratings of creativity for individual systems and overall system rankings for creativity. Figure 8.9 displays individual creativity ratings for each system.

Table 8.1: The data obtained in the evaluation study in Section 8.1.1. In these results, 5 = Completely creative, 4 = Very creative, 3 = Quite creative, 2 = A little creative but not very,1 = Not at all creative. For confidence, 5 = Very Confident, 4 = Confident, 3 = Neutral, 2 = Unconfident, 1 = Very Unconfident. The ranking points score is a weighted calculation. Items ranked first are given 3 points, second 2 points, third 1 point. All points are summed together for the overall ranking points score.

System	Mean creativity	Creativity (mean rounded to 0 dp)	Mean Confidence	Confidence (mean rounded to 0 dp)	Ranking (points)
GenJam	3.0	Quite creative	3.7	Confident	1st (250)
Voyager	2.7	Quite creative	3.5	Confident	2nd (211)
GAmprovising	2.5	Quite creative	3.5	Confident	3rd (193)

As can be seen from Table 8.1, the labels did not differentiate between systems' creativity, nor the participants' confidence in determining a creativity label. While the numeric scores fared better, finding GenJam to be most creative, followed by Voyager, and finally GAmprovising, participants collectively did not distinguish between systems to any great extent when describing how creative the

systems were. The similar shapes to the bar charts in Figure 8.9 show the variance in opinions across each system, as perceived by a large group of people (n=111).



(a) GAmprovising.

(b) GenJam.

(c) Voyager.

Figure 8.9: Ratings of creativity per system, showing the spread of opinion and the lack of any conclusive 'crowd decisions' by the 111 people surveyed. Apart from a lesser tendency for GenJam to receive ratings of 'A little creative but not very' and for GAmprovising to receive ratings of 'Very creative', there is little noticeable difference between the distribution of ratings per system.

Participants were generally 'Confident' about the answers they gave in the survey. Figure 8.10 shows that the distribution of answers for 'How confident do you feel about your answer' was fairly consistent across systems and for the overall rankings given. The average level of confidence expressed and the standard deviation between confidence ratings was close in all five cases (Voyager: mean 3.7 s.d. 0.99; GenJam: mean 3.7 s.d. 0.89; GAmprovising: mean 3.5 s.d. 0.92; Overall ranking: mean 3.6, s.d. 0.9). It can be said from Figure 8.10 that participants were generally more confident about their answers for GenJam than for Voyager and GAmprovising, but the difference is small.

A large amount of qualitative data was collected during the survey. Participants were given space to comment on their answers for individual systems and on their answers overall for the survey. Many participants took the opportunity to give further feedback, providing data both for establishing systems' creativity and feedback on the experience of the survey process as a whole. On average

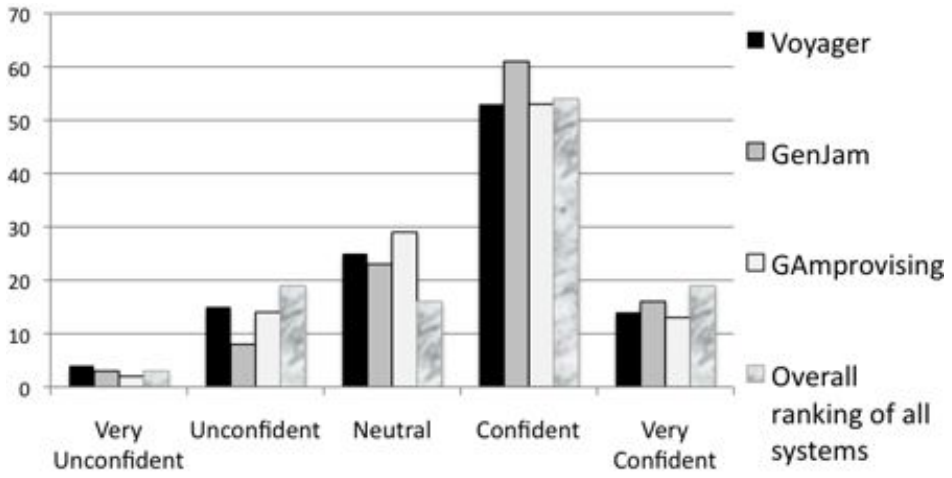


Figure 8.10: Responses to the question 'How confident do you feel about your answer', asked after rating the creativity of each of the three systems and also asked after ranking all three systems in order of creativity. The most frequent response was 'Confident', which received nearly double the responses of any other answer in the *x*-axis's Likert scale.

further comments were given on a system by 1 in 2 responses, suggesting a real engagement with the study and a willingness to provide qualitative data for the survey.

Qualitative feedback on systems' creativity This Section summarises the qualitative feedback given about each system. Positive comments are marked as '+', negative as '-' and neutral as 'n'. The number of comments making a particular point is given in square brackets after the description of that point. Some comments covered more than one point.[8]

- GAmprovising:

  One participant had heard of GAmprovising before, although they were not sure where. As the system in Jordanous (2010c) had only been given this name just prior to this study and had not previously been described as 'GAmprovising', this either indicates there is a system with a similar name or that the participant was mistaken.

  47 people chose to comment further on the creativity of GAmprovising, with comments averaging 181 characters each (s.d. 177).

  - Many people were negative about GAmprovising's performance or abilities [6].
  - The abstract, free genre of the improvisations produced confused some, due to the style of music and their preferences [3].

[8] A general point to note is that the comments seemed to centre around the systems' musicality and competence rather than the systems' creativity. This issue recurs throughout the thesis in different guises, most notably in Chapters 2 and 5 and in this Chapter in Section 8.1.1.

+ Some people liked the music produced [7].
+ GAmprovising was praised for its interesting experimentation, unconventionality and flexibility, making the listener think more [5].
- The music was criticised for being too random and lacking in coherence, direction and structure [17].
- Some saw the music as being bland or commented negatively on it sounding too computer-generated [5].
- The use of very high-pitched notes in some tracks was perceived negatively [4].
- The use of silence was criticised by its absence [1].
- Comments were made about GAmprovising lacking its own style and interpretation, or judgement [3].
- On the other hand, some people found GAmprovising to have too homogeneous a style, with too little variation [3].

n No common reactions emerged on individual tracks; people's comments on the tracks were inconsistent with each other.

- GenJam:

  Five participants had heard of GenJam before, in academic work (2 participants), a talk and demonstration (1 participant), seeing some videos in the past (1 participant) and through a musical colleague's use of a similar program (1 participant).[9]

  64 people chose to comment further on the creativity of GenJam (receiving the most extra comments of all three systems), with comments averaging 176 characters each (s.d. 175).

  - Some commented on the lack of direction and progression in GenJam's solos [2].
  + Others praised GenJam for its cohesion, melodic direction or natural flow [3].
  + Positive comments were received such as 'good idea' or 'like this, very good' [5].
  - Others drew comparisons between GenJam's solos and 'muzak'/'lift music'[10] or computer generated music, with a lack of 'feeling' [10].
  + In contrast, GenJam's soloing was described as sounding more natural, authentic or human by some participants [5].
  - Some people found that the solos sounded good at first but then became too smooth-sounding, unsurprising and boring, like an improviser who had run out of ideas or an improviser taking too conservative an approach or lacking in flair [12]. Others found

[9] This last participant's exact response was 'Computer programmer/musician where I work has done work with this, I think, and incorporated it into performances.', but as the GenJam system has proved unportable for use by other musicians, it is likely that the participant was not referring to GenJam but to a similar system (unless this is a colleague of Al Biles).

[10] The terms 'lift music'/'muzak' are negative terms referring to generated music in an easy listening style, often artificially generated and/or played in a repetitive loop, which is often used for unobtrusive, unnoticeable background music.

the system uninteresting or unmusical [2]. Questions were raised as to how much of the material in GenJam's solos was taken from pre-existing solos or repetitive [8] or whether GenJam thought about what it did and learned [1].

- Some commented that GenJam was poor at using silence in solos, a mistake often made by beginner soloists [2].

n GenJam's interplay with the trumpet player was praised by some [6] although it was sometimes criticised for not being sensitive enough to the human's playing [4].

- Comments were made about occasionally odd harmonic or rhythmic choices in GenJam's playing [5].

n Of the three pieces, some singled out the third extract (from 'The Rake') for particular positive mention [6] although although it also received some negative comments such as 'hilariously awful' [2]. The second track divided opinion between positive [2] and negative [4] and the first track was given mostly negative specific comments [6] with only one positive comment [1].

- Voyager:

  Four participants had heard of Voyager before, in academic work (1 participant), presentations (1 participant), a concert performance (1 participant) and from research after playing with a similar system (1 participant).
  53 people chose to comment further on the creativity of Voyager, with comments averaging 181 characters each (s.d. 182).

  + Voyager's interactivity was praised, in particular how it responded to the trombone player (Lewis) [7].
  - Conversely, others felt that the two players did not engage properly with each other [5].
  + The music produced by Voyager was considered to be convincing as avant-garde/abstract improvisation [6].
  + The system was described as clever [1], coherent [1] and versatile [1].
  - Some people observed that they did not like the system or its music, or did not find some of the excerpts pleasing to listen to [4].
  - Other people were confused by what the system was trying to do, particularly if they did not like the style of music [2].
  + Some participants found Voyager's music interesting, partly because of the unusual sounds and the style of the genre [2].
  - Voyager was often criticised for being too random [11].

- \- Rather than being viewed as an equal partner in the improvisation, Voyager was often perceived as being the accompanist to the human player. Sometimes this was because the participants' attention was diverted by the trombone playing [4].
- \- Comments were made that the creativity of Voyager was in the programmer, rather than in the system itself [4].
- \- Negative criticisms about computational creativity affected people's judgement of Voyager, as they felt improvisation was strictly a human activity or that it required human emotion [2].
- \+ On the other hand, some participants observed that Voyager's playing could pass for human improvisation [3].
- n Of the three tracks, if people singled out one for comment, it was usually a greater appreciation of the third track compared to the others: an excerpt from 'Duo 3' on the 1993 'Voyager' album (Avant Records) [3].
- \- The lack of structure in the music was criticised by some participants, though this may be a general reflection on the style of music rather than Voyager itself. Participants were unsure whether the unrestricted musical style of Voyager indicated greater creativity or a lack of musical identity [4].
- \- Although information was provided on the workings of the Voyager system, some participants wanted to know more about the decisions being made and reasons for those decisions. Lewis (2000) (the only publication found describing this system's functionality) is unclear on the specific details [2].

Additional findings: Participants' attitudes to computational creativity

In this survey (as in all surveys reported in this Chapter), participants were made fully aware of the fact that they were evaluating the creativity of computer systems, not human improvisers. The survey did not try to detect *subconscious* influences for or against computational creativity, preferring to retain focus on the primary information-gathering role of the surveys. Some closing questions in this survey did however attempt to gauge *conscious* attitudes towards computational creativity, by asking participants to what extent they agreed with various statements about computational creativity.

Table 8.2 presents these statements and the responses from the 111 participants.

Participants generally reported neutral or open-minded views towards computational creativity, being predisposed towards agreeing with the assumption made in this thesis that computational creativity is possible. During the surveys and during related feedback and discussions post-survey, though, some negative and occasionally defensive reactions were still to be found:

> 'I don't like the sound of this'

Table 8.2: Attitudes to computational creativity, as reported by participants in the Section 8.1.1 survey.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Computer systems cannot be creative at all and never will be	32 29%	45 41%	18 16 %	14 12%	2 10%
It is possible for computer systems to be creative	5 5%	14 13%	26 23%	58 53%	8 7%
Computer systems cannot be creative yet but could be in the future	5 5%	39 36%	41 36%	21 19%	5 5%
There are already examples of computer systems being creative	2 2%	12 11%	47 42%	46 42%	4 4%
Computer systems can produce creative behaviour / actions	5 5%	12 11%	23 20%	63 57%	8 7%
The thought of computational creativity is exciting	2 2%	11 10%	31 27%	44 40 %	23 21%
Talk of computational creativity shocks and disturbs me	43 39%	37 34%	21 18%	7 6%	3 3%
I like the idea of computers being creative	7 6%	4 4%	30 27%	50 46%	20 18%

> 'computer interfaces force people to dumb down to engage with their supersmart jazz-playing overlords.'
>
> '[On GAmprovising] Man you are nowhere. Go and listen to some real music'
>
> 'let's hope they don't replace real musicians!'
>
> 'You need to try a lot harder. What is missing in all examples so far is passion. Music is about something to say. These programs make sounds a bit like music but they have nothing to say. Like a baby babbling. Like the cargo cult. They so miss the point that it is painful.'

Generally comments such as those above were accompanied by fairly detrimental evaluations of creativity given in the survey by these participants.[11]

In attributing creativity to computational systems, many preferred instead to assign the demonstration of creativity to the people involved. Creativity was seen as being demonstrated by the programmer rather than by the computer receiving the instructions:[12]

> 'They're only as creative as the programmer who programmed them :) '

[11] This negative effect on responses is conjectured here rather than being objectively verified. The same participants could not then rate the same systems under the impression that they were human improvisers. Comparing responses to examples taken from human improvisers and from computer improvisers could be an interesting point for future work to tackle. For the purposes of this thesis, though, such experiments fall slightly outside the scope of this work, which is aimed at producing a practical evaluation methodology rather than making a significant contribution to the ongoing debate on how computational creativity is perceived by the layperson.

[12] The question of attribution of creativity has also arisen in Ritchie (2001) and Jennings (2010a).

> 'its the old tool makers as artist debate innit?'
>
> [On Voyager] 'It seems as though a lot of the creativity here is in the set of "methods" that have been hard-coded into the system. The _results_ are probably more diverse and innovative than those in the previous example; but, perhaps, less is being_learned_ from the experience with the human musicians. But, is this _learning_ the same as being _creative_? I don't know.'
>
> 'not sure it is the computer being creative - or the prog / system designer?'
>
> 'I feel fairly convinced that computational creativity is possible... although the creativity that is created is only possible because of the input of the researchers / programmers. I think that with the necessary hints and prods computers can probably produce creative music... but that this will probably only ever be possible with human input.'
>
> 'I'm wary of this becoming a linguistic exercise but if there was a computer system that produced incredible, interesting, real time and thought provoking music I probably wouldn't be praising the CPU or the ram for it's intelligence. I'd be incredibly impressed with the developer. Ultimately computers play around with numbers, they can be incredibly powerful tools but in my opinion the person who fits that tool for their purpose is creative.'

Another reason for not attributing creativity to the creative system was the reluctance to accept that a computer could replicate human skills and abilities which are considered valuable in humans. Some participants voiced their opinions that computers are mechanical automata that cannot diverge from their programming; this would require human intelligence and flexibility:

> 'I'm struggling to distinguish between "creative" and "random". I'm not sure creativity is an attribute you can ascribe to a machine? Surely it's only *aping* creativity?'
>
> 'My natural bias is toward work with live musicians, using acoustic instruments. That may have a great deal of influence on my answers. To me, when musicians improvise together, they get inside each other's heads, I don't see how this would work with a computer generated program.'
>
> 'I don't think that computers can be creative independently (i.e. they always require human intervention) – they can't lock themselves in a room for a few years, cut off their ear and then complete a masterpiece!'
>
> 'I agree that computers can probably be capable of creative "like" behaviours but i don't think that it can be thought of in completely the same as human creativity.'
>
> 'I'm doubtful about whether a computer will ever be able to match the creativity of a human, thus my neutral responses above. Until the computer's creativity can be responded to in the same way as a human's creativity, I cannot agree that a computer system is truly being creative.'

8.1.2 Human opinions of creativity in Case Study 2: Systems at ICCC'11

Similarly to Case Study 1, the evaluation results and feedback obtained for Case Study 2 (Chapter 7) were compared to human evaluations of the creativity of the case studies.

In Case Study 1, the evaluation results obtained using SPECS were compared against human judges' evaluations of the creativity of those systems. Bearing in mind that Case Study 2 was intended to simulate initial impressions of creativity, one appropriate way to gather such comparative data for this case study would be to take opinions on the systems' creativity from the audience of the ICCC'11 talks. Hence the same data would be used to formulate opinions as was used for the earlier evaluative

steps, and there would be a comparable level of research knowledge about creativity (though again differences in opinions and perspectives about creativity would influence the resulting evaluations).

In practice, however, it was unfeasible to collect this data. Of the 50-60 attendees of ICCC'11 (not all of whom were present for all five talks), several would be ruled out as evaluators due to potential bias from their involvement in work on the evaluated systems or similar systems. Even if conference attendees worked in areas unrelated to the evaluated systems, they would want to listen to the conference talks in terms of what they could learn for their own research, rather than focusing on evaluating the creative systems.[13]

It was decided, therefore, to use the same two judges and collect their overall impressions of how creative the five case study systems were. To avoid the previous evaluations influencing their impressions of the systems' creativity, a time period of three months was left between the initial evaluations and the collecting of the general evaluations. This gap in time allowed the memory of the initial evaluative details to fade whilst still being close enough to the ICCC'11 conference for the judges to remember the talks and the systems to some degree. To refresh the memory of the judges, they consulted the appropriate ICCC'11 proceedings papers (Cook & Colton, 2011; Rahman & Manurung, 2011; Norton et al., 2011; Tearse et al., 2011; Monteith et al., 2011), for seven minutes each (replicating the seven minute time constraint on gathering information on a system at ICCC'11). Whilst this meant that creativity evaluations were to some degree based on different information to the original evaluations, this compromise made sure that judges remembered the systems well enough to evaluate them. Having refreshed their knowledge of the system this way, the two judges were asked to say how creative each system was and their reasons and comments. Judges could choose from the following options to describe a system's creativity:

- Completely creative.
- Very creative.
- Quite creative.
- A little creative but not very.
- Not at all creative.[14]

Judges were also asked to rank the five systems in terms of relative creativity. The judges' rankings are listed in Table 8.3. Their feedback on individual systems' creativity is listed below.

- Cook and Colton's system: *Not at all creative* (Judge 1), *Quite creative* (Judge 2).

  'Similar to DARCI. The trivial part of the task is done by the software whereas the mean-

[13] This observation could also be attributed in part to Judges 1 and 2 at the time of the original evaluation, though the two judges were instructed to listen to the talks with the purpose of evaluation uppermost in mind, which both judges did to the best of their ability.

[14] In practice, the two judges only used the lower three options.

ingful part is provided by the humans who are involved. Also, the resulting 'artwork' is quite disappointing' - Judge 1.

'Does some cool things and has some complex processes going on but will be more creative when the opinion-forming module is added' - Judge 2.

- Rahman and Manurung's system: *A little creative but not very* (Judge 1), *A little creative but not very* (Judge 2).

  'Seems to me more as an approach to measure the quality of poems in comparison to an existing one' - Judge 1.

  'Despite not reproducing the target poem, it is writing some interesting poetry (although I liked the poem generated by the linear combination method (Manurung 2003) too). Having said that, the system doesn't do what it is designed to do - produce limericks (with associated demands of metre etc)' - Judge 2.

- Norton et al.'s system: *A little creative but not very* (Judge 1), *Quite creative* (Judge 2).

  'The system DARCI is a hybrid system, that draws its ability to create pictures that match with verbal descriptions from the community that provided the "calibration" data on the public website beforehand. So the system as a whole is creative, but the source is located in the human participants' - Judge 1.

  'Interesting to see the 40 images that *DARCI* used to make 'fiery'. When *DARCI* interacts with a human, both *DARCI*'s and the human's ratings increased - herein lies *DARCI*'s true creativity? Good that it doesn't reproduce images but learns from them and renders them slightly differently from the original' - Judge 2.

- Tearse et al.'s system: *A little creative but not very* (Judge 1), *Quite creative* (Judge 2).

  'The manipulation of parameters of the system leads to variations on a nonsense-boredom scale, there is no perspective of optimizing it for more interesting stories, that make sense' - Judge 1.

  'As a standalone system, it creates stories and does interesting things. But at the moment it's just a copy of a previously existing creative system and needs to transcend that if it is to be more creative (otherwise how can it be original?)' - Judge 2.

- Monteith et al.'s system: *Quite creative* (Judge 1), *Quite creative* (Judge 2).

  'There are parallels to the other systems but this one combines several layers of human involvement in an interesting way, so the system as a whole produces a quite creative appearing output' - Judge 1.

  'Exploration of capabilities in different scenarios from the original aims. The concept of

Table 8.3: Ordering the Case Study 2 systems by creativity: Judges' opinions.

	Judge 1	Judge 2
Most creative: 1	Monteith et al. (2011)	Monteith et al. (2011)
2	{ Rahman and Manurung (2011) / { Norton et al. (2011) / { Tearse et al. (2011) (equal)	Norton et al. (2011)
3	”	Cook and Colton (2011)
4	”	Tearse et al. (2011)
Least creative: 5	Cook and Colton (2011)	Rahman and Manurung (2011)

> this is good, as is the two systems working together. Am I influenced by programmers' decisions here? Would like to know more about the music generated. These 2 things stop me using the 'Very creative' option.' - Judge 2.

An overall ranking of systems' creativity can be generated from the data in Table 8.3. For each judge's rankings, 5 points were allocated to the system ranked most creative, down to 1 point for being ranked least creative.[15] The two sets of ranking points were then summed together:

- Monteith et al. (2011) was ranked most creative overall with 10 points (5+5).
- Norton et al. (2011) was ranked second most creative overall with 7 points (3+4).
- Tearse et al. (2011) was ranked third most creative overall with 5 points (3+2).
- Rahman and Manurung (2011) and Cook and Colton (2011) were ranked joint least creative overall with 4 points (3+1 and 1+3 respectively).

The ratings and feedback show some common opinions between the judges. For example, both judges praised the processes involved in Monteith et al. (2011) and both criticised Rahman and Manurung (2011) for their focus on replicating a target poem rather than creating new poetry. Individually, judges' opinions varied a great deal, as is perhaps to be expected with using only two judges who have differing backgrounds and expertise in the various creative domains covered. For example, Judge 2 rated the creativity of both artistic systems more highly than Judge 1 but with less artistic expertise than Judge 1 on which to base these opinions. Judge 2 was also generally more generous in these ratings, giving higher ratings overall. Some of the differences in judges' comments are revealing:

- Comments on Cook and Colton's collage generation system took different foci, showing what each judge's attention was initially drawn to. Whilst Judge 1 was unimpressed by the end product and found the software's creativity trivial, Judge 2 thought more highly of the processes

[15] Where three systems were tied over 2nd, 3rd and 4th place in the rankings, each system was allocated 3 points for 3rd place, the middle ranking of the three.

in the software.

- Both judges reflected on the impact of input from human creativity in DARCI (Norton et al., 2011) but Judge 2 saw this as positively encouraging interactive creativity whereas Judge 1 interpreted this as a weakness of the system's own creativity, probably the more appropriate conclusion to draw given the focus on evaluating the systems' creativity in the case study.
- Judge 2 used their awareness of the original MINSTREL system (Turner, 1994) to criticise the reconstruction of this system in Tearse et al. (2011). Judge 1, not having the same background knowledge of MINSTREL, judged Tearse et al.'s reconstruction on its own merits.

What this has illustrated is that individual people can form very different first impressions of systems. Taking two people's opinions was useful for directed, constructive criticism, but was too small a number of judges for any significant statements about which systems are more or less creative than each other. Although the Case Study 1 survey (Section 8.1.1) suggested a larger number of judges does not necessarily produce conclusive distinctions between systems' perceived creativity, this evaluation for Case Study 2 showed the limits of what can be taken from a small number of judges.

8.2 Using alternative evaluation methods: Computational creativity methodologies

Two other creativity evaluation methodologies were also considered for the systems in the case studies. These were the two methodologies that have been adopted most often in previous computational creativity research: Ritchie (2007) and Colton (2008b).[16]

8.2.1 Applying Ritchie's criteria to evaluate the Case Study 1 systems

Ritchie's evaluation methodology for computational creativity (Ritchie, 2007) applies 18 formally specified criteria. The criteria manipulate ratings of the typicality and value of a system's products.

Artefacts to be evaluated

The artefacts used to evaluate the creativity of the three musical improvisation systems were the nine 30-second extracts used in the online survey collecting human evaluations in Section 8.1.1. As Ritchie (2007) emphasises that creative systems should be evaluated based on its products rather than from any knowledge of the system and how it works,[17] no details were given of any of the systems or their operational details. Tracks were anonymised so that it was not clear from the filename or data what system the track was representing, to remove any potential biases being introduced from prior

[16] As shown by the survey in Chapter 2 Section 2.3.

[17] See Chapter 2 Section 2.1.2.

knowledge of a system.[18]

Obtaining ratings of typicality and value for the systems' output

In Ritchie (2007), these two ratings are defined as:

> 'Typicality: To what extent is the produced item an example of the artefact class in question?'
> 'Quality: To what extent is the produced item a high quality example of its genre?' (Ritchie, 2007, p. 73)

As Chapter 2 Section 2.1.2 described, Ritchie gives no further information on how to collect this information, or how this information should be converted into the *typ* and *val* ratings used by the criteria.[19] To collect information on the typicality and value of the products of the systems in this case study, an online survey was conducted. In this survey, participants listened to the 9 system extracts, presented in randomised order. After hearing each track, participants were asked two questions:

> Typicality: Would you agree that this is a typical example of musical improvisation?
> Value: Would you agree that this is a good musical improvisation?

This is a translation of the two definitions used in Ritchie (2007), with some slight alterations. The questions were posed in a form answerable using a 5-point Likert scale rating how strongly the participant agreed with the questions (see Figure 8.4). Following Ritchie's example, 'value' was substituted for 'quality' to avoid potential confusion as the word 'quality' in this context has connotations with the quality of the sound in the audio file, or the quality of recording. For the questions, 'musical improvisation' was substituted for the definition phrases 'the artefact in question' and 'its genre'. Finally, for simplicity, the 'value' question was slightly rephrased, replacing 'high quality example of musical improvisation' with 'good musical improvisation'.

In answering these two questions, participants were asked to ignore recording/sound quality of the tracks and the computational origins of these tracks as far as possible, instead concentrating solely on the music that was played. They were also asked not to spend too much time thinking about their answers to these questions, but to give their instinctive initial reaction to the questions.

Survey participants 89 participants in total completed the survey. Participants were recruited mainly through mailing lists targeting people interested in computer music, music informatics and other music research. The intention behind this was to recruit different participants to those that completed the survey for Section 8.1.1 and in targeting computer music researchers, hopefully attracting participants who were more tolerant of computer music programs than the general public might be. These mailing lists attract many music researchers so it is unsurprising that the majority of participants (82%) were

[18] Listeners may of course recognise a track when played as being from a particular system, which did happen during this survey, however there is little that can be done to prevent this.

[19] He also appears to use 'value' and 'quality' interchangeably, not defining 'value' separately.

educated to post-graduate level and all but 4 participants had studied at degree level. Demographically, the survey was weighted towards male participants (83.1% of the 89 respondents) with only 15 women responding to the survey. The mean age of participants was 34 years (std. dev. 11.46). The majority of the participants were from the UK[20] and the US, though several other countries in Europe, Australasia, Asia and North, Central and South America were represented amongst the participants.

Participants were asked briefly about their expertise and experience in improvisation, music and computer programming, so that their answers could be considered in context. All but 19 participants claimed some knowledge of computer programming, with 25 participants having reasonable knowledge and 19 having expert knowledge. 52 participants described themselves as computer programmers at either a professional (16), semi-professional (18) or amateur (18) level. Participants had a mean of 13.4 years computer programming experience with standard deviation 6.45 (both to 3 sf).

Musically, 79 of the 89 participants considered themselves a musician, the majority being either amateur (31) or semi-professional (30), with 18 professional musicians. In terms of improvisation, 66 participants described themselves as an improviser, either at amateur (35), semi-professional level (28), with only 3 professional improvisers. Only one participant claimed no musical knowledge at all, with the majority claiming expert (47) or reasonable knowledge (32).

27 people felt they had expert knowledge of improvisation, 28 reasonable knowledge and 25 basic knowledge, with only 9 claiming no knowledge of improvisation at all. All participants' data was considered valid for this experiment, as even if a participant has no knowledge or experience of how to improvise musically, they can understand and interpret what musical improvisation is and can judge the excerpts accordingly. Participants had a mean of 24.2 years musical experience with standard deviation 16.2 (both to 3 sf).

Application of the criteria for evaluation

Full details of how the criteria were calculated can be found in Appendix F. Tables 8.4, 8.5 and 8.6 show the results of evaluation with Ritchie's criteria.

Table 8.4: Results of applying Ritchie's criteria to evaluate GAmprovising.

5 / 18 criteria TRUE (Criteria 1, 3, 10a, 11, 12)
3 / 11 distinct criteria TRUE (Criteria 1, 3, 10a)
11 / 18 criteria FALSE (Criteria 2, 4-7, 13-18)
6 / 11 distinct criteria FALSE (Criteria 2, 4-7, 17)
2 / 18 criteria not applicable (Criteria 8a, 9, both distinct criteria)

[20] A number of the mailing lists targeted were UK-based.

Table 8.5: Results of applying Ritchie's criteria to evaluate GenJam.

13 / 18 criteria TRUE (Criteria 1-5, 10a-17)
7 / 11 distinct criteria TRUE (Criteria 1-5, 10a, 17)
4 / 18 criteria FALSE (Criteria 6, 8a, 9, 18)
3 / 11 distinct criteria FALSE (Criteria 6, 8a, 9)
1 / 18 criteria not applicable (Criterion 7)

Table 8.6: Results of applying Ritchie's criteria to evaluate Voyager.

8 / 18 criteria TRUE (Criteria 1-3, 10a-13, 15)
4 / 11 distinct criteria TRUE (Criteria 1-3, 10a)
8 / 18 criteria FALSE (Criteria 4-6, 8a, 14, 16-18)
5 / 11 distinct criteria FALSE (Criteria 4-6, 8a, 17)
2 / 18 criteria not applicable (Criteria 7, 9, both distinct criteria)

Comparative results with Ritchie's criteria for evaluation in Case Study 1

The choices of inspiring set[21] meant that several of the criteria became equivalent to each other:

Criterion 1 $\equiv$ Criterion 11.

Criterion 2 $\equiv$ Criterion 13 $\equiv$ Criterion 15.

Criterion 3 $\equiv$ Criterion 12.

Criterion 4 $\equiv$ Criterion 14 $\equiv$ Criterion 16.

Criterion 6 $\equiv$ Criterion 18.

It also meant that occasionally criteria became undefinable (criteria 7, 8a and/or 9, depending on the system). Such consequences were not unique to this choice of inspiring sets; other combinations of criteria also became undefinable when using other interpretations of the inspiring set. Undefinable criteria were discarded from analysis. If a criterion was equivalent to a previous criterion, then for this analysis it was also discarded as that aspect was already represented by the previous criterion. This analysis was based solely in the distinctly defined criteria, as opposed to the original set of 18 criteria containing duplicates and undefined criteria.

From the results of applying Ritchie's criteria, as given above, GenJam emerged most creative, as it satisfied the most distinct criteria (3 more than the next best system, Voyager) and falsified the least criteria (two or three less than the other systems). For Voyager, 4 out of 11 distinct criteria were satisfied, 5 distinct criteria were FALSE and 2 were inapplicable. GAmprovising was found to be the

[21] The concept of the inspiring set, the set of existing creative artefacts influencing the design or operation of a creative system, was discussed in Chapter 2 Sections 2.1.2 and 2.1.3. See also Section 8.3.4.

least creative system, with 6 distinct criteria being FALSE, 2 inapplicable and only 3 distinct criteria being satisfied as TRUE. The comparative creativity of the three systems matched the findings of SPECS and of the human survey.[22] The next Section investigates if this finding is replicated with Colton's creative tripod (Colton, 2008b).[23]

8.2.2 Applying Colton's creative tripod to evaluate the Case Study 1 systems

In applying Colton's framework for creativity evaluation (Colton, 2008b), data was required on each system's ability to demonstrate skill, imagination and appreciation, three qualities which were not defined in Colton (2008b) but left open for interpretation. The evaluation data already collected for SPECS could be re-used if there was crossover between the components used for analysis in Case Study 1 and the three tripod qualities of skill, imagination and appreciation. This was the case; Colton's three tripod qualities mapped to three of the 14 components, meaning that the SPECS evaluation data could be re-used for evaluation with the creative tripod:

- Skill $\approx$ *Domain Competence*.
- Imagination $\approx$ *Variety, Divergence and Experimentation*.
- Appreciation $\approx$ *Thinking and Evaluation*.

Other components also matched to parts of the three tripod qualities to some extent, though not as closely as the three components listed above. Skill can be said to include some aspects of *General Intellect* and *Generation of Results* (though skill is generally associated with domain-specific rather than general talents and abilities and is about more than producing output, focusing more on how skilful the system is in performing a particular task). *Intention and Emotional Involvement* in some ways covered the desire and ability to show appreciation, although this was a less solid link between Colton's tripod quality and corresponding Chapter 4 components than the link between appreciation and *Thinking and Evaluation*.

The question arising from making this correspondence between tripod qualities and components was: can the key parts of the evaluation data obtained through the 14 components be obtained by this subset of three components {*Domain Competence*, *Variety, Divergence and Experimentation*, *Thinking and Evaluation*}, or does this subset of the components not capture the full set of information in enough detail? Acknowledging that using fewer components gives less variety of data, the focus was on whether evaluation using this subset of components can be replicated without losing too much important evaluation information. Table 8.7 gives relevant evaluation data from Case Study 1.

[22] Though as Section 8.3 will discuss, Ritchie's criteria give less data about and insight into the creativity of the systems than other evaluation methods.

[23] Looking ahead, Section 8.3 will compare and contrast the different types of evaluation on their results and on what feedback data they generate for the system programmers.

Table 8.7: Evaluation data from Case Study 1 on the components relevant to Colton's three creative tripod qualities. Skill ≡ *Domain Competence*; Imagination ≡ *Variety, Divergence and Experimentation*; Appreciation ≡ *Thinking and Evaluation*.

System	GAmprovising			GenJam			Voyager		
Judge	J2	J4	J5	J4	J6	J1	J3	J5	J1
Domain Competence	3	6	1	6	7	9	7	6	2
Variety, Divergence & Experimentation	4	6	2	4	7	9	3	3	8.5
Thinking & Evaluation	1	5	0	7	6	8	1	1	4

Type of average	Mean				Median			
System	Ga	Gj	V	*All ratings*	Ga	Gj	V	*All ratings*
Domain Competence	3.3	7.3	5.0	*5.3*	3	7	6	*6*
Variety, Divergence & Experimentation	4.0	6.7	4.8	*4.8*	4	7	3	*4*
Thinking & Evaluation	2.0	7.0	2.0	*3.3*	1	7	1	*2.5*

Skill / (Domain Competence)

GenJam outperformed the other systems in terms of skill, with a mean rating of 7.3, a median rating of 7 across all judges, and the highest rating for this component (9/10 from Judge 1). Voyager received a mean of 5.0 and a median of 6. GAmprovising lagged behind the other systems, with average ratings of 3.3 (mean) or 3 (median) and the lowest overall rating (1/10 from Judge 2).

Qualitative feedback for the three systems from the SPECS evaluation interviews[24] highlighted the high competence and domain knowledge of GenJam on one hand and the lack of domain knowledge in GAmprovising on the other. Voyager was commended for being competent within a narrow range of styles but criticised for its ability to gain further theoretical understanding and to play in a coherent style within a higher level structure.

Imagination / (Variety, Divergence and Experimentation)

As for *Domain Competence*, GenJam scored higher ratings overall and attracted the highest individual judge rating (9/10 from Judge 1). GAmprovising had a slightly higher median rating (4) than Voyager (3), but Voyager received a slightly higher mean rating due to Judge 1's rating of 8.5/10 for Voyager's ability to diverge and experiment, compared to the 3/10 ratings from the other two judges. GAmprovising's mean rating was 2/10, the lowest for this component.

In qualitative feedback during SPECS, GAmprovising was criticised for its inhibition and lack of ideas. GenJam was also described as lacking inventiveness and variety in its thought processes, though it could diverge to some limited extent. Voyager demonstrated some variety and was consid-

[24] In using the data obtained via the SPECS evaluation in the creative tripod evaluation, both the quantitative ratings and qualitative feedback are available and both types of evaluation data can be useful.

ered less restricted by scales and formal theory, although its ability to try things out was described as random and limited rather than decisive.

Appreciation / (Thinking and Evaluation)

GenJam again demonstrated considerably higher ratings than the other systems, with its ability in this quality being rated 5 to 6 points higher than the other systems. Whilst GenJam attracted ratings with a mean of 7.0, median of 7 and a range of 6-8 in the rating, the other systems each had a mean rating of 2.0 and median of 1.

GenJam's ability to monitor itself, listen to and think about what is happening and recognise what pitches are appropriate (or not) all contributed to the better data for this system for appreciation. GAmprovising's ability to appreciate what it is doing was 'what the system needs, desperately' (Judge 2, echoed by the other judges). Voyager was given slightly more positive feedback on its ability to listen and react, although this was not seen as being very sophisticated or contributing much to Voyager's overall playing.

Combining all three tripod qualities

Tripod qualities were treated as equally important in Colton (2008b). This is in contrast to the evaluation in Case Study 1, where components were weighted according to their importance for creativity in that domain. In SPECS, *Domain Competence* was given a weight of 12.5 for musical improvisation creativity, compared to the weighting of 7.1 for *Variety, Divergence and Experimentation* and 5.1 for *Thinking and Evaluation*.

The relevant numeric information for the creative tripod qualities is highlighted in Table 8.8.

Table 8.8: Mean ratings out of 10 for the components *Domain Competence*, *Variety, Divergence and Experimentation* and *Thinking and Evaluation*, corresponding to the tripod qualities of skill, imagination and appreciation respectively.

Tripod Quality (Component)	GAmprovising	GenJam	Voyager
Skill (Domain Competence)	3.3	7.3	5.0
Imagination (Variety, Divergence & Experimentation)	4.0	6.7	4.8
Appreciation (Thinking and Evaluation)	2.0	7.0	2.0
Overall mean	3.1	7.0	3.9

GAmprovising GAmprovising performed weakly on all three tripod qualities, being marginally better perceived for its imagination and very poorly received for its appreciative ability. This low performance marked it as the least creative system of the three in Case Study 1; a finding which was replicated when using other evaluation methodologies.

GenJam In contrast to GAmprovising, GenJam demonstrated good performance on all three tripod qualities. As the qualitative feedback and the range of ratings showed, there could still be improvements along all three qualities, in particular its ability to demonstrate imagination. On the whole though, GenJam was found most creative, performing consistently well in all three tripod qualities.

Voyager Voyager demonstrated an unbalanced possession of the three tripod qualities. Although Voyager demonstrated a medium performance for two of the legs, its ability to show appreciation received the lowest mean rating overall (joint with GAmprovising's appreciation abilities), at 2/10. As was shown by the qualitative feedback from judges as well as in the quantitative ratings, Voyager needs to demonstrate more appreciation and self-evaluation of what it is doing. Even if its appreciative abilities are raised to match its abilities to be skilful and imaginative, it would still need to boost its performance in all three tripod qualities to approach GenJam's perceived levels across the tripod.

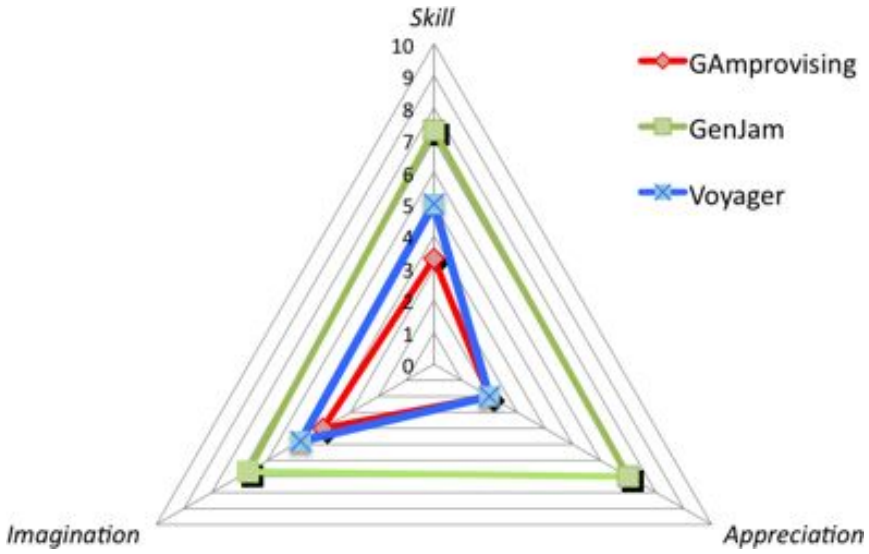


Figure 8.11: Average performance (mean ratings out of 10) on the three creative tripod qualities (skill, imagination, appreciation) for the three musical improvisation systems in Case Study 1: GAmprovising, GenJam and Voyager.

Comparative results with Colton's creative tripod for evaluation in Case Study 1

The quantitative data in Table 8.8 is pictured in Figure 8.11. As Table 8.8, the above comments and Figure 8.11 show, GenJam was considered more advanced in all three tripod qualities, according to the feedback of the judges. GenJam performed equally well in all three tripod qualities, whereas other systems were generally poorer at appreciation than at skill or imagination. In general GAmprovising performed the least well of all three systems and particularly under-performed in terms of skill.

8.2.3 Applying Ritchie's criteria to evaluate the Case Study 2 systems

A comment made in Chapter 7 Section 7.4.6 when discussing the *Originality* and *Value* components was the lack of information in the talks about artefacts produced by the systems. Ritchie's criteria (Ritchie, 2007) concentrate almost exclusively[25] on observations about the output of the system, measuring how typical that output is of the domain and how valuable the output is in the domain. If the authors of the five Case Study 2 systems had all provided examples of their systems (or links to examples) in their papers, then these could be used for evaluation using Ritchie's criteria. Ideally, details of inspiring sets would also be available for Ritchie's criteria to be fully applied. Looking at what was given in the papers:

- Cook and Colton (2011) provides one example of a generated collage (Figure 8.12, as well as some other images that are retrieved through search that could potentially be used for collages. The inspiring set for collage images is essentially the whole Flickr database of images retrievable at the time of running the collage generation process, and the Guardian newspaper website for college themes.
- Rahman and Manurung (2011) give one example of a poem produced by their system, a limerick. The inspiring set item for this particular poem is another limerick on 'relativity', originally published in 1923 and written by Arthur H. R. Buller, repeated in the paper. Both are reproduced later in this Section.
- Norton et al. (2011) give three examples of *DARCI*'s output, supplemented by some images demonstrating the use of evolutionary techniques in *DARCI* and an example population of 40 image renderings. The authors also provide an inspiring set of three images used as sources for *DARCI*'s image generation for this paper. All images are reproduced from Norton et al. (2011) in Figures 8.13(a) and 8.13(b).
- Tearse et al. (2011) gives no examples of stories produced and only brief examples of storylines used as inspiring set items for their story generation system.
- Monteith et al. (2011) provides a link to sound samples of music that represents various emotions and lists titles of five fairy stories that are used for soundtrack generation but does not give the actual wording of the stories used for the experiment.[26] Both the sound samples and fairy stories were presumably the inspiring set for artefacts reported in the paper, but no links are provided to examples of the stories being read along to a soundtrack (although the paper reports how such examples were produced and analysed for evaluation purposes).

[25] Ritchie's criteria also requires details of the *inspiring set*: the artefacts that were used as examples to construct the creative system or as initial material for the system to use.

[26] The system is designed to musically illustrate changes in the emotional content of the story as it is read aloud.

To some extent the variability in available information is understandable; it is easier to replicate images or text in a paper than to replicate music, for example. It may also be that authors were limited by length restrictions in the papers, as the papers were limited to a maximum length of 6 pages. This should not prohibit links being provided to examples if necessary, though.

As it stands, the authors of the five systems in Case Study 2 did not all provide the information necessary to apply Ritchie's criteria for evaluation. Instead we have an incomplete and imbalanced set of output examples; the same was true for the inspiring sets. An initial online search was carried out to find the missing examples but proved fruitless.

The authors could be contacted to see if they would provide output examples from their systems and details of any inspiring examples used to construct the system. It is expected that the authors would select the best examples for this, so for fair comparison, examples should be requested for each system rather than using a mixture of existing examples from papers and newly-provided examples, depending on the system. Assuming that all authors respond with examples, the output examples could then be rated for their typicality and value in their respective creative domains, as was done for Case Study 1 (Section 8.2.1). This procedure would involve considerably more time and research involvement than was permitted for this case study, though, even if the typicality and value for each output example was collected only on first impressions of the example. When using SPECS, the judges were restricted to the information supplied by the paper presenters. Judges could not seek out extra information to supplement this, even if it was considered important for that creative domain. Instead judges relied on what they had learnt in seven minutes and their own background knowledge of the creative domain.[27]

To compare the results obtained using SPECS against results obtained using Ritchie's criteria would be fair only if similar conditions were imposed on the use of Ritchie's methodology. Therefore it is difficult to see how Ritchie's criteria could be applied to all five Case Study 2 systems whilst keeping within the approach of this case study. What can be done is to evaluate the results of the systems as and when presented in the papers, with no evaluations being performed for the Tearse et al. (2011) and Monteith et al. (2011) systems. Whilst this may partly be an exercise in assessing authors' academic rigour in reporting results, it gives a way to apply Ritchie's evaluation criteria in a manner in keeping with that which SPECS has been applied so far, judging creativity based on the information immediately available.

Artefacts to be evaluated

For Tearse et al. (2011) and Monteith et al. (2011), no produced artefacts were provided or referenced in the papers, hence no evaluation shall be done of these systems using Ritchie's criteria.

[27] And in one case, brief previous acquaintance with the system: Judge 2 for Norton et al. (2011).

For Cook and Colton (2011), one example was given of a collage, generated on the theme of the current war in Afghanistan. This is reproduced in Figure 8.12. The inspiring set is a combination of all the images in Flickr[28] and news articles retrievable using the Guardian newspaper API.[29] These sets are of unknown size and given the vast nature of both sets, particularly the Flickr set, the inspiring set can be described as being infinitely large, i.e. with a size of ∞. Although the ICCC'11 paper by Cook and Colton describes *The Painting Fool* module's ability to detect pictures representing emotions of happiness and sadness when constructing collages, the image in Figure 8.12 was produced without use of this ability. As there are no more recent images available to evaluate, the example given in the paper should be used for evaluation, whilst retaining in mind this discrepancy between the content of Cook and Colton (2011) and its given illustrative example.

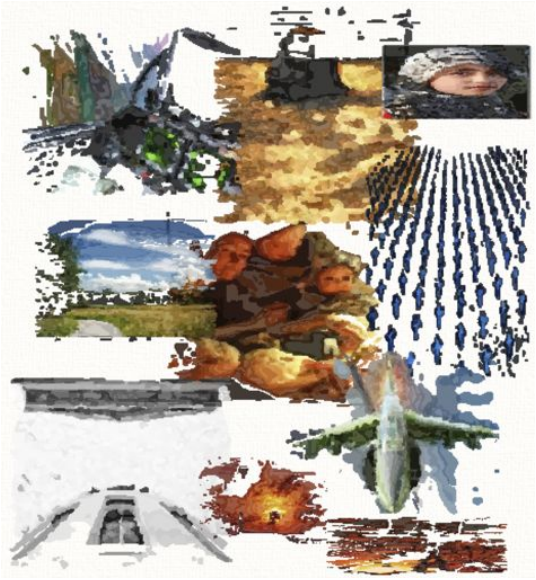


Figure 8.12: Collage generated by Cook & Colton's collage generation module in *The Painting Fool* system, on the theme of the current war in Afghanistan (Cook & Colton, 2011, Fig. 1, p. 2).

For Rahman and Manurung (2011), one example poem (a limerick) was provided (Rahman & Manurung, 2011, in Fig. 1, p. 5):

> 'There **was** the young **la**dy called **Bright**.
> They could **tra**vel much **fas**ter than **light**.
> They **re**latively **left** one **day**. It sur**vived**.
> They **left**. On the **night**, she re**turned**.'
> (Rahman & Manurung, 2011, p. 9)

[28] `http://www.flickr.com`

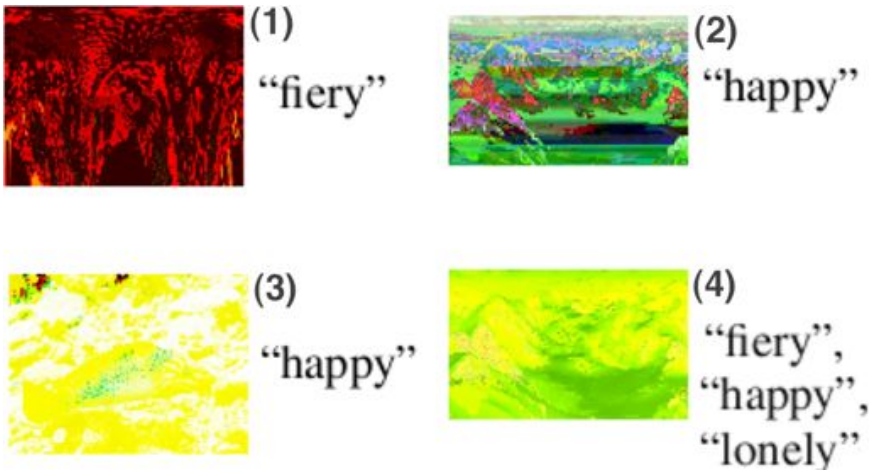
[29] No URL is given for this API, but a search reveals this may be the appropriate URL: `http://www.guardian.co.uk/open-platform`.

Where syllables are in bold, that syllable should be stressed when reading aloud the limerick.

The inspiring set for this system is this limerick, which the paper example aims to replicate:

> 'There **was** a young **la**dy called **Bright**
> who could **tra**vel much **fas**ter than **light**.
> She **set** out one **day**
> in a **rel**ative **way**
> and re**turned** on the **pre**vious **night**.'
> (Rahman & Manurung, 2011, p. 5. Originally published in *Punch* magazine (1923) and written by Arthur H. R. Buller)

For Norton et al. (2011), four images generated by DARCI were provided in the paper, as reproduced in Figure 8.13(a). Figure 8.13(b) reproduces from Norton et al. (2011) a set of images used as the sources for the images in Figure 8.13(a). The two left-most examples in Figure 8.13(a) were rendered from Image A in Figure 8.13(b), the two right-most examples from Image B, with no examples given that were rendered from Image C. The output images were each intended to represent the adjective(s) given to the right of that image.



(a) Output images, each intended to illustrate the adjective(s) listed to the right of that image (Norton et. al., 2011, Figs. 4-7, p. 14).

(b) Inspiring set images (Norton et. al., 2011, Fig. 2, p. 13).

Figure 8.13: *DARCI* output, and the inspiring set of source images used to generate this output.

Obtaining ratings of typicality and value for the systems' output

Ritchie's criteria requires that each artefact is rated in terms of its typicality of the domain and its quality/value (Ritchie, 2007).[30] Three factors influenced how ratings were obtained for the typicality and quality of the artefacts:

- The variability of domains for this case study.
- The desire to capture initial impressions of the artefacts.
- The wish to make this evaluation and other evaluation methods as comparable as possible.

An approach similar to that used in SPECS was adopted to meet these demands, with two judges asked to provide ratings. The approach was implemented in a way consistent with that used in Case Study 1, in Section 8.2.1, assisting comparison between the two case studies.[31] Using this approach:

- Human judges would have at least some acquaintance with the domains represented.
- Judges could give first impressions; neither judge had previously seen the artefacts and had no acquaintance with the systems.
- This approach is very similar to that used in the other methodologies (Chapter 7 Section 7.3.5 and this Chapter, Section 8.2.4).

Each judge was shown the artefacts one by one, and asked:

1. Would you agree that this is a typical example of [the system domain]?
2. Would you agree that this is a high quality example of [the system domain]?

To complete the questions above, text was substituted describing the domain as follows:

- The domain for Cook and Colton (2011) was 'collages illustrating a given theme'.
- For Rahman and Manurung (2011) the intended domain was less clear; the example poem was a limerick, which has quite specific requirements of metre, rhyme and structure. As Rahman and Manurung refer to their system as a poetry generator, though, the example was referred to as being in the domain 'poems'.
- The domain for Norton et al. (2011) was 'images illustrating a given adjective'.

Judges were asked to answer using a Likert scale to represent how much they agreed with each question, selecting an option from the options given in Figure 8.4. Their comments during the process were also recorded. Results are reported in Table 8.9.

[30] As a reminder, Ritchie defines 'typicality' and 'quality' as: 'To what extent is the produced item an example of the artefact class in question?' (typicality), and 'To what extent is the produced item a high quality example of its genre?' (quality) (Ritchie, 2007, p. 73). Chapter 2 Section 2.1.2 noted that the words 'quality' and 'value' seemed to be used interchangeably in Ritchie (2007) and that a separate definition of 'value' was not included.

[31] See Section 8.1.

Table 8.9: Judges' ratings of typicality and quality of artefacts produced by Case Study 2 systems.

	Typicality		Quality	
Artefact	Judge 1	Judge 2	Judge 1	Judge 2
Rahman and Manurung (2011)				
Poem	Disagree	Agree	Disagree	Disagree
DARCI: Norton et al. (2011)				
Image 1	Agree	Agree	Agree	Neutral
Image 2	Strongly agree	Neutral	Neutral	Agree
Image 3	Strongly disagree	Disagree	Strongly Disagree	Disagree
Image 4	Neutral	Strongly disagree	Neutral	Disagree
Collage generation module of *The Painting Fool*: Cook and Colton (2011)				
War collage	Strongly disagree	Agree	Agree	Disagree

Criteria results for each system

Full details of the criteria calculations for each system are given in Appendix F. Here, a summary of results is presented in Tables 8.10, 8.11 and 8.12. Insufficient data was available in Tearse et al. (2011) and Monteith et al. (2011) to apply Ritchie's criteria for evaluation.

Table 8.10: Results of applying Ritchie's criteria to evaluate Cook & Colton's collage generator.

1 / 18 criteria TRUE (Criterion 10a)
7 / 18 criteria FALSE (Criteria 1, 2, 3, 4, 6, 7, 9)
10 / 18 criteria not applicable (Criteria 5, 8a, 11-18)

Table 8.11: Results of applying Ritchie's criteria to evaluate improScreenshotMusicalExpertise's poetry generator.

1 / 18 criteria TRUE (Criterion 10a)
6 / 18 criteria FALSE (Criteria 1, 2, 3, 4, 6, 9)
11 / 18 criteria not applicable (Criteria 5, 7, 8a, 11-18)

Comparative results with Ritchie's criteria for evaluation in Case Study 2

DARCI (Norton et al., 2011) was the only system for which two criteria (5, 10a) rather than one (10a) were true. It also had the fewest inapplicable criteria; the only inapplicable criteria were Criteria 11-18 which, it was noted earlier, could not be applied for any of these systems if the results set did not include items from the inspiring set.

Table 8.12: Results of applying Ritchie's criteria to evaluate Norton et al.'s collage generator.

2 / 18 criteria TRUE (Criteria 5, 10a)
8 / 18 criteria FALSE (Criteria 1, 2, 3, 4, 6, 7, 8a, 9)
8 / 18 criteria not applicable (Criteria 11-18)

The two criteria that *DARCI* satisfied were:

5. A decent proportion of the output are both suitably typical and highly valued.

10a. Much of the output of the system is not in the inspiring set, so is novel to the system.

The two other evaluated systems (Cook & Colton, 2011; Rahman & Manurung, 2011) also satisfied the second of these criteria, 10a.

It is unclear in Ritchie (2007) how the criteria results should be analysed. Is *DARCI* (Norton et al., 2011) the most creative because it satisfied 2 criteria as opposed to 1, or is Rahman and Manurung's poetry generator most creative because it had least false criteria (6 as opposed to 7 for Cook and Colton (2011) and 8 for Norton et al. (2011))? Or should the number of inapplicable criteria be taken into account? It was decided that for this analysis, the percentage of applicable criteria that were true would be calculated for each system and this would be used to compare the systems' creativity. Therefore if a criterion is not applicable to a system, it is removed from the set of criteria for that system. This matches how undefined criteria were treated in Case Study 1 (Section 8.2.1).

- Cook and Colton's collage generator satisfied 1 out of 8 applicable criteria (12.5%).
- Rahman and Manurung's poetry generator satisfied 1 out of 7 applicable criteria (14.3%).
- Norton et al.'s image generator satisfied 2 out of 10 applicable criteria (20%).

These percentages find the *DARCI* image generator by Norton et al. to be the most creative system of the three, followed by Rahman and Manurung's poetry generator and then Cook and Colton's collage generation module for *The Painting Fool* system. For all three systems, though, only a small percentage of criteria were satisfied. In comparison, when applying Ritchie's criteria to the systems in Case Study 1 (Section 8.2.1), between 5 and 13 criteria were satisfied per system.[32]

8.2.4 Applying Colton's creative tripod to evaluate the Case Study 2 systems

In applying the creative tripod (Colton, 2008b) for evaluation, Section 8.2.2 notes how Colton's tripod qualities mapped to three of the 14 components used for SPECS evaluation:

- Skill $\approx$ *Domain Competence*.

[32]This may partly be because judges were given more resources, more freedom and more time to investigate the systems in Case Study 1, whereas the information available for Case Study 2 was generally quite limited.

- Imagination ≈ *Variety, Divergence and Experimentation*.
- Appreciation ≈ *Thinking and Evaluation*.

The evaluation data gathered on these three components could therefore be used to evaluate the ICCC'11 systems using the creative tripod.[33]

Skill / (Domain Competence)

Judges provided data on *Domain Competence* for all systems except that of Cook and Colton. All the other four systems showed a good grasp of domain expertise and skills, especially Rahman and Manurung which received rating of 9 and 8 out of 10. *DARCI* (Norton et al., 2011) was rated 7/10 by Judge 2 (not rated by Judge 1) and the *MINSTREL*-based storyteller (Tearse et al., 2011) received a ranking of 8/10 from Judge 1 (but was left unrated by Judge 2). Monteith et al.'s music soundtrack generator was seen as possessing only average *Domain Competence* by Judge 1, who gave it 5/10, however Judge 2 saw it as equal to *DARCI* for *Domain Competence*.

Imagination / (Variety, Divergence and Experimentation)

All five systems were rated by at least one judge. In general the *Variety, Divergence and Experimentation* demonstrated by systems was considered to be average across all systems, with no system standing out, with the exception of *DARCI* (Norton et al., 2011); although *DARCI* received 5/10 from Judge 2, Judge 1 gave this system the maximum rating of 10/10.

Appreciation / (Thinking and Evaluation)

Only three of the five systems were rated by judges for *Thinking and Evaluation*. It was felt by judges that they were not given enough information about this component for *The Painting Fool*'s collage generation module (Cook & Colton, 2011) and the music soundtrack generator (Monteith et al., 2011) to form an opinion on the systems' abilities on this component. The poetry generator system (Rahman & Manurung, 2011) was rated highly by both judges (9 and 7), and received the highest ratings on this component overall. Whilst Judge 2 thought the systems by Norton et al. and Tearse et al. performed as well on this component as the poetry generator, Judge 1 gave *DARCI* (Norton et al., 2011) a rating of 5, as opposed to the rating of 9 given to the poetry generator, and did not give a rating for the *MINSTREL* reconstruction (Tearse et al., 2011).

Combining all three tripod qualities

The relevant results for the creative tripod qualities are highlighted in Table 8.13.

Cook & Colton's collage generator Although Colton proposed the creative tripod evaluation framework, in a paper that is only a few years old (Colton, 2008b), Colton's system presentation at ICCC'11

[33] Using the SPECS data this way will help to see if this subset of the 14 components covers the full set of components adequately to give us the same level of information as the full set, or if there is important evaluative information missing.

Table 8.13: Ratings out of 10 (in the format: Judge 1's rating / Judge 2's rating) for the components *Domain Competence*, *Variety, Divergence and Experimentation* and *Thinking and Evaluation*, corresponding to the tripod qualities of skill, imagination and appreciation respectively.

Tripod Quality (Component)	Cook & Colton	Rahman & Manurung	Norton et al.	Tearse et al.	Monteith et al.
Skill (Domain Competence)	- / -	9 / 8	- / 7	**8 / -**	5 / 7
Imagination (Variety, Divergence & Experimentation)	5 / -	5 / -	**10 / 5**	5 / 6	- / 5
Appreciation (Thinking and Evaluation)	- / -	9 / 7	5 / 7	- / 7	- / -
Overall mean of all rated qualities	n/a	7.2	6.8	6.8	n/a

did not report enough information for the judges to feel capable of rating its skill and appreciation (or *Domain Competence* and *Thinking and Evaluation*) abilities.[34] Although the system was deemed by Judge 1 to demonstrate average imaginative ability (*Variety, Divergence and Experimentation*), Judge 2 felt unable to form an opinion on this component either.

Rahman & Manurung's poetry generator The poetry generator was judged skilful by both judges, as was its ability to appreciate what it was doing (*Domain Competence* and *Thinking and Evaluation* respectively). Its capacity for imagination was considered average by one judge, although the other judge did not feel able to comment on this (*Variety, Divergence and Experimentation*). Generally though, Rahman and Manurung's system performed well on the creative tripod qualities, with appropriate information available to evaluate all three tripod qualities at least to some extent.

Norton et al.'s image generator DARCI All the ratings received by judges for *DARCI* were 5 or above, and one judge considered its imaginative abilities to be worth a maximum rating of 10. In general, *DARCI* gave a reasonably competent performance on all three components and there was enough information given to judge its performance on all tripod qualities (with only one piece of missing information from the judges: Judge 1's rating of *DARCI*'s skill).

Tearse et al.'s reconstruction of the MINSTREL *story-telling system* Enough information was available in the presentation of this system for it to be rated by at least one judge on each component. The story teller received middling ratings for imagination from both judges and was rated above average for its skill and appreciation, though by only one judge in each case.

Monteith et al.'s music soundtrack generator For both judges, there was insufficient information about this system's ability to show appreciation and Judge 1 also felt they could not rate the system's imagination, though Judge 2 rated the system's imagination as average (5 out of 10). This judge gave

[34] During the ICCC'11 presentation, Colton concentrated on the intentionality being shown in the collage generator, spending time on this aspect rather than giving information on others. It is unclear as to how this priority fits into Colton's creative tripod framework, but it would be interesting to see how this could be incorporated into the creative tripod.

a slightly higher rating for skill (7 out of 10), alongside the rating of 5 out of 10 from Judge 1 for skill. In general the soundtrack generator can be described as showing average to good skill and average imagination, but that no conclusions can be drawn about the system's ability to appreciate its work.

Comparative results with Colton's creative tripod for evaluation in Case Study 2

Looking at the discussions above and the data in Table 8.13, three systems emerge as 'balanced', i.e. with all three 'legs' present (Rahman & Manurung, 2011; Norton et al., 2011; Tearse et al., 2011). Monteith et al.'s system could not be evaluated on its appreciative abilities and Cook and Colton's system presentation lacked information on both its skill and appreciation. Both systems received only middling ratings in all cases where ratings were provided, apart from a 7/10 for Monteith et al.'s skill from one judge (accompanied by a 5/10 from the other judge).

Of the three systems with data on each of the tripod qualities, mean ratings were calculated for each quality (ignoring unspecified ratings):

- Rahman and Manurung (2011) - Skill: 8.5, Imagination: 5, Appreciation: 8.
- Norton et al. (2011) - Skill: 7, Imagination: 7.5, Appreciation: 6.
- Tearse et al. (2011) - Skill: 8, Imagination: 5.5, Appreciation: 7.

Figure 8.14 compares these results graphically. Unlike for Case Study 1 (Section 8.2.2) all three systems averaged at least 5 out of 10 in all cases, with means of 7.5 or above for the skill and imagination in the poetry generator, the appreciation of *DARCI* and the skill of the *MINSTREL* reconstruction. Taking the mean of the three qualities for each system, overall averages were 7.2 for Rahman and Manurung (2011) and 6.8 each for Norton et al. (2011) and Tearse et al. (2011). These observations indicate that Rahman and Manurung's system was found to be more creative than the other systems, as it had a higher mean overall and the highest ratings for two out of three qualities. Its performance for appreciation, though, was only average (5/10). It could be argued that *DARCI* demonstrated a better all-round performance and was therefore found to be more creative. Given that Colton (2008b) does not investigate how to use the creative components for quantitative comparison, exact conclusions are open to interpretation. No such usage of the creative tripod was found to use as an example, so the decision taken here was to highlight Rahman and Manurung's poetry generator as the most creative system, but to note its weakness in imagination abilities, and also to acknowledge the all-round abilities of *DARCI* (Norton et al., 2011) and the similar performance of the *MINSTREL* reconstruction (Tearse et al., 2011) to the poetry generator.

The other two systems (Cook & Colton, 2011; Monteith et al., 2011) were considered less creative overall than these three systems, because they did not demonstrate clear abilities on some of the tripod qualities (given the information in the presentations on the systems). Of these two, Monteith et al.'s system may have been slightly superior to that of Cook and Colton because it demonstrated

some aspects of both skill and imagination and received one rating of 7/10 (from Judge 2 for skill) in comparison to the rest of the ratings for these two systems (either left blank or rated as 5/10).

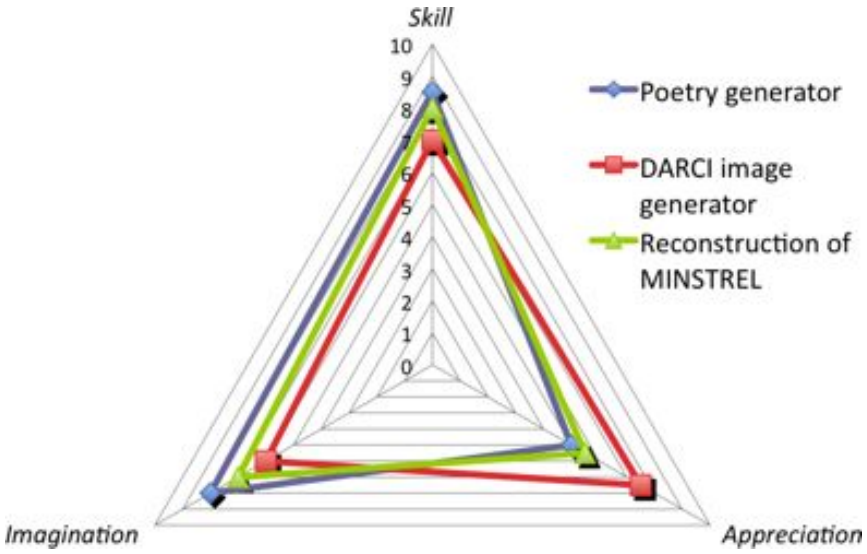


Figure 8.14: Average performance on the three creative tripod qualities (skill, imagination, appreciation) by Rahman & Manurung's poetry generator, Norton et al.'s image generator *DARCI* and Tearse et al.'s reconstruction of the story-telling system *MINSTREL*, as rated by two judges after hearing 7-minute talks presenting each system.

A key part of Colton's framework is in identifying strengths and weaknesses of the systems in the three tripod qualities as opposed to making measured comparisons. Using the creative tripod showed:

- Although Rahman and Manurung's system is skilful and imaginative, Rahman and Manurung need to work on its ability to appreciate what it is doing in order to be creative.
- Similar conclusions can be drawn by Tearse et al., whilst noting that the skill and imagination of the *MINSTREL* reconstruction is good but not at the level of the poetry generator.
- *DARCI* is perceived to be slightly weak (relative to the other qualities) at showing imagination. Overall all three tripod qualities are reasonably well demonstrated by *DARCI* but that improvements can be made.
- Monteith et al. can make improvements to the skill and imagination of their system, seen as only average to fairly good, whilst giving more information on appreciation in their system.
- Cook and Colton's main feedback from the application of the creative tripod is that to better promote their system's creativity, the information they provide should contain more emphasis on the three tripod qualities; this was highlighted as key for creativity by one of the co-authors of this paper three years previously (Colton, 2008b).

8.3 Comparing the success of different evaluation methods in the two case studies

Four different evaluation methods have been now used to evaluate the creativity of systems in each of the case studies:

1. SPECS (using the Chapter 4 components as a basic definition of creativity).
2. Human opinions of creativity (the entire concept, as opposed to creativity when broken down into contributory components).
3. Ritchie's empirical criteria.
4. Colton's creative tripod.

Each generated evaluative information on each system and comparisons of creativity between systems within each case study.[35] It is difficult to arrive at a conclusion of which evaluation method is most accurate, due to the lack of any 'correct answer' or baseline to compare against (as has been remarked upon several times in this thesis). What can now be done, though, is to compare evaluative results obtained through different methods for uniformity and discuss how useful each of these evaluations are to the computational creativity researcher, in terms of gathering effective feedback to assist further development and identify contributions to knowledge made by each system.

8.3.1 Comparing evaluation results and feedback in Case Study 1

To recap on the results obtained through each of the evaluation methods used, individual feedback on each system is summarised below and Table 8.14 contrasts overall comparisons.

1. SPECS using the Chapter 4 components:
    - GAmprovising was good at creating results, developing its results and progressing as a system. It lacked abilities to interact and communicate socially and to carry out reasoned self-evaluation.
    - GenJam's interactive abilities were highly praised and it was seen as productive in creating results, although it could be more spontaneous and improve its originality.
    - Voyager's interactivity and communicative abilities were praised and the value in the system was perceived as high. It was criticised for its weaknesses in developing what it does and progressing over time, and also for its lack of self-evaluation or the ability to think through its processes.
2. Survey of human opinions:
    - GAmprovising was considered 'quite creative' overall with a quantitative mean rating

[35] It is not appropriate to compare systems across the two case studies, due to the differences in approach and levels of detailed knowledge used.

of 2.5 / 5.0 and 193 ranking points collected from overall comparisons. Its music divided opinion and aesthetic preferences, particularly due to its more free and avant-garde style. Whilst receiving praise from some for its unconventionality, GAmprovising attracted heavy criticism for its randomness and lack of development of coherent structure.

- GenJam was considered 'quite creative' overall with a quantitative mean rating of 3.0 / 5.0 and 250 ranking points collected from overall comparisons. It was considered reasonably competent although perhaps unadventurous and a little boring after extended listening. The interaction with the trumpet player was highlighted for praise.
- Voyager was considered 'quite creative' overall with a quantitative mean rating of 2.7 / 5.0 and 211 ranking points collected from overall comparisons. Many enjoyed Voyager's interaction with the human player and found the music convincing as avant-garde-style improvisation, although this style of music was not to everyone's tastes. It was often seen as an accompanist to the human player rather than a co-performer in the improvisation, though, and was criticised somewhat for its randomness and occasional lack of engagement with its co-performer.

3. Ritchie's empirical criteria framework:
   - For GAmprovising, of 11 distinct criteria, 3 evaluated as TRUE (Criteria 1, 3, 10a), 6 as FALSE (Criteria 2, 4-7, 17) and 2 were not applicable (Criteria 8a, 9).
   - For GenJam, of 11 distinct criteria, 7 evaluated as TRUE (Criteria 1-5, 10a, 17), 3 as FALSE (Criteria 6, 8a, 9) and 1 was not applicable (Criterion 7).
   - For Voyager, of 11 distinct criteria, 4 evaluated as TRUE (Criteria 1-3, 10a), 5 as FALSE (Criteria 4-6, 8a, 17) and 2 were not applicable (Criteria 7, 9).
4. Colton's creative tripod framework:
   - GAmprovising was poor in all three tripod qualities both in ratings and in qualitative feedback, with mean ratings out of 10 of 3.3 for skill, 4.0 for imagination and 2.0 for appreciation.
   - GenJam demonstrated good all-round ability in all three tripod qualities, with mean ratings out of 10 of 7.3 for skill, 6.7 for imagination and 7.0 for appreciation. Qualitative feedback marked GenJam as being consistently good, if not excellent, in all three tripod qualities, but GenJam could improve in all three qualities (particularly its imaginative capacity).
   - Voyager also had an imbalanced possession of the three tripod qualities, with a mean rating out of 10 of 5.0 for skill, 4.8 for imagination and a lower rating of 2.0 for appreciation. Voyager's weakness in appreciative abilities was noted in the qualitative feedback from

judges as well as in the quantitative ratings. In the other two tripod qualities Voyager's performance was average.

Table 8.14: Overall comparisons of the relative creativity of each system in Case Study 1 from 1: Most creative to 4: Least Creative. Results as reported in: the use of SPECS and components, surveys of human opinions, Ritchie's empirical criteria framework, and Colton's creative tripod framework.

	Evaluation Method			
Ranks	SPECS using components	Survey of opinions	Ritchie's criteria	Colton's tripod
1	GenJam	GenJam	GenJam	GenJam
2	Voyager	Voyager	Voyager	Voyager
3	GAmprovising	GAmprovising	GAmprovising	GAmprovising

Table 8.14 shows how most evaluation methods agree on rankings, at this shallow level of detail. GenJam was always found to be the most creative of the three systems, followed by Voyager, with GAmprovising the least creative system. It must be noted that Table 8.14 is somewhat contrived and overly summative, hiding much evaluative detail of potential interest. The table summarises the information gathered but it is acknowledged that the evaluative findings within methodologies were not always consistent, such as Judge 4's overall rating of GAmprovising as higher than GenJam. Table 8.14 is therefore orientated towards brief summative comparison, not final results.

Where the methodologies significantly differed was the amount and type of feedback obtained for evaluation. There were also issues about the ease of applicability of each methodology, although it is noted here that all four methods involved a similar-length process of data collection (relative to the case study requirements).

SPECS with the Chapter 4 components obtained detailed feedback about the systems' strengths and weaknesses. In particular, the feedback was tuned towards musical improvisational creativity and what is important in this domain. This gave the system authors information as to what to focus attention on in creativity improvements, for most effective results. Judges for this evaluation method mostly found the components applicable to evaluation of the systems, with only a few exceptions for some judges in the more human-like components. The component-based ratings collectively matched judges' overall opinions quite closely.

When consulting people's opinion on how creative each system was, constructive qualitative feedback was collected from people's comments as well as quantitative feedback in the form of ratings. Participants often reported difficulties in rating the systems' creativity without a definition of creativity to use, but generally felt quite confident about their ratings.

Ritchie's empirical criteria was the only evaluation methodology to make no use of qualitative feedback, relying purely on quantitative ratings of typicality and value of the products of systems. Whilst this methodology is the most formally specified approach, there were still a number of parameter values left unspecified, for example for thresholds. Decisions about how to set these parameters could affect the results of applying the criteria. Another issue with Ritchie's criteria was that although it was clear to see which criteria have been satisfied or not, the criteria have been specified at a level of abstraction away from the actual systems, meaning that they needed to be recast in the context of the systems before the feedback could be useful to the system authors.

Colton's creative tripod framework used the same basic evaluation data as was collected for SPECS, but interpreted it in the same way. To a certain extent this showed that the subset of the three components *Domain Competence*, *Variety, Divergence and Experimentation* and *Thinking and Evaluation* could together collect a similar result (in this case study) as the full set of 14 components. What was missing from the data from the subset of three components was information noted as highly important for creativity in musical improvisation, such as the systems' ability to interact socially and communicate. This ability was noted as a distinguishing feature of more creative systems in the feedback from several participants, across the different methods of evaluation.

8.3.2 Comparing evaluation results and feedback in Case Study 2

To recap on the results obtained through each of the evaluation methods used for Case Study 2, individual feedback on each system is summarised below and Table 8.14 contrasts overall comparisons.

1. SPECS using the Chapter 4 components:
    - The collage generator (Cook & Colton, 2011) performed well at creating results, demonstration of intention and social abilities, but could improve its originality, value and ability to be spontaneous.
    - The poetry generator (Rahman & Manurung, 2011) was good at creating results in a domain-competent way but needs to improve its ability to experiment and diverge and to a lesser extent could improve upon its originality, value and spontaneity.
    - *DARCI* (Norton et al., 2011) showed strengths in social interaction, spontaneity, self-evaluation and production of results, but could perform better on originality and value.
    - The story generator's (Tearse et al., 2011) abilities to be original and to produce results were praised, though it could improve upon its inherent value, its ability to progress and develop and to work independently.
    - The soundtrack generator (Monteith et al., 2011) was considered valuable and competent in its domain, but could improve its ability to diverge and experiment.

2. Survey of human opinions:
    - The collage generator: 'Not at all creative/Quite creative'. The complexity of the processes used was praised by one judge but seen as trivial for creativity by the other.
    - The poetry generator: 'A little creative but not very/A little creative but not very'. It generated interesting poetry but did not generate what was intended and was more aimed at generating a target example.
    - *DARCI*: 'A little creative but not very/Quite creative'. *DARCI*'s ability to learn was highlighted as a useful attribute by one judge. The system may be more useful for interactive creativity with humans than as a standalone system, though, with one judge seeing the creativity of DARCI as located within the knowledge obtained from people.
    - The story generator: 'A little creative but not very/Quite creative'. Whilst creating stories that seem to be fairly interesting but slightly nonsensical, the process did not seem to be optimised for increasing the interestingness of its stories, but for replicating as closely as possible a previous system (MINSTREL).
    - The soundtrack generator: 'Quite creative/Quite creative'. Judges liked the intentions behind the system and its ability to combine two existing systems and layer human involvement, but needed more information for a fuller opinion.
3. Ritchie's empirical criteria framework:
    - For the collage generator, of 8 applicable criteria, 1 evaluated as TRUE (Criterion 10a) and 7 as FALSE (Criteria 1-4, 6, 7, 9).
    - For the poetry generator, of 7 applicable criteria, 1 evaluated as TRUE (Criterion 10a) and 6 as FALSE (Criteria 1-4, 6, 9).
    - For *DARCI*, of 10 applicable criteria, 2 evaluated as TRUE (Criteria 5, 10a) and 8 as FALSE (Criteria 1-4, 6-9).
    - Neither the story generator or the soundtrack generator could be evaluated due to lack of information on their respective inspiring sets.
4. Colton's creative tripod framework:
    - The collage generator showed average imaginative abilities and there was a lack of information on other qualities, with mean ratings out of 10 of 5.0 for imagination but no data for skill or appreciation.
    - The poetry generator demonstrated very good skilfulness and appreciation with average imagination, with mean ratings out of 10 of 8.5 for skill, 5.0 for imagination and 8.0 for appreciation.
    - *DARCI* showed average to good all-round performance on the tripod qualities, with mean

ratings out of 10 of 7.0 for skill, 7.5 for imagination and 6.0 for appreciation.

- The story generator performed reasonably well on all three tripod qualities although could improve its imaginative abilities, with mean ratings out of 10 of 8.0 for skill, 5.5 for imagination and 7.0 for appreciation.
- The soundtrack generator gave partial information on the tripod qualities, demonstrating average skill and imagination, with mean ratings out of 10 of 6.0 for skill, 5.0 for imagination but no data for appreciation.

Table 8.15: Overall comparisons of the relative creativity of each system in Case Study 2 from 1: Most creative to 5: Least Creative. Results as reported in: the use of SPECS and components, surveys of human opinions, Ritchie's empirical criteria framework, Colton's creative tripod framework. NB 'gen' is an abbreviation of 'generator'.

Evaluation Method	SPECS using components	Survey of opinions	Ritchie's criteria	Colton's tripod
1	*DARCI*	soundtrack gen	*DARCI*	poetry gen
2	story gen	*DARCI*	poetry gen	*DARCI* / story gen
3	poetry gen / soundtrack gen	story gen	collage gen	*DARCI* / story gen
4	poetry gen / soundtrack gen	poetry gen / collage gen	- (other two systems unrated)	soundtrack gen
5	collage gen	poetry gen / collage gen	- (other two systems unrated)	collage gen

Some of the observations about the applicability of the different methods from Case Study 1 (Section 8.3.1) also apply here:

- The SPECS methodology gave feedback most finely tuned towards specific requirements for creativity in each domain.
- Judges who volunteered opinions on the systems' creativity noted the difficulty in labelling each system's creativity but were able to perform this task with some confidence.
- Ritchie's criteria again disregarded qualitative feedback and gave results that were summative but a level of abstraction removed from the actual systems.
- Colton's framework, using a subset of the data collected for SPECS, collated some findings but again missed out on areas noted as being highly important for creativity.

The difference between Case Studies 1 and 2 was in the overall findings; for Case Study 2, no two methodologies produced the same results. Typically, *DARCI* (Norton et al., 2011) was judged one of

the more creative systems and the collage generation module for *The Painting Fool* was judged one of the less creative systems. Otherwise, there was considerable disagreement between findings. The lack of a 'ground truth' or baseline to judge the systems by was exacerbated by the different domains that the systems work in, meaning that comparison between these five systems was of limited value. The priority in this case study was therefore formative feedback for the system authors rather than comparisons of creativity between systems.

The consequences of judging a system given limited and perhaps incomplete information meant that occasionally important information for evaluation is missing. This affected the use of all the evaluation strategies employed. It is interesting to see which methodologies were most robust when dealing with missing information.[36] Human opinions seemed best at coping with missing information, as might be expected given that little was specified for the humans to use as a definition of creativity. SPECS was relatively robust, as was Colton's tripod framework. Ritchie's criteria approach was the most affected by missing information, as various criteria could not be applied and the absence of information on inspiring sets and example outputs had significant effects.

8.3.3 Reflections on using human opinion surveys on creativity of systems

Surveys of human opinion provided more useful summative reflections on the relative creativity of each system in Case Study 1, compared to Case Study 2, because of the number of judges involved. Useful evaluative feedback was generated in both case studies by consulting human opinion. More consistency emerged overall between judges in Case Study 1 due to averaging over more judges, but a general consensus of opinion still did not clearly arise, with all systems overall being considered 'quite creative' and with contradictions and discrepancies in overall opinions on rankings.

Several relevant findings arose from the surveys of human opinion[37] that were not directly connected with judging systems' creativity, but that related to the wider aspects of performing this task. Advertising the survey resulted in a number of discussions being struck up on social networking pages and online forums, where people discussed what constituted computational creativity and gave their opinions about computational creativity. Some quotes from these more general discussions are included with the survey comments where appropriate in this Section.

Troubles evaluating creativity: The need for a clearer definition Several people questioned how creativity should be defined in the context of these systems or requested a definition of 'creativity' to be supplied in the survey, rather than relying on their own understanding. This was a slightly

[36] This is particularly relevant when evaluating other researchers' systems for comparison, or to learn from what other researchers have achieved, as it may not be possible to retrieve desired information that is not available through publications or online resources about the system. More detailed discussion of resource availability issues was seen in Chapter 7 Section 7.1 and will be returned to in Chapter 9 Sections 9.1.1 and 9.1.3.

[37] The surveys in Sections 8.1.1 and 8.1.2.

surprising but significant finding of the survey.[38]

> 'My biggest problem was in trying to use the term "creative" as a quantifiable concept.'
>
> 'I do think that there should be a definition of "creativity" right up front!'
>
> 'Perhaps a brief description of what it takes for a computer to be creative?'
>
> 'All depends on your definition!'
>
> 'This raises the question "What is creative?" '
>
> 'I think it depends on the definition of creativity - is it just creating something? or creating something that makes the appropriate amount of "sense", for want of a better word, for people to appreciate? I'm using the latter definition!'
>
> 'Creative quite an ambiguous term'
>
> 'this [survey] is a classic example of how many different interpretations there can be of the word "creative".'

Rating on what they liked, rather than what was creative One particular definitional problem during the survey was summed up by these two participants:

> [On the system removed from Case Study 1] 'I liked this one better than the other ones, but am really struggling to distinguish between "like" or "approve" and "think it's creative".'
>
> 'I kept going with my gut instinct which was basically to rate it on how much I *liked* it... but I don't think that really equates to how creative it was... but I'm not even sure a computer *can* be creative, which is why I had to just keep reverting to "like".'

The second participant went on to explain how they struggled with applying the concept of creativity to computers when they saw creativity as 'a uniquely human thing.' Thus they had fundamental problems with evaluating the computational creativity systems, instead resorting to a conceptually easier measure of aesthetic, even though they were aware of the difference:

> [On GAmprovising] 'I preferred these samples to the previous ones, but that isn't really a measure of creativity!'

This reaction to computational creativity[39] introduced bias to human evaluation of the creativity of computational systems, whether conscious, as for this participant, or subconscious, as found in the study by Moffat and Kelly (2006). Significant problems can result for evaluation of computational creativity, unless participants were given more tangible and less controversial metrics to evaluate, as is the intent with using the components derived in Chapter 4.

Matching computational creativity to human standards Many people dealt with the issue of computational creativity by comparing the systems' performance to human standards. This was done in both a positive and negative way. The positive comments took one main form of comparing the system to a good human player, as some representative quotes demonstrate:

[38] Some extra confusion was possibly introduced by asking participants to evaluate computational creativity rather than human creativity, generally the more familiar manifestation of creativity.

[39] Also discussed in Chapter 1 Section 1.4.

> [On Voyager] 'This feels the most "human" in terms of creativity to me.
>
> [On the system removed from Case Study 1] ' I thought that the flute improv was brilliant. I'd have no idea that this was generated by machine unless I was told.'
>
> [On the system removed from Case Study 1] 'All three excerpts had a sense of musical cohesion such that it is not difficult to imagine the composition being performed live by a person.'

Negative comments were disbelieving of the ability of systems to match a human level of competence, or likened the system to a poor or novice human musician:

> 'Probability calculations do not a jazz-cat make!'
>
> 'I'm doubtful about whether a computer will ever be able to match the creativity of a human'
>
> 'Good luck in trying to programme musical creativity it takes years of dedication for humans to realise their creative potential I suspect it would take a series of programmers across many generations to catch up.'
>
> [On the system removed from Case Study 1] 'It is following rules like most beginner jazzers. It's "knowing" when to break the rules that make jazz interesting.'
>
> [On the system removed from Case Study 1] 'This system seems to be creative in a way that a lot of "workaday" performing musicians are - they have spent a long time studying a particular style of music and have learned to abstract some ideas from that about how to play in that style.'
>
> 'they may be creative but they are not producing good music. they sound like a clueless beginner who has got lost and is following a chord sequence slavishly until they get a good idea.'

Intention and Emotional Involvement One criticism of systems made repeatedly by participants, across all systems, was that there was no demonstration of the component *Intention and Emotional Involvement* from Chapter 4. This was seen as a major flaw. As the following variety of quotes show, participants were critical of the systems' abilities to demonstrate personality, be emotionally involved or express feelings, passion, warmth, enjoyment or soul:

> [On GenJam] 'Overall the ideas were strong, just lacked expression and personality. This is an important component of jazz improvisation.'
>
> 'Part of enjoying music is about feeling that you are sharing something with its creator. For me, at any rate. And I don't think there's any doubt that our ability to enjoy and connect with music is intimately connected with emotion, even if that music has been created without emotion. ... Maybe, because I am an emotive being, I just *want* it [creativity] to be about emotion.'
>
> 'it's nice to think that there's a little bit of the composer's soul in there..... but then maybe the soul is just part of the illusion of self and we are just computers that think they're people. ;) '
>
> 'It requires a significant amount of human intervention for any music created on a computer to have feeling, depth, warmth, emotion, or any of those other things that give music valency.'
>
> 'What is missing in all examples so far is passion. Music is about something to say. These programs make sounds a bit like music but they have nothing to say. Like a baby babbling. Like the cargo cult. They so miss the point that it is painful.'
>
> 'even though i enjoyed [Voyager] i wouldn't call it creative as it lacks soul and true idea.'
>
> [On GenJam] 'Sounds clever but cold. ... There's a lack of magic and spontaneity'
>
> 'some of the music I listened to in that survey reminded me of the way you can sometimes find a Jackson Pollack [sic] on a dust sheet at the end of an extensive interior decorating sesh.

> Purely random, but nice to look at. Was it the result of a creative process. Not really. Or not intentionally.'
>
> 'When computers can enjoy jazz is the time when they will be able to play it. ... Have you got the computers to answer this survey to see if they think they know what they are doing?'

Randomness and creativity Some systems, notably GAmprovising and Voyager, were heavily criticised for sounding random. This was seen as negative almost universally across the survey, implying a lack of decision-making and consideration in the construction of improvisations. Though perhaps not a fair representation of Voyager's complex decision-making processes (Lewis, 2000), this impression was accurate for GAmprovising, which was originally designed as a proof-of-concept model testing how random generation can be used as a heuristic for computational creativity.[40]

In discussions on an online forum prompted by this evaluation survey, discussing the feasibility of a computer system being creative, one response was:

> 'if it [the computer system] is based on random number generation then maybe no.'

It was observed during this discussion that 'Most jazz players just play a bunch of random notes anyway :-)'. Another participant mentioned during the survey:

> 'I'm struggling to distinguish between "creative" and "random". I'm not sure creativity is an attribute you can ascribe to a machine? Surely it's only *aping* creativity?'

In his discussion of randomness, Cope (2005) discredits its usage as a technique for simulating creativity. Though he acknowledges the link between creativity and unexpectedness, which can be simulated using random search, he claims that randomness is a hindrance to creativity models rather than an aid. One aspect of randomness that Cope fails to recognise, though, is that an undirected search of possibilities helps new options to be discovered that may otherwise have been missed, especially in larger search-spaces such as the 'conceptual spaces' advocated by Boden (2004).

If random generation mimics spontaneous human ideas and divergent, unconstrained thinking, as GAmprovising was intended to explore, then randomness (or more accurately, pseudo-randomness, as complete randomness can only be simulated in deterministic systems like computers) becomes an acceptable heuristic for computational creativity. From the findings of this survey, though, survey participants did not perceive random generation to be an adequate computational replacement for human creative inspiration and spontaneity.

Practical issues in evaluating the musical improvisers Participants were asked to evaluate the computer system rather than the whole performance, focusing attention on the system's performance in the context of the music being produced. Participants were also asked to judge the music not in terms of its sound quality or computational origins, but on its creativity; if a system used MIDI sounds then

[40] See Chapter 6 Section 6.2 for further discussion.

ideally this should not have influenced their judgement unduly. Some participants understandably found these instructions difficult to follow, acknowledging this in their responses.

A number of participants commented that the tracks were not long enough to judge, or that they would have liked to hear the full track to hear the soloing in context. A decision had been taken to limit the lengths of tracks to 30 seconds, partly due to different lengths in the systems' tracks[41] and partly to make sure the survey was not over-long for participants. In hindsight, perhaps it would have been appropriate to provide links to full track(s) for people to listen to if desired, with a warning that this would increase the length of time the survey took.

8.3.4 Reflections on using Ritchie's criteria

In applying Ritchie's criteria for evaluation, there were issues over how the inspiring set I should be defined.[42] In Ritchie (2007) the inspiring set is defined as a subset of all the possible basic items that could be produced by the system (Ritchie, 2007, p. 77), although more specific definition is vague:

> 'The construction of the program is influenced (either explicitly or implicitly) by some subset of the available basic items. This subset, which we will call the inspiring set, could be all the relevant artefacts known to the program designer, or items which the program is designed to replicate, or a knowledge base of known examples which drives the computation within the program.' (Ritchie, 2007, p. 76)

As mentioned in Section 8.2.3, inspiring set information was often not available for systems in Case Study 2. Various interpretations exist of the inspiring sets I for systems in Case Study 1:[43]

- The set of all musical improvisations that could possibly have influenced the system programmer, of a size that was impractically large and which is considered here as being of size ∞.
- The set of all potential input from co-musicians in a live performance, again of an impractically large size which shall again be treated as being of size ∞. Other musicians' playing is used by Voyager (Lewis, 2000) and GenJam (Biles, 2007).
- The empty set $\emptyset$, as a practical solution where the inspiring set is unknown.
- The licks (short melodies) used as seeds for GenJam, as musical fragments to be mutated and linked (Biles, 2007).

The scenarios where $|I| = \infty$ may be closest to Ritchie's original intentions, according to the quote above, but were less tractable in initial experiments and produced the same results for all three systems for criteria 9-18, without differentiating between the three systems:

[41] Track lengths ranged from 30 seconds for a GAmprovising track to several minutes for some Voyager tracks.

[42] Although Ritchie's criteria are presented as formal and objective, varying interpretations are still possible in a number of ways such as the definition of the inspiring set, the interpretation of typicality and value in a specific creative domain, as well as the need to set parameters and weights.

[43] This is discussed further in Appendix F.

9. $ratio(I \cap \mathcal{R}, I) > \theta \quad \longrightarrow \quad \frac{3}{\infty} \rightarrow 0 \not> 0.5 \therefore FALSE$
10a. $(1 - ratio(I \cap \mathcal{R}, \mathcal{R})) > \theta \quad \longrightarrow \quad 1 - \frac{3}{3} = 0 \not> 0.5 \therefore FALSE$
11. $AV(typ, (\mathcal{R} - I)) > \theta \quad \longrightarrow \quad AV(typ, \emptyset) = undefined \therefore$ not applicable
12. $AV(val, (\mathcal{R} - I)) > \theta \longrightarrow \quad AV(val, \emptyset) = undefined \therefore$ not applicable
13. $ratio(T_{\alpha,1}(\mathcal{R} - I), \mathcal{R}) > \theta \quad \longrightarrow \quad \frac{0}{3} = 0 \not> 0.5 \therefore FALSE$
14. $ratio(V_{\gamma,1}(\mathcal{R} - I), \mathcal{R}) > \theta \quad \longrightarrow \quad \frac{0}{3} = 0 \not> 0.5 \therefore FALSE$
15. $ratio(T_{\alpha,1}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \quad \longrightarrow \quad \frac{0}{0} = undefined \therefore$ not applicable
16. $ratio(V_{\gamma,1}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \quad \longrightarrow \quad \frac{0}{0} = undefined \therefore$ not applicable
17. $ratio(V_{\gamma,1}(\mathcal{R} - I) \cap T_{\alpha,1}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \quad \longrightarrow \quad \frac{0}{0} = undefined \therefore$ not applicable
18. $ratio(V_{\gamma,1}(\mathcal{R} - I) \cap T_{0,\beta}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \quad \longrightarrow \quad \frac{0}{0} = undefined \therefore$ not applicable

Treating the inspiring set as an empty set would not make use of what we know about the inspiring sets for GenJam. This option was however appropriate for Voyager and GAmprovising, neither of which used stored musical material[44] as seeds for generating their improvisations, especially given that the alternative of working with an inspiring set of size ∞ is not very practical.

The chosen option for GenJam was to represent the inspiring set as a non-empty set of unknown size for which the elements were unknown, but contained none of the output artefacts, for the reasons explained above. This sufficed for implementation of the criteria. All systems in Case Study 1 could be evaluated using Ritchie's criteria. In Case Study 2, the systems in Tearse et al. (2011) and Monteith et al. (2011) could not be evaluated as desired using Ritchie's criteria, as the related papers did not report enough details of results. In some respects this problem has been artificially introduced by the wish to capture first impressions of creativity rather than considered and detailed investigation of how creative the systems are. The point remains, however, that sometimes adequate information to evaluate systems using Ritchie's criteria is not available. If evaluating other researchers' systems to compare against one's own, or in gathering information, occasionally the only information available on a system is in published papers on that system. This is particularly the case if a researcher has left academia or is not contactable for other reasons, or if the system is no longer operational and there are no archived results to provide. In such cases, Ritchie's criteria approach becomes inapplicable.

During the collection of data from judges for Ritchie's criteria application in Case Study 2, one judge remarked that they would like to know how the artefacts were produced; something which Ritchie's criteria does not allow for. This sentiment was echoed in some of the qualitative comments offered while collecting ratings for Ritchie's criteria, in the survey for Case Study 1 (Section 8.2.1):

> [On Voyager] 'you don't specify what is the relation between the trombonist and the computer. did the computer react?'
>
> [On the system removed from Case Study 1] 'Obviously constructed by concatenating small rhythmic cells with pitches chosen to match the chord changes'

[44] The inspiring set contains the artefacts which *inspire* or seed the creative system's operation or output.

> 'I would like to know a little more about the parameters for the computer in selecting pitches, rhythm, durations, attacks, etc. for its improvisations. I would also have liked to know more about your thinking and purpose in pursuing this–what your expectations are.'
>
> 'more detailed explanation of the interactions, if any, between human and computer'
>
> 'I have trouble with the phrase "evaluate the computer systems". I guess the answer is definitely not, I could be hearing something regurgitated from their stored database of human input. How would I know if I can't see how the system works? Philosophically, evaluating a computer system in a musical realm is fraught with complications. Of course I can evaluate the music itself, but more or less independent of where it comes from.'
>
> 'I would like to know what the purpose of the program is, i.e. which parameters of the music (e.g. pitch, rhythm, dynamics, agogics, tone color etc.) are really improvised.'
>
> 'I would have liked some info about what the systems were trying to achieve with each improvisation. I mean, if you were listening to a person improvise, they might provide you with a title, or tell you that they were improvising around a particular theme or something like that.'

Chapter 3 repeatedly highlighted how creativity entails more than just the end product, evidencing this flaw with Ritchie's methodology. As one participant succinctly pointed out:

> 'I thought the questions were about the improvisations, not about the computer systems that produced them. I now feel confused.'

Even with this objection, though, judges in Case Study 1 reported a generally high level of confidence in the typicality and value ratings they provided for Section 8.2.1. Figure 8.15 shows that most judges felt 'Very Confident', 'Confident' or 'Neutral' in performing the task of rating the systems for typicality and value. This level of confidence was demonstrated across all systems' tracks and also in confidence levels reported overall for the survey. Judges providing data on typicality and value of products in Case Study 2 also reported feeling confident about the majority of their provided ratings.

Again, significant amounts of qualitative information were volunteered by the judges which could not be incorporated at all in applying Ritchie's criteria formally. Although the SPECS evaluation in Case Study 2 did not collect qualitative feedback provided by the two judges, this was directly related to how SPECS was implemented for this case study; the judges did not have time to provide additional feedback due to the time limits imposed. For Case Study 1, however, both qualitative and quantitative feedback were incorporated into the implementation of SPECS.

Many participants reported problems with addressing the questions about typicality and value. These questions were deliberately phrased in a form very close to Ritchie's original definitions (Ritchie, 2007), both to follow Ritchie's methodology as closely as possible and also to test how clear that phrasing was. As the quotes below typify, there were questions over exactly what the participants were intended to evaluate: the whole product, including the human performances, or the part of the end products produced by the computer system. For systems like GAmprovising which produced solo improvisations, this issue did not arise; however some confusion was caused in evaluating interactive systems such as Voyager, the subject of these quotes:

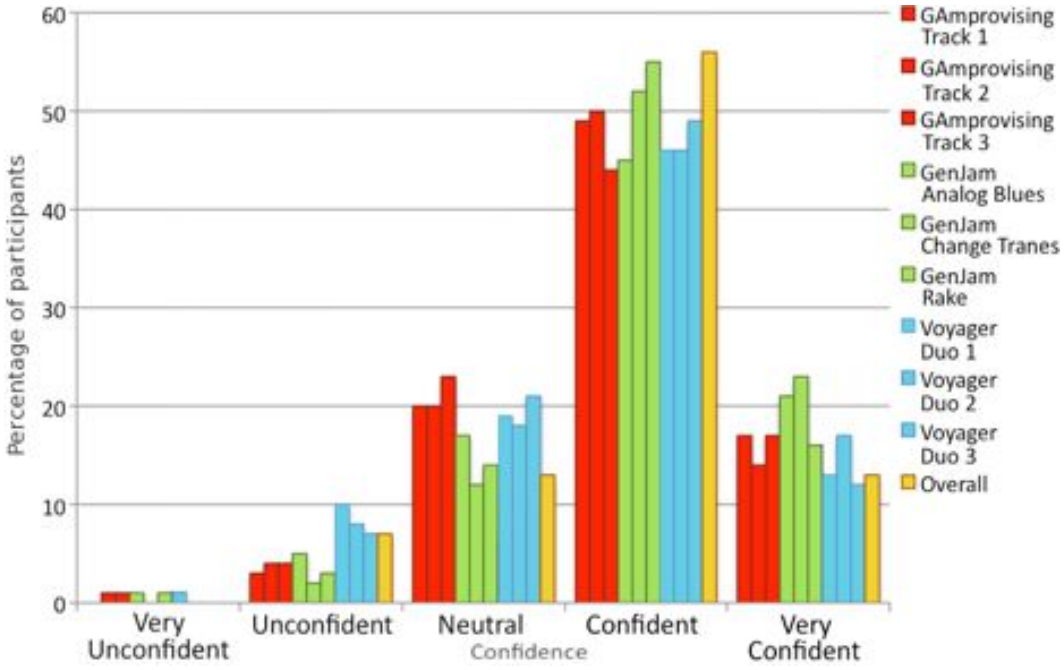


Figure 8.15: Levels of confidence reported by survey participants providing ratings of typicality and value of artefacts produced by Case Study 1 systems. If 'Very Unconfident' = 1, and 'Very Confident' = 5, then GAmprovising ratings were given with a mean of 3.8 confidence, GenJam with a mean of 4.0 confidence and Voyager 3.7 mean confidence.

> 'not sure if I was supposed to ignore the trombone playing. It would have been difficult to ignore.'
>
> 'The question is not entirely clear - do you mean the track *as a whole*, or, am I to evaluate only the ability of the computer to improvise? I've gone with the track as a whole - in which case I say, the trombone is good, the computer less so... but this is based on the trombone "leading" so it's subjective'

A related issue was summarised by these participants: what constitutes a 'typical' improvisation? There is no simple answer to this question, highlighting a flaw in Ritchie's criteria, at least for evaluating musical improvisation creativity. What a typical improvisation is depends on the context and the opinion of the people evaluating:

> 'it's not clear, given the variety of situations presented, what is a "typical" and "good" improvisation...'
>
> 'There is almost definitionally no such thing as a typical improvisation, and the best-regarded improvisers are respected for their atypicality. Therefore I think 'is this improvisation typical' is a very ambiguous question to ask, which will confound multiple interpretations together.'

Another issue arose with using thresholds for values of *typ* and *val* to pass, rather than referring to the absolute values. For example, it was not taken into account that GenJam's Change Tranes track scored 0.8 for *typ*, the highest mean rating, whereas Voyager's Duo 3 track only narrowly passed the

threshold of 0.6, with a mean rating of 0.62 for *typ* (Appendix F Table F.1). This did simplify matters somewhat, given the number of criteria to evaluate, but did not make full use of the information obtained. The participants' distinctions between 'Agree' and 'Strongly Agree' for example became irrelevant, even though it was felt important to offer participants a Likert Scale with five options rather than a more limited three options.

8.3.5 Reflections on using Colton's creative tripod framework

Essentially the three tripod qualities were treated as a subset of the fourteen components. This approach was appropriate as each quality translated closely to a corresponding component. Colton makes little suggestion on how to evaluate the three tripod qualities, beyond giving his own descriptive examples. It seemed reasonable to use the evaluative information already obtained for SPECS, rather than duplicating this effort unnecessarily.

Colton's approach enabled a wider array of evaluation data to be used than for Ritchie's criteria,[45] allowing evaluation to be informed by more of the 'Four Ps' than observations on just the end product (Chapter 3 Section 3.4.2). A subtler, more fine-grained comparison could also be achieved.[46] The size of differences in quantitative data could inform conclusions in Colton's framework (and in SPECS), whereas for Ritchie's criteria it was sufficient for numeric data to pass a threshold value; after that point, relative differences in numeric data were ignored in Ritchie's criteria.

The question of whether Colton's creative tripod is more appropriate and adequate for evaluation than SPECS became a question of whether the subset of {*Domain Competence*, *Variety, Divergence and Experimentation*, *Thinking and Evaluation*} was appropriate and adequate to represent the 14 components without losing important information. In other words, whilst accepting that some information would be lost as a result of evaluating fewer components/qualities, could all the *important* evaluative findings from the 14 components in SPECS be replicated using the set of tripod qualities?[47] Colton (2008b) defines creativity in all domains to be defined by the set of qualities {skill, imagination, appreciation}, without reflecting on their relative importances. In Chapter 7, Figure 7.2 showed that the judges in Case Study 2 disagreed. A similar finding was demonstrated in the emphases on what constitutes creativity, as shown in the findings of the improvisational creativity questionnaire for Case Study 1.[48] Priorities in Case Study 1 were:[49]

[45] SPECS also enables wider evaluation.

[46] The lack of detail in the criteria-based evaluation is a limitation acknowledged in Ritchie et al. (2007).

[47] This is a critique of how SPECS has been implemented in this particular case study, following the recommendations to adopt the Chapter 4 components as the base model of creativity. Considering the SPECS approach more generally, Colton's creative tripod could be incorporated into SPECS if desired, as long as the researcher gave justification for defining creativity in their domain as the combination of skill, imagination and appreciation, used these three qualities as the standards to test for creativity and applied reasonable evaluative measures to test these three standards.

[48] See Chapter 6 Section 6.3.2.

[49] These three components collectively accounted for 41.4% of the responses' subject matter.

- *Social Interaction and Communication* was mentioned in the questionnaire responses in Chapter 6 Section 6.3.2 a total of 44 times in a positive way (the highest tally of any of the components), with no negative mentions.
- *Domain Competence* was mentioned positively 43 times in the questionnaire responses, though it was also mentioned as potentially detrimental for creativity 6 times if too rigidly adhered to.
- *Intention and Emotional Involvement* received 41 positive mentions, with no negative mentions.

Priorities in Case Study 2:

- *Domain Competence* was considered by both judges in this case study to be at least 'Quite important' for creativity in all the systems' domains. In the cases of poetic creativity and musical creativity, one of the two judges raised their opinions on this to consider this component 'Crucial' for creativity.
- Somewhat appropriately, *Variety, Divergence and Experimentation* varied in its perceived importance for creativity, from 'Crucial' / 'Quite important' (Judge 1 / Judge 2) for poetic creativity and musical creativity, 'Quite important' (both judges) for story-telling creativity, through to 'A little important' / 'Quite important' for artistic creativity.
- *Thinking and Evaluation* was usually considered 'A little important' by Judge 1 and 'Quite important' by Judge 2, though for poetic creativity Judge 1 agreed with Judge 2 that this was 'Quite important'.

Figure 7.2 in Chapter 7 also shows that other components made 'Crucial' contributions to creativity, both across all domains and for specific domains.[50] Not all these components were not covered by the subset of {skill, imagination, appreciation}. (N.B. * = *analogous to tripod qualities*)

- Components crucial for creativity in all domains:
  - *Generation of Results*.
  - *Originality*.
  - *Spontaneity and Subconscious Processing*.
  - *Value*.
- Additional crucial components for creativity in artistic creativity: none.
- Additional crucial components for creativity in poetic creativity:
  - *Domain Competence**.
  - *Variety, Divergence and Experimentation**.
- Additional crucial components for creativity in story-telling/narrative creativity:

[50] In this context, a component is considered to make a crucial contribution to creativity if one judge deems it to be 'Crucial' and the other judge deems it to be at least 'Quite important'.

  - *Independence and Freedom*.
  - *Progression and Development*.
  - *Social Interaction and Communication*.

- Additional crucial components for creativity in musical creativity:
  - *Active Involvement and Persistence*.
  - *Domain Competence**.
  - *Progression and Development*.
  - *Variety, Divergence and Experimentation** .
  - N.B. In musical creativity, *Value* was only considered 'Quite important' by both judges.

It is clear to see that the three tripod qualities did not fully cover the key components for creativity in each of the systems' domains. To some degree, it could be argued that some tripod qualities incorporated aspects of the key components, for example skill could be linked to *Generation of Results* and *General Intellect* (Sections 8.2.2). It is difficult to see how components like *Spontaneity and Subconscious Processing*, *Social Interaction and Communication* or *Active Involvement and Persistence* could be represented using the tripod qualities. The greater definitional guidance afforded with the 14 components generated more detailed information for the system authors to work with if wanting to improve their systems' creativity.

In terms of the amount of information gathered for evaluation, the set of tripod qualities was smaller (3 as opposed to 14 for the set of components). For this case study there was restricted time available to evaluate the systems. If focusing solely on the tripod qualities then the judges would have been more likely to collect sufficient information on all three tripod qualities, if information was provided. When collecting information on 14 components under time pressures, it was likely that some relevant information may have been missed, even if provided, especially if the information provider was attempting to deliver as much information as possible in that time. In general when encountering a large amount of information, we filter out the parts that are irrelevant and pay attention to the parts that are most important for the intended purpose (Pollack & Pickett, 1957). Focusing on skill, imagination and appreciation would collect useful evaluative information; there is little doubt from the results in this thesis that these three qualities are generally important for creativity. This focus would however direct attention away from key aspects of creativity in this case study such as originality, value and the ability to produce results; the creative tripod overlooks this other important information and does not account for it. Hence the problem of incomplete information would reappear in a different form.

8.4 External evaluation of different methodologies

The above comparisons between methodologies are drawn from evidence obtained in the applications of the different evaluation methodologies in the two case studies. Previously, it has been highlighted that external evaluation should be employed where possible, rather than using evaluators from the research team involved in the system (Chapter 5 Section 5.1.5). In this regard, external evaluation was sought to consider and perform meta-evaluation on SPECS and other key existing methodologies (Ritchie, 2007; Colton, 2008b, surveys of human opinion). As details of the FACE model (Colton et al., 2011) had been published in the intervening period between the previous evaluations and the seeking of external evaluation, FACE models were also constructed for Case Study 1 systems (Section 8.4.1).

A large amount of feedback has already been obtained for systems evaluated in the case studies, especially Case Study 1. Therefore it was decided to invite the researchers behind the systems for Case Study 1[51] to view all the evaluative feedback obtained and give their opinions on various aspects of each methodology and its corresponding evaluative results.[52] The criteria for meta-evaluation of the evaluation methodologies are derived and presented in Section 8.4.2 and the methodology used for the meta-evaluation is described in Section 8.4.3. The obtained meta-evaluations are reported and discussed in Section 8.4.4.

8.4.1 Comparisons with the FACE model

Unlike the Case Studies, this external evaluation study occurred after the first publication of research on the FACE/IDEA model (Pease & Colton, 2011b; Colton et al., 2011; Pease & Colton, 2011a, also see Chapter 2 Section 2.1.5 of this thesis). The FACE model in particular has been used for creative system evaluation (Colton et al., 2011, 2012; Collins, 2012) and makes a potentially important con-

[51] Al Biles (GenJam) and George Lewis (Voyager). Bob Keller was also invited, as a key researcher behind the ImproVisor system (Gillick et al., 2010) that was originally also included in Case Study 1. Although this system was removed from the Case Study due to incorrect musical examples used, Bob Keller is still a relevant candidate to offer opinions on the evaluation results in Case Study 1 as an interested party, through his research into and development of a musical improvisation system. The author of GAmprovising - myself - was excluded from this evaluation process; I could not be an external independent evaluator of SPECS.

[52] Case Study 1 was selected in preference to Case Study 2 partly because the evaluation feedback collected was in more detail. It was felt that the researchers for systems in Case Study 1 should be made aware of this extensive feedback, if they were not already aware, particularly any information gleaned from comparisons with the other systems, as comparisons between systems were more relevant for Case Study 1 than Case Study 2. There was also less likelihood that these researchers would hold existing personal biases towards or against a particular creativity evaluation methodology, as there was no evidence of them having applied any of the computational creativity evaluation methodologies on their systems, or (critically) of them having been involved in a methodology's development. Additionally, in illustration of comments made about time limitations for evaluation, which will be found in Chapter 9 Section 9.1.3, unfortunately there was not time available to involve system researchers from both case studies in external evaluation studies. As will be reported in Chapter 10 Section 10.3, though, one of the primary tasks for future work will be to elicit feedback from the researchers behind systems in Case Study 2.

tribution to the creativity evaluation literature; hence a question arose on whether to include FACE evaluations in this current evaluation study. If included, this would generate useful comparative data between SPECS, the existing methodologies and the FACE model. The length of time available for evaluation at this stage, however, prohibited the application of FACE for evaluation in the same depth as for SPECS and the other methodologies previously examined in Case Studies 1 and 2.[53] As evaluation methodologies would need to be applied to all Case Study 1 systems, and as I was the author of one of these systems (GAmprovising), this would lead to potential bias being introduced as the author of GAmprovising (myself) would be responsible for constructing FACE models for all the Case Study 1 systems, including GAmprovising. Nonetheless, this last option was chosen, to apply FACE evaluation to all Case Study 1 systems. This affords the advantage of gaining comparative feedback between the methodologies explored so far and the FACE models, which compensates for the issues in applying FACE evaluation. Analysis was conducted with the understanding that these FACE models are based on one (potentially biased) opinion rather than on the opinions of more than one independent judge.[54] The resulting FACE models are given below for each of the three systems in Case Study 1:

GenJam GenJam's creativity can be represented as the ability to generate creative acts following the description: $< A^g, C^g, E^g >$. It does not provide natural language descriptions of what it has done so the F(rame) criterion is not present. In its more recent incarnation, GenJam does use its own self-evaluation (based on pre-existing melodies that have been judged to be good), though it does not generate novel aesthetic measures, so can be attributed the A^g criterion. It employs a creative process and generates example outputs, so can also be attributed with C^g and E^g, though it performs no meta-generation of methods to create new products or new creative concepts for artefact generation.

Voyager Voyager's creativity can be represented as the ability to generate creative acts following the description: $< C^g, E^g >$. It does not provide natural language descriptions of what it has done so the F(rame) criterion is not present. Voyager can listen to and respond to other musicians, but it appears from Lewis (2000) that Voyager's choice of responses to incoming musical input are not guided using aesthetic measures of the incoming musical material. Voyager has some 15 generation methods to use in its creative process, so ably satisfies C^g (but not C^p, unless it can generate new overarching creative processes in addition to its current strategy, at runtime). As it can generate artefacts but not generate new methods for generating new artefacts, Voyager can be attributed E^g.

[53] This acts as another pre-emptive example of the points in Chapter 9 Section 9.1.3 about time availability issues and their impact on performing evaluation.

[54] In Chapter 10 Section 10.3, one suggested area of further work is to construct FACE models for all Case Study systems using similar resources and time constraints as for the other methodologies in each Case Study.

GAmprovising GAmprovising's creativity can be represented as the ability to generate creative acts following the description: $< A^g, C^g, E^p, E^g >$. It does not provide natural language descriptions of what it has done so the F(rame) criterion is not present. GAmprovising performs self-evaluation (based on applying Ritchie's criteria to evaluate the creativity each Improviser), though it does not generate novel aesthetic measures, hence it can be attributed the A^g criterion but not A^p. GAmprovising employs generation methods in its creative process but although it can generate novel parameters for the existing method of creation (based on random generation), it cannot generate novel methods of creation, so satisfies C^g but not C^p. As GAmprovising can generate both new artefacts (i.e. new improvisations) and new methods of generating new artefacts (via generating new populations of Improvisers during an evolution cycle), it can be given the attribution of E^p and E^g.

Comparing the FACE models, no system produces creative activity that satisfies all four of the criteria, as none provide framing information, but GenJam and GAmprovising satisfy three criteria. Voyager does not use or generate aesthetic measures, hence only satisfies two criteria. For further differentiation of GenJam and GAmprovising, GenJam only meets the ^{g} part of each of the three criteria as it generates items to satisfy the criteria but not the meta-generation of methods. GAmprovising does better, with the generation of new methods for generating new artefacts (E^p), though no system is able to generate new aesthetic measures (A^p). Thus from examining FACE models, we can conclude that GAmprovising is evaluated as most creative, followed by GenJam, and lastly Voyager. This ordering contradicts the other methodologies, where GAmprovising was found least creative and GenJam most creative.

8.4.2 Criteria for meta-evaluation of creativity evaluation methodologies

A central theme of this work is that criteria for evaluation should be clearly stated and justified. This theme is also applicable to the meta-evaluation criteria used for gathering external evaluation feedback on the various creativity evaluation methodologies.

Certain areas have arisen for discussion during the thesis which could be suggested as meta-evaluation criteria for assessing creativity evaluation methodologies, such as the accuracy and usefulness of the feedback to a researcher, or ease of applicability.

Pease et al. (2001) identify two candidate meta-evaluation criteria:

> 'Firstly, to what extent do they reflect human evaluations of creativity, and secondly, how applicable are they?' (Pease et al., 2001, p. 9)

More recently, Pease has suggested the set of {generality, usability, faithfulness, value of formative feedback} as candidate criteria (Pease, 2012, personal communications). In relevant literature on

evaluation and related literature on proof of hypotheses in scientific method, other contributions could also be used as criteria for measuring the success of computational creativity evaluation methodologies, as outlined below.

Criteria for testing scientific hypotheses and explanatory theories

Sloman (1978) outlined seven types of 'interpretative aims of *science*' (Sloman, 1978, p. 26, my emphasis added), of which the third aim is the forming of explanatory theories for things we know exist. In the context of this thesis, an example of the explanatory theories mentioned in the third aim would be a theory that allows us to explain if or why a computational creativity system is creative.[55] Ten criteria were offered by Sloman (1978) as criteria for comparison of explanatory theories.

> 'So a good explanation of a range of possibilities should be definite, general (but not too general), able to explain fine structure, non-circular, rigorous, plausible, economical, rich in heuristic power, and extendable.' (Sloman, 1978, p. 53)

Within these criteria there is some significant interdependence and Sloman advises that the criteria are best treated as a set of inter-related criteria rather than distinct yardsticks, with some criteria (such as plausibility, generality and economy) to be used with caution.[56]

Thagard (1988) defined a 'good theory' as 'true, acceptable, confirmed' (Thagard, 1988, p. 48). These criteria were later expressed in the form of 'the criteria of consilience, simplicity of analogy' (Thagard, 1988, p. 99) as essential criteria for theory evaluation:[57]

- *Consilience* - how comprehensive the theory is, in terms of how much it explains.
- *Simplicity* - keeping the theory simple so that it does not try to over-explain a phenomenon. Thagard mentions in particular that a theory should not try to 'achieve consilience by means of ad hoc auxiliary hypotheses' (Thagard, 1988, p. 99). In other words, the main explanatory power of the theory should map closely to the main part of that theory, without needing extensive correction and supplementation.
- *Analogy* - boosting the 'explanatory value'(Thagard, 1988, p. 99) of a theory by enabling it to be applied to other demands. This is especially appropriate where theories can be cross-applied in more established domains where knowledge of facts is more developed.

[55] Such a mapping between scientific theories and the aims of this thesis has already been explored in Chapter 2 Section 2.2 and Chapter 5 Section 5.5; here we remember the conclusions in Chapter 5 that computational creativity evaluation is not directly mappable to scientific method, but that the shared elements between the two mean that we can learn from the study of scientific method in our aim for best computational creativity evaluation.

[56] This may go towards explaining why Sloman's list of criteria is longer than others mentioned in this Section.

[57] Thagard also saw a link between better explanations and the familiarity of what is used in the explanation: 'The use of familiar models is not essential to explanation but it helps' (Thagard, 1988, p. 95). As familiarity is seen as non-essential for good explanations, though, it was not included in the criteria that Thagard emphasised throughout discussions on theory evaluation.

Guidelines for good practice in research evaluation

Suggestions for good practice in performing evaluation in research can be interpreted as criteria that identify such good practice. For example, in his 'Short Course on Evaluation Basics', John W. Evans identifies four 'characteristics of a good evaluation'[58]: a good evaluation should be objective, replicable, generalisable and as 'methodologically strong as circumstances will permit'. In considering what constitutes good evaluation practice, the MEERA website ('My Environmental Education Evaluation Resource Assistant')[59] describes 'good evaluation' as being: 'tailored to your program ... crafted to address the specific goals and objectives [of your program'; '[building] on existing evaluation knowledge and resources'; inclusive of as many diverse viewpoints and scenarios as reasonable; replicable; as unbiased and honest as possible; and 'as rigorous as circumstances allow'. From a slightly different perspective on research evaluation, the European Union FP6 Framework Programme describes how FP6-funded projects are evaluated in terms of three criteria: a project's *rationale* relative to funding guidelines and resources; *implementation* effectiveness, appropriateness and cost-effectiveness; and *achievements* and impact of contributions of objectives and outputs.

Dealing with subjective and/or fuzzy data: Blanke's specificity and exhaustivity

In computational creativity evaluation the frequency of data being returned is low and the correctness of that data is generally subjective and/or fuzzy in definition, rather than being correct or incorrect, present or missing. Blanke (2011) looked at how to evaluate the success of an evaluation methodology for measuring aspects like precision and recall, in cases where the results being returned were somewhat difficult to pin down to exact matches due to fuzziness in what could be returned as a correct result.[60] As an evaluation solution, Blanke (2011) proposed *component specificity* and *topical exhaustivity*, following from Kazai and Lalmas (2005). Exhaustivity 'is measured by the size of overlap of query and document component information' (Blanke, 2011, p. 178). Specificity 'is determined by counting the rest of the information in the component [of an XML document] that is not about the query' (Blanke, 2011, p. 178), such that minimising such information will lead to maximising the specificity value, as more relevant content is returned.

Identifying meta-evaluation criteria

Drawing all the above contributions together, five criteria can be identified for use in meta-evaluation of computational creativity evaluation methodologies. These are presented here, with relevant points

[58] Evans' course is published at `http://edl.nova.edu/secure/evasupport/evaluationbasics.html`, last accessed November 2012.

[59] All quotes from the MEERA website are taken from `http://meera.snre.umich.edu/plan-an-evaluation/evaluation-what-it-and-why-do-it#good`, last accessed November 2012.

[60] The specific case Blanke considered was in XML retrieval evaluation, where issues such as hierarchical organisation and overlap of elements, and the identification of what was an appropriate part of an XML document to return, caused problems with using precision and recall measures. There was also an issue with relatively low frequencies in what was being returned.

from the comments above being grouped under the most relevant criterion, as far as possible.[61]

- **Correctness:**
  - MEERA's *honesty of evaluation* criterion.
  - MEERA's *inclusiveness of diverse relevant scenarios* criterion.
  - Evans' *objectiveness* criterion.
  - MEERA's *avoidance of bias in results* criterion.
  - Sloman's *definiteness* criterion.
  - Sloman's *rigorousness* criterion.
  - Sloman's *plausibility* criterion.
  - Thagard's *consilience* criterion.
  - Blanke's *exhaustivity* criterion.
  - Evans' *methodological strength* criterion.
- **Usefulness:**
  - Pease's *value of formative feedback* criterion.
  - FP6's *rationale*, *implementation* and *achievements* criteria.
  - Sloman's *heuristic power* criterion.
  - Thagard's *analogy* criterion.
- **Faithfulness as a model of creativity:**
  - Pease et al. (2001)'s *reflection of human evaluations of creativity* criterion.
  - Pease's *faithfulness* criterion.
  - MEERA's *tailoring of the method to specific goals and objectives* criterion.
  - Blanke's *specificity* criterion.
- **Usability of the methodology:**
  - Pease et al. (2001)'s *applicability* criterion.
  - Pease's *usability* criterion.
  - Evans' *replicability* criterion.
  - MEERA's *replicability* and *rigorousness of a methodology* criteria.
  - Sloman's *non-circularity* criterion.
  - Sloman's *rigorous* and *explicitness* criteria (in how to apply the methodology).
  - Sloman's *economy of theory* criterion.
  - Thagard's *simplicity* criterion.

[61] Some overlap across criteria is acknowledged, for example Thagard's *analogy* criterion can be interpreted as being concerned with both 'usefulness' and 'generality'.

- **Generality:**
  - Pease's *generality* criterion.
  - MEERA's *inclusiveness of diverse relevant scenarios* criterion.
  - Evans' *generalisability* criterion.
  - Sloman's *generality* criterion.
  - Sloman's *extendability* criterion.
  - Thagard's *analogy* criterion.

8.4.3 Methodology for obtaining external evaluation

Each external evaluator was given a feedback sheet reporting the evaluation feedback obtained for their system from each creativity evaluation methodology being investigated: SPECS; Ritchie's criteria; Colton's creative tripod; survey of human opinion; and the FACE model. For each methodology, the sheets also included brief comparisons between systems according to the systems' evaluated creativity. An example of these feedback sheets, Appendix G presents the sheet provided to Al Biles to report the evaluation results for GenJam. A similar set of feedback was prepared and sent to George Lewis as evaluative feedback relating to Voyager. Methodologies were presented under anonymous identifiers in the feedback sheet to prevent any bias from being introduced, for example if the evaluator had heard of that methodology before or if the evaluator could identify SPECS as the methodology belonging to the person asking them for feedback.

Evaluators were first asked if they had any initial comments on the results. They were then asked to provide full feedback for each methodology in turn, on the five criteria derived above in Section 8.4.2. They looked at all five criteria for the current methodology and then were asked for any final comments on that methodology before moving onto the next methodology. Methodologies were presented to the evaluators in a randomised order, to avoid introducing any ordering bias.

For each criterion, questions and illustrating examples were composed to present the criterion in a context appropriate for computational creativity evaluation. These questions and examples, listed below, were put to external evaluators to gather their feedback on each criterion as meta-evaluation of the various evaluation methodologies.

- **Correctness:**
  - How correct do you think these results are, as a reflection of your system?
  - For example: are the results as accurate, comprehensive, honest, fair, plausible, true, rigorous, exhaustive, replicable and/or as objective as possible?
- **Usefulness:**
  - How useful do you find these evaluation results, as an / the author of the system?

  - For example: do the results provide useful information about your system, give you formative feedback for further development, identify contributions to knowledge made by your system, or give other information which you find helpful?

- **Faithfulness as a model of creativity:**
  - How faithfully do you think this methodology models and evaluates the creativity of your system?
  - For example: do you think the methodology uses a suitable model(s) of creativity for evaluation, does the methodology match how you expect creativity to be evaluated, how specifically does the methodology look at creativity (rather than other evaluative aims)?
- **Usability of the methodology:**
  - How usable and user-friendly do you think this methodology is for evaluating the creativity of computational systems?
  - For example: would you find the methodology straightforward to use if wishing to evaluate the creativity of a computational creativity system (or systems), is the methodology stated explicitly enough to follow, is the method simple, could you replicate the experiments done with this methodology in this evaluation case study?
- **Generality:**
  - How generally do you think this methodology can be applied, for evaluation of the creativity of computational systems?
  - For example: can the methodology accommodate a variety of different systems, be generalisable and extendable enough to be applied to diverse examples of systems, and/or different types of creativity?

For each criterion, evaluators were asked to rate the system's performance on a 5 point Likert scale (all of a format ranging from positive extreme to negative extreme, such as: [Extremely useful, Quite useful, Neutral, Not very useful, Not at all useful]). They could also add any comments they had for each criterion.

Evaluators were asked about the correctness and usefulness of the methodology's results, before learning how the methodology worked. This gave the advantage of being able to hear the evaluators' opinions considering the feedback results in isolation, without any influence from how the results were obtained. Nonetheless, as this thesis has argued a number of times (most notably in Chapter 3 Section 3.4.2 and Chapter 5 Section 5.1.2), the process by which a product was generated is important to consider alongside that product, for a more rounded and informed evaluation. Evaluators were given details on how that methodology worked after evaluating the correctness and usefulness criteria. They

were then asked to provide feedback for the final three criteria (faithfulness, usability and generality). The descriptions of each system are given in Appendix G.

Finally, evaluators were asked to rank the evaluation methodologies according to how well they thought the methodologies evaluated the creativity of their system overall. Although the formative feedback is, again, probably more useful in terms of developing the various methodologies, it was interesting to see evaluators' opinions on how the methodologies compared to each other. The rankings, completed by Al Biles and Bob Keller, are reported in Table 8.16. At this point, evaluators were also given a change to add any final comments, before finishing the study.

8.4.4 Results and discussion of meta-evaluation of Case Study 1 methodologies

Al Biles [AB] completed a full evaluation of all methodologies and George Lewis [GL] provided evaluations of Colton's creative tripod and the SPECS methodology.[62] The initials by each criterion discussion below act as a reminder of which evaluators could provide comments for that criterion.[63] As well as using a 5-point Likert scale to rate methodologies on each criteria, evaluators sometimes gave comments to accompany their ratings. Where given, these extra comments are acknowledged in the discussions below.

Bob Keller [BK] also provided comments on some aspects of all methodologies, from the perspective of his research on the development of the musical improvisation system Impro-Visor. Given that Keller was commenting from a different perspective to the other two evaluators, his comments are considered at the end of this Section 8.4.4 alongside Biles' and Lewis' more general comments prior to and after the main meta-evaluation.

Correctness (AB [all], GL [SPECS & creative tripod]) FACE evaluations were considered to be 'completely correct' by Biles, more so than SPECS which both Biles and Lewis considered to produce 'quite correct' results. While Biles agreed with most of SPECS' feedback, particularly the relative weaknesses identified in *Spontaneity and Subconscious Processing* and *Intention and Emotional Involvement*, Biles corrected some judges' comments on *Dealing with Uncertainty* by explaining that 'GenJam is incapable of playing a theoretically wrong note'; he describes GenJam's robustness at dealing with unpredictable events in live performance, though he conjectures this may be because GenJam 'doesn't realize there was any uncertainty'.

The results reported by Colton's creative tripod were considered 'completely correct' by Biles and 'quite correct' by Lewis. Ritchie's criteria produced 'quite correct results' according to Biles, who agreed with most criteria (particularly on the constrained typicality of GenJam's results) apart from

[62] These methodologies were shown to Lewis first and second in the randomly ordered presentation of methodologies.

[63] Discussions are presented with the acknowledgement that not all methodologies could be considered by both evaluators.

criterion 9; Biles argued that GenJam can replicate members of the inspiring set, though the likelihood of this occurring may not be enough to exceed the threshold value for satisfying the criterion.

Biles was neutral on the correctness of the opinion survey results, agreeing with the difficulty of evaluating such a subjective concept as creativity in this open-ended way, without a definition. While Biles agreed with the validity of most of the reported comments, positive and negative, and appreciated the personal feedback, the numeric results 'were a bit mushy' and fairly unreliable. It is notable that despite being considered as a potential 'ground truth' for evaluating creative systems (e.g. Zhu et al., 2009), the survey of human opinion received the lowest rating for correctness of results.

Usefulness (AB [all], GL [SPECS & creative tripod]) Biles was 'neutral' on the usefulness of the FACE model information. GenJam is intended to meet the goal of improvising 'well enough to be a stimulating sideman for me [Biles] when I play gigs with it') rather than to explore any 'meta-level "understanding" '; hence Biles was not interested in seeing 'if GenJam can "prove" that it understands or can explain jazz the way a human could'. On the other hand, both Lewis and Biles found SPECS' feedback to be 'extremely useful'. Though Biles found that 'the level of detail and sheer comprehensiveness of all the criteria is imposing', he noted that the results provided confirmatory evidence for many of his thoughts about GenJam.

Colton's creative tripod was also found to be 'extremely useful' by both Biles and Lewis, with Biles commenting that the set of tripod qualities mapped well to the set of terms he has often used to guide his thinking ('competence, spontaneity and taste') and that he appreciated the more structured feedback of the feedback compared to the opinion survey results. Biles found Ritchie's criteria to be 'quite useful' and appropriate, although the Boolean nature of the criteria prompted Biles to use the analogy 'Assume a spherical chicken' to describe the reductive constraints of Boolean criteria.

As GenJam has been presented to audiences for several years now, Biles has already collected much feedback from surveying opinions. As the opinion survey returned feedback that Biles had already encountered, he was 'neutral' about its usefulness.[64]

Faithfulness as a model of creativity (AB [all], GL [SPECS & creative tripod] FACE received 'neutral' evaluations from Biles for how faithfully it modelled creativity. Biles commented that FACE did not acknowledge the collaborative aspect of creativity, which is very important in GenJam.

Biles thought SPECS was 'very thorough' and modelled creativity 'extremely faithfully', while Lewis was 'neutral' on this aspect of SPECS. Biles also considered the creative tripod to be an 'extremely faithful' model of creativity, as it matched his own previous thoughts (see the above comments on Usefulness). Lewis was 'neutral' about the faithfulness of the creative tripod model, commenting on the difficulties involved in conflating judgements of the creative process with aesthetic judgements.

[64] For a less established system, the opinions survey data may be considered more useful.

Lewis suggested that creativity could be evaluated by 'people who are competent in both areas [the creative process and aesthetic judgements], but also people who are blind to one or the other.'

The opinion survey was deemed to 'quite faithfully' model creativity by Biles, as 'you get unadulterated opinions from folks, which are very appropriate in an inherently subjective domain like "creativity".' Biles gave the same 'quite faithful' verdict to Ritchie's criteria (in the latter case, basing his verdict on the constraints imposed by using boolean rather than continuous criteria).

Usability of the methodology (AB [all], GL [SPECS & creative tripod], BK [all]) Some interesting points were raised in thoughts about this criterion based on participants' personal qualities and how this would affect usability in a number of cases. For example, Biles commented that the abstract nature of Ritchie's criteria raised the 'empirical question as to whether respondents could map the questions onto their perceptions of the music', therefore remaining 'neutral' overall on the usability of Ritchie's criteria. The opinion survey was deemed 'quite user-friendly' by Biles, observing that 'nothing is simpler than just cutting to the chase and asking whether something is creative or not.' but adding the caveat that 'this would be extremely user friendly for folks who have a clear idea in their head of what "creative" means, but it would be more difficult for "indecisive" folks.'

Biles also highlighted the need for expertise in the methodology itself in order to apply SPECS correctly, seeing it as 'a tool that requires a lot of practice before it can be used productively'. 'On the other hand [Biles noted], if this instrument is used by knowledgeable and motivated evaluators, it could yield a lot of useful information', but overall Biles found SPECS to be 'not very user-friendly'.[65] Lewis was more positive about SPECS, finding it 'quite user-friendly', and expressed the same opinion about the creative tripod, although he reflected on the influence of using evaluators with experience in the specific genre of music that was being improvised, as well as the influence of the ' "computerized" nature of the music', rather than hearing more acoustic performances.[66] Biles was most positive about the creative tripod, seeing it as 'extremely user-friendly' because the tripod qualities are accessible, understandable concepts. FACE was found by Biles to be 'quite user-friendly', with'just a 4X2 table to fill in, and the determination for each cell should be pretty easy'.

Generality (AB [all], GL [SPECS & creative tripod], BK [all]) Biles was 'neutral' about the generality of the FACE model and Ritchie's criteria, due to seeing both positive and negative evidence for the generality of each methodology. Biles noted that the FACE model was reported in (Colton et al., 2011) to be derived from reflections on systems in both mathematical and visual creativity, 'certainly different domains that likely conceptualize creativity differently', but also observed that the

[65] This highlights the importance of choosing appropriate tests to implement SPECS with the Chapter 4 components; this is considered further in Chapter 10 Section 10.3.

[66] Lewis reported how his current use of Disklaviers 'has made it more difficult to tell the human from the computer players'.

FACE model of explanation of creative process would be failed by 'many artists [who] are unable to articulate how they are creative, much less explicate their process objectively'. Biles' neutral opinion on the generality of Ritchie's criteria was due to how generality would be influenced both positively and negatively by the generic nature of the terminology used in the criteria evaluation.

Lewis and Biles disagreed on both SPECS and the creative tripod. Biles thought SPECS could be applied 'quite generally', due to SPECS being a 'somewhat configurable' approach; Lewis saw SPECS as 'not very generally' applicable, although he did not comment further on this view. On the creative tripod, Lewis commented that the creative tripod was 'not very generally' applicable across different domains due to issues with how to make judgements of qualities such as skill:[67]

> 'artists are trying to make work that hasn't been encountered before, rather than trying to replicate existing models of musicality. That makes judgements, particularly of skill, even more difficult, particularly in an era when virtuosity-based models of aesthetics aren't as influential in other fields of art making as they continue to be in music-language poetry, conceptual art, etc. ... [or] when some artists are likely to look askance upon any work purporting to be creative that replicates existing genres or is generally seen as non-exploratory or questioning.' (Lewis)

Lewis's opinion on the creative tripod was contrasted by Biles, who was not sure how the tripod would apply in less subjective domains, but noted that the (Colton et al., 2011) discussion of domains other than artistic domains made him 'suspect [the creative tripod] would hold up well', therefore considering the creative tripod to be 'quite generally' applicable across domains.

On the opinion survey, Biles felt this could be 'quite generally' applicable across different domains, echoing his previous comments that while 'reliability is sacrificed for validity [in the opinion survey] ... thats a defensible trade in artistic domains.[68]

Further comments on the methodologies Some points were raised by evaluators when there were opportunities to make more general comments on the methodologies, or overall. Additionally, Bob Keller provided some evaluation feedback based on his experiences in developing the musical improvisation system Impro-Visor.

Biles mentioned issues with the small numbers of judges involved in SPECS and in Colton's creative tripod (which used the same data as SPECS).

Keller considered the demands of stylistic constraints on a system and how to determine whether a system should be considered more creative or less musical, should it break a given constraint. For example, 'outright creativity in the absence of constraints, such as the generated results should "swing" can allow a very creative, but not very "swinging" system to be marked higher.'

Commenting on specific methodologies, Keller remarked that the term 'appreciation', used in

[67] This comment may also be intended to refer to Lewis's opinion of SPECS as 'not very generally' applicable.

[68] It could be speculated that Biles' opinions here are based around 'artistic domains' rather than including domains such as mathematical creativity which he had mentioned earlier.

Colton's creative tripod, could be confusing. Having come across this term before in the context of computational creativity, Keller questioned if 'any system in today's state of the art can truly appreciate, in the sense that it would be applied to a human.' He also noted, for Ritchie's criteria, that a better understanding of how a system generated artefacts would impact on the accuracy of data obtained from evaluation; hence Keller places emphasis on the creative process, which Ritchie's criteria do not account for. On SPECS, Keller commented that 'if anything could be done to reduce the number of categories, that might it more attractive', echoing a previous similar comment made by Biles on the volume of data generated by SPECS.[69]

Commenting on the Impro-Visor system, Keller observed that the aspects of FACE had not necessarily been a priority for current development but that he would want to study the criteria more carefully to see if it would aid future development towards more creativity. He saw the FACE model as more for guiding development of the creativity of Impro-Visor than for post-hoc evaluation of a system which had not necessarily been designed with the FACE criteria in mind. Additionally, Keller did not feel the FACE model could be useful for addressing other goals of the system, such as providing information on the educational aspects of Impro-Visor.

A final point mentioned by Biles and Keller was how creativity related to the goal of the system development. Creativity was seen as an interesting aspect to develop within these systems, but not the main end goal for any of the three systems. GenJam was developed as a performer to play with, complement and inspire Al Biles in live improvisations. Impro-Visor, Keller's system, was designed for educational purposes, to develop the improvisation skills of human players. From Lewis's comments, his motivations for Voyager also seem to focus more on exploring musical virtuosity and aesthetics rather than creativity. Development efforts for these systems have therefore been concentrated on maximising other aspects of the system; Keller notes that 'it is hard to use [the methodologies] to make cross comparisons of systems when the systems were designed with different goals and assumptions'. In their responses, however, all evaluators treated creativity as a positive aspect to acknowledge within their system, even if they had previously not given it as much attention as other aspects.

Summary of external evaluation results Al Biles summarised the meta-evaluation of the five different methodologies with: *'Five very different approaches, and each bring something to the table.'* In the comparisons between methodologies and the overall rankings listed in Table 8.16, SPECS was either considered the best methodology overall (ahead of the creative tripod) or the second best (behind Ritchie's criteria) for evaluating a system's creativity. Echoing a recurring theme throughout this thesis though, the more useful information comes from the more detailed formative feedback and comments rather than a single summative ranking.

[69] See Chapter 10 Section 10.3 for proposals on how to address this issue of volume.

Table 8.16: Comparisons of the methodologies (where 1=best, down to 5) in answer to the question: *'Please could you rank the evaluation methodologies according to how good you think they are, overall, for evaluating how creative your system is?'*. SPECS appears either first or second.

Position	Al Biles GenJam	Bob Keller [Impro-Visor] (not in Case Study 1)
1	Method SB (SPECS)	RC (Ritchie's criteria)
2	Method CT (Creative Tripod)	SB (SPECS)
3	Method RC (Ritchie's criteria)	Method FD (FACE)
4	Method OS (Opinion survey)	Method OS (Opinion survey)
5	Method FD (FACE)	Method CT (Creative Tripod)

SPECS was evaluated by both Biles and Lewis, with some additional comments from Keller. SPECS generated 'extremely useful' and 'quite correct results', in both of the main evaluators' opinions. One evaluator found SPECS to be an 'extremely faithful' model of creativity, though the other was 'neutral' on this matter. While one evaluator found SPECS 'quite user-friendly', the other questioned how user-friendly the SPECS methodology would be, given the steep learning curve in understanding the components. In terms of generality, evaluators disagreed on how generally SPECS could be applied, further comments illustrated how methods like SPECS were more appropriate for taking into account other system goals, compared to more limited views on creativity such as in the FACE model. Biles and Keller in particular commented on the lack of accommodation of other system goals in the FACE model, though it is to be acknowledged that such accommodation does not form one of the goals of the FACE model and is more of an unintended but useful consequential result in models such as SPECS. FACE was placed third in the overall rankings by Biles and last by Keller. Biles, the main evaluator for FACE, found the results generated by FACE to be 'completely correct',[70] were neutral on the usefulness of FACE model feedback, the generality of the FACE model across domains and its faithfulness as a model of creativity. FACE was deemed 'quite user-friendly' due to its simplicity; this opinion was repeated, more strongly, for the other creativity evaluation framework Colton was involved in, the creative tripod. Lewis and Biles both evaluated the tripod; they disagreed as to whether the tripod would be generally applicable across many domains, and also as to how faithfully the tripod modelled creativity. Both evaluators agreed, however, that the feedback from the tripod was 'extremely useful' and either 'completely correct' or 'quite correct'. Biles ranked the creative tripod as the second best creativity evaluation methodology overall, though Keller placed it last.

Ritchie's criteria methodology was fully evaluated by Biles. Biles found the criteria to produce

[70]This in itself is useful feedback, as in contrast to the other methodologies, the FACE models were constructed through one person's interpretation of the systems through research of the systems and their associated documentation, so their correctness would be more prone to be called into question.

'quite correct', 'quite useful' feedback that was 'quite faithful to creativity' (despite raising issues with enforced simplifications of the data due to the boolean rather than continuous nature of the feedback). Biles was 'neutral' on the usability of applying the criteria for creativity evaluation and on their generality, questioning how the generic terminology used to solicit ratings of typicality and value could be applied to different domains successfully. Keller considered Ritchie's criteria to be the best methodology overall for creativity evaluation, though Biles gave it a middling ranking.

The opinion survey was ranked overall to be the fourth best methodology out of the five. It received a few negative comments from Biles, the main evaluator for this system, despite Biles noting that 'nothing is simpler than just ... asking whether something is creative or not' and that the survey solicited spontaneous,'unadulterated' opinions rather than restructuring the feedback (though Biles also noted that the tripod feedback was clearer than the survey feedback due to its more structured presentation). Biles was guided in a number of comments by an observation that the opinion survey sacrificed reliability/consistency of results for greater validity in terms of the personal qualitative feedback. He thought that the survey approach could be applied 'quite generally' and was 'quite user-friendly' and 'quite faithful' to what it means to be creative. The success of this methodology would depend on the type of person participating, and whether they were clear on what 'creative' means. Given that the GenJam system has been publicly presented many times before, though, Biles felt he learned nothing new from the feedback from the survey, unlike the other methodologies. He was 'neutral' on the correctness of the methodology, confirming observations made earlier in this Chapter (Section 8.1) that human opinion cannot necessarily be relied on as a 'ground truth' to measure evaluations against, due to varying viewpoints.

8.5 Comparing and contrasting methodologies

Five meta-evaluation criteria have been identified for meta-evaluation of creativity evaluation methodologies, as reported in Section 8.4.2. Below, the criteria are applied to analyse all the methodologies investigated earlier in this Chapter, using findings from both of the Case Studies.[71] Such considerations on the methodologies allow us to compare if, and how, SPECS represents a development of our knowledge on how to evaluate the creativity of computational systems. The considerations below complement the findings in Section 8.4 by accounting for more detailed information and observations that may not have been detected by the external evaluators, but which should still be considered, as well as the results from Case Study 2, which the external evaluators were not presented with.[72]

[71] The FACE model (Colton et al., 2011) is mostly excluded from these observations, as it was applied in different circumstances to the other methodologies and hence comparisons would not be completely fair.

[72] As Chapter 10 Section 10.3 discusses, one area of future work will be to conduct an external evaluation similar to that reported in Section 8.4 of this Chapter, soliciting opinions and feedback on the evaluative results from the researchers behind the Case Study 2 systems.

Correctness Showing that human opinion cannot necessarily be relied on as a ground truth, even on a large scale, some participants in opinion surveys (especially in Case Study 1) admitted that they were likely to be evaluating the systems based on the system's quality rather than its creativity, which would affect the overall correctness of the results of evaluations from the human opinion survey.

SPECS performed better than Ritchie's criteria for correctness. Although the 18 criteria have a comprehensive coverage of observations over the products of the system, criteria evaluation is based solely on the products of the creative system, not accounting for the system's process, or observations on the system or how it interacted with its environment.[73] Colton's tripod model was found to be reasonably accurate in terms of identifying and evaluating important aspects in the two case studies, but it has disregarded aspects such as social interaction, communication and intention, which have been found to be very important in understanding how musical improvisation creativity is manifested.

It should be noted that 'correctness' does not imply that the results from evaluation match common human consensus as a 'ground truth', or 'right answer'; both case studies have demonstrated that these are not reliable goals in creativity evaluation. Instead, correctness is concerned with how appropriate the feedback is and how accurately and realistically the feedback describes the system.

Usefulness The methodologies differed in the amount of feedback generated through evaluation. A fairly large volume of qualitative and quantitative feedback was returned through the application of SPECS, unlike Ritchie's criteria which only returned a set of 18 Boolean values, one for each criterion. Colton's creative tripod generated feedback for 3 components, rather than 14 components, so was shorter than SPECS. The human opinion surveys generated similar quantities of feedback to SPECS, from more people but a shallower level of detail.

The human opinions surveys returned less detailed feedback than SPECS, which generated a large amount of detailed formative feedback. The opinion surveys' feedback also may have been skewed towards the systems' quality rather than creativity, according to participant feedback (Section 8.1.1).

Ritchie's criteria returned a set of boolean values rather than any formative feedback, in a fairly opaque form given the formal abstraction of the criteria specification; if there was a lack of output examples, Ritchie's criteria could not generate any feedback at all, even based on other observations about the system. Colton's creative tripod returned information at the same level of detail as SPECS per component/tripod quality, but less information overall, as several useful components of SPECS were overlooked because they did not map onto the set of tripod aspects.

Faithfulness as a model of creativity Participant feedback for the human opinion surveys acknowledged that evaluations may have related more to the quality of the system, not its creativity, with several participants requesting a definition of creativity to refer to in the Case Study 1 survey. SPECS

[73] See Chapter 3 Section 3.4.2 for details of the Four Ps of creativity: Person, Product, Process and Press (environment).

requires researchers to base their evaluations on a researched and informed understanding of creativity that takes into account both domain-specific and domain-independent aspects of creativity. In this way it is the only methodology that directly accounts for specific informed requirements for creativity in a particular domain. Human opinion surveys would acknowledge this but only tacitly, without these requirements necessarily being identifiable or explainable. Although the parameters and weights in Ritchie's criteria could be customised to reflect differing requirements for creative domains, in practice no researchers have attempted this when applying Ritchie's criteria, probably due to the formal and abstracted presentation of the criteria. In Colton's creative tripod, all three tripod qualities are treated equally in previous examples (including those in Colton (2008b)) regardless of their contribution in a specific creative domain and no further qualities can be introduced into the tripod framework.

Usability of the methodology Less information was collected for Colton's creative tripod than for SPECS or the other methodologies, impacting on usability of the methodology in terms of time taken. Coupled with the informal nature of performing creativity evaluation with the tripod framework, Colton's creative tripod emerged as the most easy-to-use of the methodologies evaluated. Data collection for the other methodologies was of a similar magnitude, although data analysis for Ritchie's criteria was slightly more involved and more specialist than the other methodologies, requiring a specific understanding of the criteria.

Feedback from Section 8.4 reflected on the volume of data generated by using the components as a base model of creativity, as recommended for SPECS. If SPECS is applied without using the Chapter 4 as the basis for the adopted definition of creativity, then SPECS becomes more involved and more demanding in terms of researcher effort, negatively affecting its usability. Hence the recommendation for using the components within SPECS becomes further supported and perhaps strengthened. Although the SPECS methodology makes no formal requirements to use external evaluators, the accompanying commentary to SPECS has strongly encouraged researchers to follow this path in order to capture more independent and unbiased results (as well as a larger variety of opinions). Using external evaluators increases the time demands of the experiment in the same way as for the human opinion surveys, as both require studies to be carried out and introduce extra work to be done such as planning experiments for participants or applying for ethical clearance for conducting experiments with people. These extra demands are not necessarily encountered when performing evaluation as recommended using Colton's tripod or Ritchie's criteria (or FACE evaluation), though it is important to acknowledge that there is a trade-off in terms of potential bias being introduced for these three methodologies due to the lack of external evaluation being performed.

Generality SPECS, Colton's tripod and to some extent, Ritchie's criteria and the human opinion surveys, could all be applied to different types of system, providing that the system produces the appropriate information relevant to the individual methodologies. Case Study 2 in particular illustrates this. Lack of output examples affect the generalisability of Ritchie's criteria, as the criteria cannot be applied to systems with no discernible output. There is also some question of whether opinion surveys could be carried out for evaluating all types of creativity, particularly where creativity is not manifested outwardly in production of output, affecting the generality of opinion surveys.

Overall comparisons From the above discussions, we can focus particularly on how SPECS performs in each aspect and see whether SPECS outperforms other methodologies or encompasses the good performance of other methodologies. This helps to understand if SPECS represents progress in computational creativity evaluation methodologies.

Considering all the observations made in this Chapter from the perspective of the five meta-evaluation criteria derived in Section 8.4.2, SPECS performed well in comparison with the other evaluation methodologies on its faithfulness in modelling creativity. SPECS also performed better than Ritchie's criteria for usefulness and correctness and produced larger quantities of useful feedback than Colton's creative tripod (because less information was collected for Colton's creative tripod, although this had the effect of making Colton's creative tripod the easiest to use of the methodologies evaluated.)

Somewhat counterintuitively, SPECS (and generally all the other methodologies) were more likely to generate correct results compared to the surveys of human opinion. Many participants in the opinion surveys reported that they evaluated systems based on quality rather than creativity, due to difficulties in evaluating creativity of the Case Study systems without a definition of creativity to refer to. There is also some question of whether human opinion surveys could be carried out for evaluating all types of creativity, particularly where creativity is not manifested outwardly in copious production of output, affecting its generality. Lack of output examples would also affect the usability and generalisability of Ritchie's criteria.

Although it must be remembered that the FACE evaluations were conducted by fewer judges, without independent verification and for Case Study 1 systems only, it is conjectured that SPECS outperforms the FACE model for correctness in Case Study 1, given inconsistency between the comparative creativity evaluations using the FACE model and all the other evaluation methodologies considered.

Taking more specific detail from comparisons in this Chapter, the main advantages of SPECS, compared to existing creativity methodologies, are:

- Customising evaluation standards towards specific requirements of the specified domain, allowing (and encouraging) researchers to account for domain-specific variances.

- Requiring researchers to be able to clearly state and justify what standards they use for creativity evaluation, based on an informed and researched view of what it means to be creative.
- Prioritising detailed and formative constructive criticism to assist the researcher in future system development, instead of concentrating on generating cursory summative feedback.

8.6 Summary

Several evaluation methods were applied to the systems evaluated in Case Studies 1 and 2. As well as SPECS (Chapters 6 and 7), Section 8.1 consulted human opinion to try and capture a 'ground truth' for creativity evaluation (Zhu et al., 2009). Two key existing methodologies for computational creativity were also applied in Section 8.2 (Ritchie's criteria and Colton's creative tripod, reported in Ritchie, 2007; Colton, 2008b, respectively). Results were compared in Section 8.3; it was noted that few 'right answers' or 'ground truths' for creativity were found (especially in Case Study 2).

In Case Study 1, GenJam was generally found most creative overall, with GAmprovising as the least creative system and Voyager ranked in between. Summative comparisons are of limited value to the researcher, though, especially in terms of identifying key contributions to knowledge from the system development and weaknesses of the system to be improved. Ritchie's criteria particularly illustrated this, as the level of abstraction away from the system itself in the feedback obscured what could be learnt from that feedback. Information loss was magnified as qualitative feedback was disregarded and as results were reported as Boolean rather than continuous values.

For the purposes of progressing in research, learning from advances and improving what has been done, formative evaluation feedback is more constructive (as Chapter 1 Section 1.3 has explained). Colton's creative tripod framework, the opinion surveys and the SPECS methodology all performed well at providing this feedback. When implementing SPECS using the Chapter 4 components, SPECS gave feedback in the most detail but needed the most information gathered for evaluation.

Colton's creative tripod performed relatively well but did not evaluate all systems completely when information was limited, as was the case in Case Study 2. Also, this tripod assumes that creativity can always be represented sufficiently and completely by *skill*, *imagination* and *appreciation*. The case studies found evidence otherwise; other aspects were often as important or more so for creativity in a given domain, while the relevance and contribution of different components varied according to the creative domain being investigated.

In the surveys of human opinion in this Chapter, participants reported difficulties in evaluating the systems' creativity. In particular, several people wanted a definition of creativity to refer to in evaluation rather than relying on their own intuitive understanding. This may be due to biases against computational creativity or a lack of acquaintance with a computer being creative, as considered in

Chapter 1. Most participants did however appear to be positive or at least neutral towards computational creativity; this might not stop subconscious biases affecting evaluations (Moffat & Kelly, 2006) it would reduce the likelihood of overt negative biases being demonstrated. The difficulties may instead be due to people finding it hard to objectively rate a subjective concept like creativity when asked to (Chapter 3), although participants generally reported feeling confident about their responses.

External evaluation was solicited from the authors of the Case Study 1 systems and one other researcher with interests in creative musical improvisation systems. Five criteria were identified from relevant literature sources for meta-evaluation of important aspects of the evaluation methodologies (Section 8.4.2): *correctness, usefulness, faithfulness as a model of creativity, usability of the methodology*, and *generality*. The methodologies were compared based on the external evaluators' feedback concerning the evaluations performed on their system and the comparative feedback generated by each methodology considered so far. Given the timing of this meta-evaluation study, FACE evaluation models could also now be constructed for the systems in Case Study 1, hence were included in this study. The results showed that SPECS compared favourably to the other methodologies both overall and on most of the five meta-evaluation criteria, though the volume of data produced by SPECS raised questions on SPECS's usability and presentation of this data had some negative influence on the initial usefulness of the results compared to more succinct presentations.[74] Comparative feedback was given in Section 8.4.4 on SPECS, Ritchie's criteria, Colton's creative tripod, the FACE model and human opinion surveys as computational creativity evaluation methodologies.

The five meta-evaluation criteria in Section 8.4.2 were also applied for further analysis of the data accumulated in this thesis for both case studies. Considering all the case study findings along these five meta-evaluation criteria: SPECS performed very well on the criterion of faithfulness as a model of creativity and useful feedback, and was an improvement on some other methodologies in correctness and generality, though it was outperformed by Colton's tripod for usability.

SPECS compared well against the other evaluation methods used, especially in terms of the amount of detailed formative feedback generated and its faithfulness in evaluating creativity rather than closely related but distinct alternative aspects. A number of points did however arise for consideration during the implementation of SPECS in Case Studies 1 and 2. These will be examined in Chapter 9, alongside reflections on reactions so far to earlier versions of this work in the computational creativity research community.[75]

[74] These questions will be returned to again in Chapter 10 Section 10.3.

[75] Content in the current Chapter and in Chapter 9 also prompt suggestions for future work; these suggestions will be discussed in Chapter 10 Section 10.3.

Chapter 9

Evaluation: Reflections on and reactions to the SPECS methodology

---

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012).

Figure 9.1: *Wordle* word cloud of this Chapter's content

Overview

The SPECS methodology has been applied to evaluate several creative systems. Section 9.1 critically reflects on the success of these implementations of SPECS, considering relevant issues and the overall methodology itself. Section 9.1 looks at various issues surrounding the use of the fourteen components of creativity reported in Chapter 4 (Section 9.1.1), such as training judges to use the components, coping with scenarios where information for a given component was missing or where judges had difficulty rating component(s). Reflections are made on the appropriateness of the components as a base model for creativity. Section 9.1 also considers issues with identifying baseline standards for comparison of evaluation results (Section 9.1.2); concerns about time pressures in the two case studies (Section 9.1.3); a number of issues related to the use of human judges and subjective evaluation data (Section 9.1.4); practical issues in the evaluation process (Section 9.1.5) and more general reflections on using SPECS for evaluation of creativity (Section 9.1.6). Section 9.2 considers how SPECS deals with the adoption of inadequate evaluation criteria or approaches seen in creativity evaluation, as reviewed in Chapter 2.

Another consideration of SPECS is in how it has been received by the community it is primarily intended as a tool for, the computational creativity research community. Section 9.3 considers this, based on the reactions of the research community to date to the work in this thesis and interest that has been expressed in the work, both at research events and in personal communications. From this point, Section 9.4 takes a brief look ahead to suggest how SPECS may be adopted in the future as a standard tool for evaluating creativity of computational systems.

9.1 Critical reflections upon applying the methodology

Much can be learned about a methodology through applying it. In the case studies for this thesis, various points arose, highlighting both SPECS' strengths as a tool and also some weaknesses to be addressed. Section 9.1.6 examines the SPECS methodology itself, independently of how it was implemented in the case studies, in terms of experience gained from applying the methodology for evaluation. The other parts of Section 9.1 evaluate the methodology as implemented in the case studies in Chapters 6 and 7, using the components as the basis for Step 1 (as recommended in Chapter 5). This recommendation itself is examined in Section 9.1.1.

9.1.1 Using the 14 components as a definition of creativity

Looking specifically at how SPECS was implemented in Case Studies 1 and 2, creativity was defined as a combination of the 14 components defined in Chapter 4. The set of components was customised to the individual creative domains either so that components considered more important for creativity

were weighted higher than others, or with analysis focusing on more important components after a more general analysis. Issues surrounding the use of these components will now be examined further.

Training judges to use the components

Judges in both case studies were given details of the components to consult while they were conducting evaluations. Definitions of the components were also outlined to participants verbally at the start of the experiment. In a few cases though, interpretations of the components still differed amongst raters, as shown through verbal feedback during interviews with judges in Case Study 1 and with occasional large differences in ratings for a system's component, for example with one judge giving 2/10 and the other giving 8/10 for *Social Interaction and Communication* in the poetry generator in Rahman and Manurung (2011). Aside from differences of opinion and background knowledge, discrepancies may have arisen because some components still shared small areas of crossover, despite the amount of grouping performed via clustering and inspection of the results in Chapter 4 during the components' derivation.[1] For example, either *Social Interaction and Communication* or *Thinking and Evaluation* components would cover how systems took note of feedback. It was however noted during interviews with judges for Case Study 1 that people accounted for the same factors overall during their evaluations even if they included them in different components; observations were still captured and the components played an important role in guiding the judges to give more detail.

Missing information for evaluation

In Case Study 2, judges had to deal with the scenario where the desired information for evaluation was not provided. Two possible strategies were suggested for tackling this: either to leave the ratings blank or to put in default ratings, such as a middle rating of 5/10. It was thought best to leave ratings blank where information was insufficient to rank a particular component, to avoid confusion with situations where components were given a 5/10 rating, representing an informed view that the system demonstrates average performance on a component. Evaluations were based on what material was available, even if the availability of material was inconsistent across different systems.[2]

The problem of availability of system information was also encountered in Case Study 1. For example, Judge 4 was unsure how GenJam might cope when *Dealing with Uncertainty*. Availability of information also impacted on the choice of systems evaluated for Case Study 1.[3] For Case Study 1, other systems were considered for evaluation, including EarlyBird and other systems by Hodgson (Hodgson, 2006a, 2006b, 2006c) and 'Band out of a Box' (Thom, 2000). Resources for these systems, however, were not as straightforward to obtain as the three systems selected for the case study:[4]

[1] This point will be returned to later in this Section.

[2] As is within the capabilities of people performing evaluation, when full information is not available.

[3] Section 9.1.3 will return to consider this issue and related issues in more depth.

[4] The issue of availability of evaluation materials was also noted for related commercial/proprietary software such as

- Though the development and workings of Hodgson's systems were outlined in publications (Hodgson, 2006b, 2006a) and in his thesis (Hodgson, 2006b), no output examples were available and the source code was not publicly available. As Hodgson currently does not work in academia,[5] his systems were not easily available. Communications about his work relied on finding Hodgson's new contact details and Hodgson being keen to communicate on his old research. Initial enquiries to this end met with no response, so Hodgson's systems were not used for Case Study 1. Some of the work in Hodgson (2006c) was done many years ago (Boden, 1998; Hodgson, 1998; Poole, 1998), so the source code may now be lost, or Hodgson may not remember his systems' functionality in detail.
- Obtaining details of BoB (Thom, 2000) would probably have been easier than obtaining systems by Paul Hodgson, particularly examples and/or source code to run. As Thom currently works in academia, there was more motivation for her to communicate details of her system for research projects, for academic recognition and knowledge sharing. Having said that, as alternative musical improvisation systems existed, with more readily available data, those systems were chosen ahead of BoB for evaluation in Case Study 1.

At a broader level, the time constraints imposed in Case Study 2 enabled some investigation of issues related to formative evaluation: the question of whether systems could be evaluated usefully and accurately with SPECS where only partial information is available; and the usefulness of information that can be gained in a short time. As was highlighted in motivating discussion for Case Study 2 in Chapter 7 Section 7.1, the availability (or lack of) of software and data impacts upon researchers' ability to learn from, develop and improve upon previous research. The second case study illustrates an alternative set of experiments implementing SPECS that have been carried out to obtain feedback on the creativity of systems in situations where there are time pressures (as discussed above) and/or missing information. Though such situations could also be evaluated using most other evaluation methods tested in Chapter 8, it was not possible to evaluate all systems for Case Study 2 using Ritchie's criteria (Ritchie, 2007) due to missing crucial information on system outputs.[6]

As considered in Chapter 7 Section 7.1, sometimes systems may be particularly relevant to current research interests but may not be fully available. If there is partial information available on the system, then some findings can still be collected from evaluation, as Case Study 2 has demonstrated. This was also demonstrated to some extent in Case Study 1, where not all the systems were available to be

'Band-in-a-Box' (`http://www.pgmusic.com`).

[5] I believe Paul Hodgson may be considering a return to academic work (Torrance, 2012, personal communications), so his previous systems may be easier to obtain in the future.

[6] Although only partial evaluations could be performed using the creative tripod (Colton, 2008b) and SPECS, both these methods still could be carried out to generate evaluative feedback, whereas lack of output examples meant that Ritchie's criteria could not be calculated at all.

run and tested during the study; for example GenJam was written for legacy hardware (that was not available for the case study) and had dependencies on supporting software which has since become obsolete. Findings from both case studies, and especially Case Study 2, may be more limited than we would ideally like.[7] When considering the results of evaluation, such limitations should be borne in mind, to put the feedback into context. The case study findings show, however, that we can still collect information that helps us in future research, learning from what has been done in previous research (particularly for very relevant or significant systems). Learning from previous work can help contribute useful information to develop the research progress of both the individual researcher(s) and their research area as a whole. Such contributions are fundamental to the development of knowledge should not be overlooked.

Difficulties in rating some components

For both case studies and particularly in Case Study 1, judges reported that some ratings were more difficult to give than others. As mentioned previously, the acknowledged issues with defining a number of components of creativity (Chapters 3 and 4) meant that the meanings related to the components necessarily had some crossover and overlap. For practicality in minimising the number of components where possible, some components incorporated a spectrum of related points. Judges questioned the definitions of some of these components, for example Judge 5 debated whether *Originality* should be equated with randomness, or given another definition. The same judge also considered different interpretations of *Progression and Development*, in the context of individual agents within a system developing or the whole system developing as an entity. Judge 1 asked how *Progression and Development* differed from *Independence and Freedom*.

As the components were derived using papers about human creativity as well as computational creativity, some components had a particularly human-like quality. Difficulties were noted by judges in applying some of these aspects of creativity to computational systems, particularly for *Spontaneity and Subconscious Processing* and *Intention and Emotional Involvement*. As investigated in Chapter 3, in this work computational creativity was generally treated as computational activity which would be considered creative if seen in humans,[8] matching the predominant interpretation in computational creativity research (Chapter 3 Section 3.4.1). As computational creativity has been linked to human creativity in this way, the inclusion of components that focused on human-like creative characteristics could be justified.

To analyse how simple the judges found it to provide ratings of each of the components in Case Study 1, timings were extracted from the recordings of the interviews conducted during the evaluation

[7] This was also relevant when considering the amount of time available for evaluating research, as will be discussed later in this Chapter in Section 9.1.3.

[8] This was debated during discussion at ICCC'11; further details are in Section 9.3.

study (Chapter 6 Section 6.3.5). The time taken for each judge to rate each component was recorded, timed from the start of the discussion on that component to the point where the judge gave their final rating for that component. Judges were encouraged to give their ratings as quickly as possible, but to feel confident about their ratings.

This timing metric was based on the premise that the longer that the judge took to think about and give their rating, the more complex they found that component to rate. While other possible factors could have delayed the judge's response, for example fatigue or boredom during the experiment, loss of concentration, external interruptions, etc., efforts were made during interview to minimise these factors by attempting to keep the interview stage fun and relaxed, providing progress checks such as '4 components left to rate' and removing external distractions wherever possible.

Table 9.1: Data on response times for judges' ratings of components for Case Study 1 (given to 1dp).

Component	Mean response time (seconds)	Standard deviation
Social Interaction and Communication	137.3	123.0
Domain Competence	92.0	64.9
Intention and Emotional Involvement	90.6	51.6
Active Involvement and Persistence	115.5	98.3
Variety, Divergence and Experimentation	108.8	58.3
Dealing with Uncertainty	84.5	50.9
Originality	94.2	53.5
Spontaneity and Subconscious Processing	119.0	87.3
Independence and Freedom	91.3	50.5
Progression and Development	81.1	39.0
Thinking and Evaluation	73.4	39.8
Value	87.2	57.4
Generation of Results	69.1	36.4
General Intellect	47.5	24.8

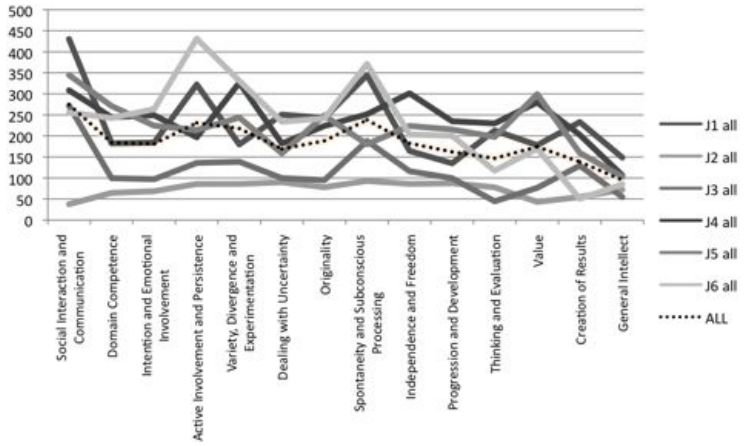
Here, two projections of the timing data are examined. Table 9.1 shows the timing data obtained and Figures 9.2(a) and 9.2(b) show how a judge's time to give a rating varies between components. Table 9.1 and Figures 9.2(a) and 9.2(b) list components in the order they were presented to judges. In interview, components were presented in order of descending contribution for this type of creativity, as identified in Chapter 6 Section 6.3.2. This was in anticipation of judges becoming fatigued during the experiment; components are likely to receive most focused attention if dealt with earlier in the list of 14 components. As the mean data for each component shows, the time taken to judge components generally decreases as later components are reached. Potentially this could have been because the latter components were simply easier to rate, but it is more likely that the number of components introduced a level of fatigue. The standard deviation also decreased in a similar way, showing that

over the course of the experiment, the variability of time taken between judges reduced.[9]

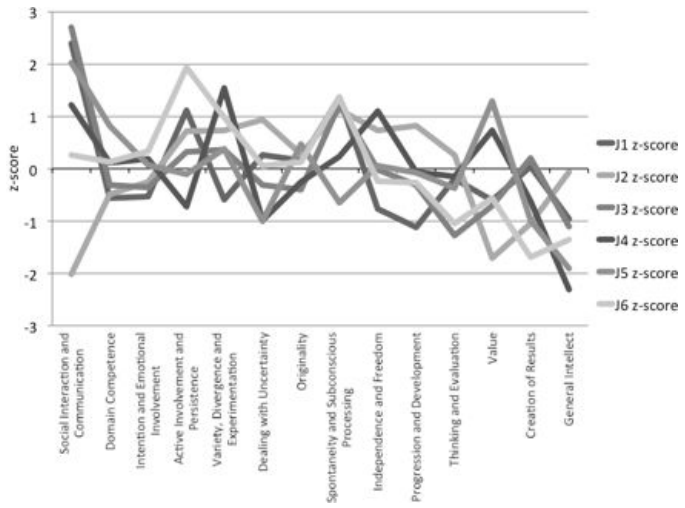
Figure 9.2(a) shows how some components went against this general trend, standing out in the dotted line in the figure that represents the mean of all individual ratings. These components were: *Active Involvement and Persistence*, *Spontaneity and Subconscious Processing*, *Originality* and *Value*. Reflecting on the interviews with judges, different reasons are posited for this:

- With *Active Involvement and Persistence* the judges tended not to have considered how the system would work if it had encountered problems, partly because reports and papers on a system tended to present what happens when the system worked correctly, rather than when it ran into difficulties. To address this component, the judges needed to re-consider the systems from a new perspective.
- Judges had difficulty applying *Spontaneity and Subconscious Processing* to the systems, partly due to its human-like nature (which also affected judges to a lesser degree in rating *Intention and Emotional Involvement*) and partly because judges did not always comprehend how these two aspects fitted together. To some extent this reflects the less-than-natural clustering of these two aspects of creativity, which could perhaps be clustered differently. Together the two parts of this component represent the spontaneous 'aha' moment of *inspiration* in creativity that comes after a period of *incubation* or subconscious processing of information (Poincaré, 1929; Wallas, 1945, Chapter 3 Section 3.4.2). This combination emerged in the 'creativity words' identified in Chapter 4; while the combination would seem appropriate to those who are familiar with the afore-mentioned theory, it may appear somewhat arbitrary to those who are not.
- *Originality* has often been strongly linked with creativity (Chapter 3 Section 3.4.1), so it was somewhat surprising that judges occasionally found this component difficult to use to rate the systems. In general judges were reluctant to attribute great originality to an improvisation if they were not so familiar with examples of that particular musical style, or if they felt that there were many examples of musical improvisation in that style that they had not encountered. Subconsciously, most judges took an interpretation of originality more closely resembling *h-creativity* rather than *p-creativity* (Boden, 2004, Chapter 3 Section 3.6.3). That is, judges decided if an artefact was original compared to all the artefacts of that type that could exist (c.f. h-creativity), rather than compared to all the artefacts of that type that the judge had encountered (c.f. p-creativity).
- Usually a longer length of time represented difficulty in rating a component, but one exception to this generalisation was with the *Value* component. Generally judges spoke a lot about different value that they saw within the system, or why they liked or disliked the system and its

[9]The set of components is reflected upon further in this Section and also in Chapter 10 Section 10.3.



(a) Combined timings for each judge: how long (secs) each judge took to rate a component for their first system, plus time taken for that component for their second system. The dotted line shows the mean time taken for all individual (i.e. not combined) ratings per component. *N.B. The data includes the timings for the fourth system (later removed from Case Study 1 - see Chapter 6 Section 6.2), as the reason for its removal (incorrect samples used) would not largely affect the timings data, but removing timings data for that system could introduce discrepancies and bias due to how many systems each judge evaluated or in which order systems were evaluated.*



(b) Z-scores of the combined timings data from Figure 9.2(a). Variability in ratings for each judge is shown by standardising the Figure 9.2(a) ratings: subtracting the mean of that judge's timings and dividing by std. dev. A component with z-score > 0 took relatively longer for that judge to rate and vice versa. *Again this includes the ratings for the system removed from this case study to avoid introducing ordering bias or discrepancies between judges' data.*

Figure 9.2: Graphical representations of Case Study 1 judges' response times.

output, before giving a rating. In the other components in this list, judges tended to pause more and have large gaps between points whilst they thought, but that was not observed for *Value*; judges generally had more to say and seemed to be more comfortable rating this component. One could conjecture that this was because people might expect to be asked more about the value of what they are rating, when evaluating[10] something, as opposed to questions relating to many of the other components.

Appropriateness of the use of customised components as a definition of creativity

As per the requirements for SPECS, it was clearly stated that the components from Chapter 4, were used (hence following the recommendations in Chapter 5). This transparency and informativeness about the evaluation performed makes the evaluation more analysable and repeatable[11] for new systems. But how appropriate did the set of 14 components prove to be for evaluation?

In Case Study 2, no components were rated as *Not at all important*. Similarly, in Case Study 1, every component was mentioned in some way at least a few times during the information gathering stage for Step 1b. So all the components proved to be relevant for creativity to at least some degree.

9.1.2 Baseline standards for comparison

It is unclear what could be used as a standard or baseline to compare the individual performance of systems and collective performance of all systems. In creativity evaluation, finding a ground truth to measure the correctness of evaluations is difficult, if not a fruitless task, given the subjectivity and general ambiguities involved in creativity (Chapter 3).[12] Chapter 8 explored other creativity evaluation methodologies (Ritchie, 2007; Colton, 2008b) and comparisons with human opinion.

In the human opinion surveys, several people used human performance as a baseline to measure against. Questions were raised through this about the fairness of comparison given the relative performance of humans to computers on several of the examined tasks and the delivery of results (for example an excellent computer improvisation could sound very mechanical if performed using MIDI). The surveys also found inconsistencies as to whether people compared the systems to human experts in that domain, or people with some competence, or novices in that domain.

9.1.3 Time pressures in the Case Studies

In both case studies, judges had a limit placed on the time they had available to learn about the systems before evaluating them. The time pressure was greater in Case Study 2, where judges had

[10] Particularly because the word 'evaluate' incorporates most of the word 'value'.

[11] Another system has since been offered for evaluation by Spector, who was curious how another system (Spector & Alpern, 1994) would evaluate compared to the other Case Study 1 systems (Spector, 2011, personal communications).

[12] These are issues that the SPECS methodology tackles.

seven minutes to learn about a system. It could also be questioned whether the time period of 30 minutes per system in Case Study 1 was long enough to learn enough about each system to provide appropriate feedback; these points will be considered below.

Ethics issues As Chapter 2 (especially Section 2.3) has shown, previous computational creativity evaluation literature has tended to overlook the question of who should evaluate a system. A resulting recommendation in Chapter 5 was that the developer(s) of a creative system should not be the sole evaluators of that system, to help avoid any biases affecting the accuracy of evaluation and to enable more independent evaluations to be collected (McAllister, 2012, personal communications).

If participants are involved in evaluation, the issues surrounding participant ethics must not be overlooked, particularly those regarding the duration of evaluation studies and experiments. As mentioned in Chapter 5 Section 5.1.4, issues arise around financial compensation for participants (for example, what budget is available, or is a fair rate of recompense being offered) and/or the reliance on goodwill to ask people to evaluate a system without being reimbursed for their time. In both scenarios (especially in the latter case), the amount of time one can ask participants to spend on evaluation should be carefully considered. There may be a negative effect on participant recruitment for longer studies (particularly if not being reimbursed for their time).[13] In the UK Civil Service's ethical guidelines for research,[14] researchers are advised to 'avoid placing an unnecessary burden on respondents (p. 9), and that '[t]he potential impact of choices in research design (such as sample design, data collection method and so on) on participation should be considered. ... Consideration should be given to issues likely to act as a barrier to participation, and reasonable steps taken to address these' (p. 11). The guidelines also recommend 'avoiding unnecessarily long interviews' to protect the well-being of participants (p. 12). Research councils and other funding bodies issue advice for cases where participants' well-being might be affected. For example the ESRC guidelines[15] recommend asking 'Will the study involve prolonged or repetitive testing?' as part of the process of experiment design and ethics approval. A further effect of longer studies is that participant fatigue would be likely to increase, potentially leading to poorer results as participants become bored, tired or distracted. Hence the allocation of participants' time is an important factor to balance.

Resource allocation in research Further time-related issues arise regarding the allocation of researchers' time taken up in preparing and performing evaluation, and collating/analysing results:[16]

[13] This point was raised during personal communications (2012) with Hilary Ougham, Academic Research Officer at the University of Brighton, whose responsibilities include advising on ethics issues in research.

[14] 'Ethical Assurance for Social Research in Government', available at *http://www.civilservice.gov.uk/wp-content/uploads/2011/09/ethics_guidance_tcm65782.pdf* (last accessed November 2012).

[15] The Economic and Social Research Council, a key funding body in the UK, issues comprehensive guidelines regarding research ethics. These are published at *http://www.esrc.ac.uk/aboutesrc/information/researchethics.aspx* (revised 2012) (last accessed November 2012).

[16] This was also discussed in the motivations for Case Study 2 (Chapter 7 Section 7.1.

- How much time is allocated to evaluation from the research budget (for funded research)?
- How much time does the researcher(s) have available, and how many hours should be spent on evaluation instead of other tasks that are important for the research project or for the researcher's overall work responsibilities?
- Are there forthcoming deadlines for conference or publication submissions, or forthcoming events where evaluation results should be presented?
- In a larger project,[17] are time pressures introduced by other phases of the wider project scope that require tasks associated with the creative system to be completed before those other phases can be worked upon?

Unfortunately, researchers often cannot use as much time as they would like to perform tasks such as evaluation. Several factors affect the amount of time that can reasonably be spent on these tasks. Availability of people involved in the evaluation is a valuable resource, to be spent wisely.

Considering time issue in Case Study 1 In Case Study 1, the participants were given 30 minutes to research each of the two systems they were evaluating. The whole evaluation process took roughly 2 hours per participant, in one session, including time for interviews straight after the research period to collect ratings for that systems, plus a break between systems to allow the participants to be refreshed before the second system. Participants were contributing to this study on a goodwill basis and were not reimbursed financially. As participants were also required to have knowledge in both musical improvisation and computer music/computational creativity, there were already a number of difficulties in recruiting enough suitable participants. To give the participants more time, or to ask them to evaluate more systems, two alternative options were considered: asking participants for more than two hours of their time or asking participants to attend a second session at another time. It was decided, however, to structure the case study evaluations as reported so that the evaluations would happen in the same session. An advantage of this study design is that participants were more likely to remember the evaluation criteria across different systems and to apply the criteria more consistently without becoming too fatigued over a longer period of time.[18]

Is thirty minutes enough time to learn about a system deeply enough to be able to evaluate it? Personal communications (2012) with researchers who specialise in software usability testing[19] have highlighted that participants need enough time to get a representative impression of the system's capabilities; a key factor in gauging this amount of time is whether more time would allow the participants to access content novel enough to change their first impressions. When participants are unlikely to

[17] Larger projects involving computational creativity will hopefully occur, given a recent inclusion of computational creativity in the European Union FP7 framework call for research funding, as mentioned in Chapter 1 Section 1.5.1.

[18] These decisions have been discussed with a number of people, including Blay Whitby, who has had expertise in research ethics and responsibility for ethics clearance for research at the University of Sussex.

[19] Particularly video game testing, through the *Vertical Slice* company, based in Brighton.

experience anything substantially different about a system in further research time, they do not need to spend more time learning about the system. Case Study 1 participants were always asked if they were ready to conclude the research period and to provide evaluation ratings after 20 minutes had elapsed, or if they would like more time. Though this can be difficult to anticipate and self-monitor, in all cases, participants reported being ready to continue to the evaluation stage and no participants requested more time for research.

Considering time issue in Case Study 2 The time constraints imposed in Case Study 2 were partially inspired by the decision of the organising committee of the 2011 International Conference on Computational Creativity (ICCC'11) to impose a time limit of seven minutes per talk at this conference, where the Case Study 2 SPECS evaluations took place (Chapter 7 Section 7.2) This interesting experiment by the ICCC'11 committee raised a relevant question in the context of this thesis, namely: what can be learnt about a system's creativity in this amount of time? Case Study 2 was partly designed to address that question and partly designed to demonstrate a shorter and less involved application of SPECS compared to Case Study 1, allowing for scenarios where the amount of time available for evaluation of systems is more limited than for the Case Study 1 evaluation, or where system developers wish to get a quick 'snapshot' impression of a system's creativity during development, not wishing to invest too much time in evaluating a partially developed system at this stage but seeking some formative feedback that can be gained from first impressions. Under these time constraints, the amount of information that can be imparted about the system is limited, and there is a risk that incorrect or inaccurate impressions can be formed if the system is not presented well. A skilful presentation of the system, however, can allow researchers to extract useful formative feedback with a shortened time investment, which may prove invaluable at crucial stages of development.[20] An evaluation measuring the participant's first impressions of the creativity of the programs may produce results that differ somewhat from their long term impressions; however these results would still be important if their first impressions were strong enough to form judgements.[21]

9.1.4 Using judges to provide subjective ratings

Some issues were introduced by the decision to use human judgement for evaluation: the subjectivity of the judgements provided, particularly if influenced by biases about computational creativity, the choice of the rating scale and the accuracy and consistency of ratings from different judges.

[20] The Case Study 2 results can in fact be a useful guide here, to help the researcher maximise the information delivered about their system's creativity in a short amount of time.

[21] From personal communications (2012) with Gareth White and other members of *Vertical Slice*, a video game usability testing company based in Brighton.

Using subjective ratings

Both case study evaluations relied on subjective ratings. To some extent, having more than one judge smoothed out any inconsistencies, however to obtain evaluation data for the methodology, practical considerations limited the number of judges to 6 in Case Study 1 and 2 in Case Study 2.

Difficulties in finding appropriate judges reflected more on the decisions made in how to apply SPECS, rather than a reflection on the methodology itself. As it stands, SPECS makes no demands on how standards for creativity should be tested, merely requiring that these standards are justified and clearly stated. A future extension to the work in this thesis would be to explore several other ways in which systems could be tested. One intriguing way of implementing SPECS would be to use empirical tests where possible, for objective measurements rather than subjective opinions based on short acquaintance with the systems. Exploring such metrics, one would need to be careful that the metrics did indeed measure the intended component (or part of that component, supplemented with other tests) and that important aspects were not overlooked. It may be necessary to use domain-specific rather than general tests, for example recording the number of 'licks'[22] a musical improvisation system 'knows', through examination of a licks database such as that which GenJam uses[23] (Biles, 2007). Also, if the computational creativity system is producing artefacts for a human audience, it becomes difficult to justify a choice of evaluation not including judgement by human standards, particularly as those standards may change over time and according to context.

Chapter 5 Section 5.1.7 discussed these issues, concluding that both quantitative and qualitative tests should be incorporated in evaluation, with some triangulation of results for composite feedback. Whilst both quantitative and qualitative feedback were used in evaluation for Case Studies 1 and 2, the question of what tests to include for Step 3 of SPECS invites further attention in future research. In particular, in seeking standardisable and flexible tests, this question would probably benefit from multiple examinations by several researchers, on different creative systems.

Biases about computational creativity

All judges were aware that the systems they were evaluating were computer-based. In the surveys in Sections 8.1.1 and 8.2.1, participants were told that they were evaluating computer systems that improvised music. Selection criteria for judges for both case studies was designed to minimise any negative biases. Judges for Case Study 1 had studied about creative systems at least at undergraduate degree level. Judges for Case Study 2 were computational creativity researchers.

The questions at the end of the Section 8.1.1 survey looked at participants' attitudes towards computational creativity. Generally most thought that computational creativity was possible, although

[22] Licks, as mentioned in Chapter 6 when describing the GenJam system, are short phrases which can be used to partially construct an improvisation. Licks are most commonly used in jazz and similar styles.

[23] See GenJam's description in Chapter 6 Section 6.2.

opinion was divided as to whether computers were already demonstrating creativity or whether this was for the future. Most liked the idea of computational creativity, finding it exciting rather than shocking or disturbing. Having said that, it is possible that judges' ratings could have been subconsciously influenced by the knowledge that they were evaluating computational rather than human creativeness (Moffat & Kelly, 2006); certainly some negative bias was voiced in the comparative evaluation surveys in Sections 8.1.1 and 8.2.1.[24]

Choice of rating scale

A rating scale of 0 to 10 was chosen because it was felt that people would usually be familiar with an x-out-of-ten scale. There are differences in how people may choose to rate systems out of 10. Some may be happy to use the full range 0-10, others may prefer a more cautious approach of using the middle of the range and avoiding outliers. The connotations of giving something a score of 10/10 implies absolute perfection that cannot be surpassed (with an opposite implication from a 0/10 rating).

When judges are required to give ratings in evaluation, an alternative would be to give them a Likert scale. With Likert scales, judges choose a rating based on looking at several descriptive labels, ordered by an underlying scale from low to high, and select the most appropriate option for what they want to express. The 0-10 rating scale used in Case Studies 1 and 2, though, was essentially a 21-point Likert scale (as .5 scores were permitted), which gave the judges far more options to choose from without overwhelming them with unfamiliar options, or asking the judges to match their opinions to descriptions of an ordinal scale.

Accuracy and consistency of judges' ratings

The accuracy and consistency of the judges' evaluations may have varied, for several reasons. Judges may have had differing interpretations of creativity in musical improvisation (although Amabile (1996) bases her Consensual Assessment Technique on the assumption that experts in a given domain have similar interpretations of creativity in that domain). Judges may also have become fatigued during the experiment or lost interest, as the study lasted 2 hours (though much effort was made to prevent boredom during interviews). They may also not have fully understood the components.[25]

Figure 9.3 shows how judges made different use of the ratings scale in the case studies.

- In Case Study 1, judges' ratings may have been affected to some degree by the quality of the systems they rated, as they did not all rate the same systems. It is feasible that a judge rating two systems which were ranked lower overall would use lower ratings over all the components than a judge rating two systems which were ranked higher overall.[26]

[24] Section 8.3.3 gave some examples and also discussed questions of bias towards computational creativity in more detail using feedback from forums, social networking and case study surveys as evidence of this effect during evaluation.

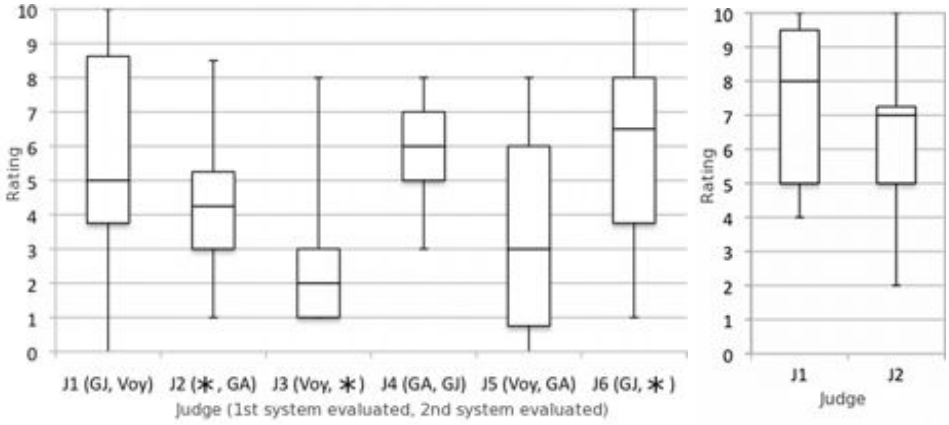
[25] This will be discussed further in Section 9.1.1.

[26] The mean unweighted ratings were 3.6/10 (GAmprovising), 7.0/10 (GenJam) and 4.3/10 (Voyager) (Chapter 6 Table

- In Case Study 2, judges' ratings could have been affected by how many ratings each judge supplied; out of a possible total of 70 (14 components, 5 systems), Judge 1 supplied 27 ratings and left 43 components unrated. Judge 2 supplied 44 ratings, leaving 26 components unrated.

Some Case Study 1 judges (Figure 9.3(a)) clearly made more extensive use of full range of ratings than others, e.g. Judge 1 used ratings from 0-10 while Judge 4's ratings were restricted to between 3-8. Judge 3's ratings tended to be the most negative and Judge 4's ratings the most positive. Judges 3 and 4 were also most likely to keep within a small range of ratings, compared to Judges 5, 1 or 6.

Within Case Study 2 (Figure 9.3(b)), neither judge made full use of the 0 - 10 scale, with minimum ratings given of 4 (Judge 1) and 2 (Judge 2), although both judges gave maximum ratings of 10 in some cases. In general Judge 1's ratings were wider ranging and more positive than Judge 2. Judge 2's ratings were negatively skewed, meaning that most ratings were higher than the mean (6.3), whereas Judge 1's ratings were more evenly distributed (with a slight negative skew; more ratings were higher than the mean of 7.5 than if the data had no skew).



(a) Case Study 1. (b) Case Study 2.

Figure 9.3: Box and whisker plot showing the use of the 0-10 rating scale by each judge in the case studies, across all their ratings combined. These plots show the maximum, upper quartile, median, lower quartile and minimum ratings given by each judge. For Case Study 1, GA = GAmprovising, GJ = GenJam and Voy = Voyager. ** represents the fourth system that was later removed from this case study (see Chapter 6 Section 6.2). This fourth system is included here to give the full picture of the judges' collective overall ratings behaviour, as was also done in Figures 9.2(a) and 9.2(b).*

Inter-judge variation is to be expected, as people use rating schemes differently. Such variation can cause inconsistency when a small number of judges are used, as was the case in both case studies. In the case studies, issues such as this were anticipated but it was felt the approach adopted in the case studies would also generate useful feedback on the effectiveness of the actual SPECS methodology

[6.3).

and its implementation (particularly in the time-intensive but flexible and information-rich interview style adopted for Case Study 1). It is acknowledged, though, that the provision of data from more judges could be preferable, particularly in future situations where there is no especial need to seek feedback on the methodology itself. Section 9.1.4 of this Chapter and Chapter 10 Section 10.3 discuss how automation of tests and the use of empirical rather than subjective evaluation may also help here.

To analyse how the judges' evaluation ratings matched their overall opinion of the creativity of the two systems, at the end of the experiment in Case Study 1 judges were asked which of the two systems that they had evaluated was more creative in their opinion, giving reasons why. The judges' preferences were compared to the sum of the ratings given, both before and after weighting the ratings, to see if judges' ratings were higher overall for the system that they considered more creative. Such analysis highlights several aspects:

- If judges' ratings reflected their impression of the creativity of the systems.
- If a judge gave the system they considered more creative lower ratings than the system they considered less creative.
- If judges' priorities for musical improvisational creativity matched the weighted components.
- If weighting the ratings makes a difference on the above points.

Table 9.2: Data on how the judges' component evaluations collectively matched their overall impressions of the creativity of Case Study 1 systems. *As before, the fourth system, now removed from this case study, is included in these results as an asterisk *, for a fuller view of judges' opinions.*

Weighted ratings			
Judge	More creative	Less creative	Difference
J1	GenJam (74.5)	Voyager (43.1)	31.4
J2	* (53.7)	GAmprovising (34.3)	19.4
J3	Voyager (43)	* (18)	25
J4	GAmprovising (57.4)	GenJam (63.7)	**-6.3**
J5	Voyager (51.5)	GAmprovising (10.9)	40.6
J6	GenJam (73.6)	* (34.3)	39.3
Unweighted ratings			
Judge	More creative	Less creative	Difference
J1	GenJam (104.5)	Voyager (57.5)	47
J2	* (75.5)	GAmprovising (51)	24.5
J3	Voyager (50)	* (27)	23
J4	GAmprovising (83.5)	GenJam (87)	**-3.5**
J5	Voyager (71)	GAmprovising (17)	54
J6	GenJam (103)	* (57)	46

This data gives evidence to consider both the reliability and accuracy of the methodology and also

the reliability and accuracy of the judges, as reported in Table 9.2. The results in Table 9.2 show that for five out of six judges, the system that they preferred also received higher ratings, with a difference of between 23 and 47 for the original ratings and between 19 and 41 for the weighted ratings. The slightly higher differences between original ratings, compared to weighted ratings, suggest that weighting the ratings does reduce the effect of less important components on the overall value.

Judge 4, who evaluated GAmprovising and GenJam, thought GAmprovising was more creative than GenJam, but overall rated GenJam higher both in their original ratings and when the ratings were weighted according to importance. Though the difference was small (-3.5 for the original ratings, -6.3 for the weighted ratings) compared to the differences recorded in other judges' ratings, this anomaly should be investigated further. There are two possible explanations for this small anomaly:

- Judge 4's opinion of what was important for creativity was at variance with the other judges and with the priorities identified in Chapter 6 Section 6.3.2.
- The components (weighted or unweighted) do not align with this judge's overall impressions.

Given that the other 5 judges' summed ratings matched their overall impressions, both when weighted and unweighted, it is suggested that Judge 4's interpretation of creativity differed from other judges and from the questionnaire findings of Chapter 6 Section 6.3.2. In their reasons for finding GAmprovising more creative than GenJam, Judge 4 mentioned in particular how GAmprovising had a large amount of freedom and still seemed to refer back to itself. This relates to the *Independence and Freedom* and *Thinking and Evaluation* components, weighted as having a 5.4% and 5.1% (respectively) contribution to the overall creativity of the system.

To investigate this discrepancy in judges' opinions of creativity, it may be worth relaxing the constraints on who qualifies as a computer musical improvisation expert, in an effort to recruit more judges to compare between. Should more judges be recruited, it would be preferable to find a shorter way to deliver the information about the systems, or delivering information to several judges at once, such as in a lecture or class. For the purposes of this research, though, it was thought important to use judges who had expertise in both computer music and musical improvisation, so that elements of bias against computer music creativity systems could be removed and the judges could base their evaluations on solid knowledge both of computer music and musical improvisation.

9.1.5 Practical concerns in evaluation practice

The case studies showed that the methodology can be applied to systems in several different domains. Even within Case Study 1, systems supposedly in similar domains produced stylistically different improvisations, ranging from avant-garde to jazz standards. The systems were not designed to meet the same aims; for example Voyager and GenJam interact with other musicians in real time during live

improvisation, whereas GAmprovising does not. The flexibility of the SPECS methodology allowed it to be applied across this variety of creative systems.

Regarding the utility of the evaluative comparisons made in the case studies, whilst beneficial information can be learnt from making comparisons, especially in Case Study 1, only limited benefit could be drawn from comparing systems from the different domains in Case Study 2.[27] If the evaluative aim was to summatively measure overall creativity of a system, perhaps in some competitive way, a single evaluative score could be obtained, for example through dimensionality reduction as performed in Sauro and Kindlund (2005). For Case Study 1 it would also be possible to sum the weighted component ratings together for a single score.[28] For the purposes of summative evaluation and quantitative comparison, though, the comparison results in Case Study 2 show limited benefit in comparing system performances competitively. In ICCC'11 discussions and elsewhere (Cheng, 2011, personal communications), the point was raised that chasing a small numeric increase in a single evaluative score can be somewhat tedious and unproductive. Where benefits can be gained from cross-domain system comparisons is from formative evaluation and constructive feedback. Information on how one system excels in one area can often be applied to improve other systems; for example the interactivity in GenJam and Voyager could inspire developments in a system that generates images.

Reflections on the ability of the SPECS methodology to adapt over time are appropriate here. How can SPECS adapt over time to changes in evaluation standards or in contextual aesthetics? Perceptions of creativity can change over time and can be context-sensitive. The key to SPECS' particular robustness here is that the standards for creativity must be clearly stated in SPECS. If priorities for creativity in a given domain change over time, then previous evaluations are not necessarily nullified but can be adapted, provided evaluative tests for each standard are clearly stated.[29]

One practical issue in applying the components for evaluation was that there were 14 components to evaluate; a significant number to consider. It may be that some components do not contribute to creativity in a particular domain at all, though this was not found in the results for the two case studies in this thesis. To some extent, though, evaluation could be concentrated on those components which are identified as most important, with the acknowledgement that some more minor information will be omitted from the evaluation.[30] Alternatively, a dimensionality reduction technique such as principal components analysis could be used to identify which combinations of components are most critical, as in Sauro and Kindlund (2005) or Nielsen (1994).

[27] Section 9.3 will report discussions at ICCC'11 on what extent similar-enough systems can be found for comparison to each other, and what can be gained from the comparisons.

[28] Such collation has very deliberately not been performed in the case studies, to emphasise and retain the focus on the individual component feedback generated during the SPECS evaluation.

[29] One reviewer of my ICCC'11 conference paper on this work (Jordanous, 2011a) noted that the methodological recommendations made for clarity of evaluation could also be applied to how to write a good paper on creative systems.

[30] Though Section 8.2.4 showed that reducing the components to a domain-general subset is not always appropriate.

9.1.6 Reflections on applying SPECS for evaluation

It should be noted that all of the evaluation methods in Chapter 8, and more, could be applied within the framework of SPECS. The case studies in this thesis follow the recommendations from Chapter 5 to use the Chapter 4 components as the base model of creativity, customised towards the domain in question. A system evaluator could however choose to use Colton's creative tripod framework (Colton, 2008b) or Ritchie's empirical criteria framework (Ritchie, 2007) as the adopted definition of creativity, if this decision is stated and justified by the evaluator. A definition of creativity may also be based around 'what people find creative', leading to the consultation of human opinion as the evaluation method chosen.

Putting exact implementation matters aside for the moment, a key message from using SPECS is the importance of being clear about what creativity is in the domain being examined, adopting an appropriate definition and evaluating creativity based on testing standards derived from that definition. Following the SPECS methodological steps developed from the recommendations advocated in the *Evaluation Guidelines* (Chapter 5) allows computational creativity researchers to perform evaluation of creativity of their systems in a standardised and appropriate manner.

In the future it would be interesting to see how and if other researchers use SPECS, in particular:

- How others choose to apply the SPECS methodology (and whether they choose to adopt the Chapter 4 components for evaluation, as recommended in Chapter 5).
- How easy others find the SPECS methodology to use.

This could be done either by circulating the methodology to the research community and seeing how many people choose to use it and in what way, as was reported in Ritchie (2007). An alternative would be to directly recruit some researchers to apply the methodology to their systems and monitor their progress in doing so. While there are some non-trivial practical issues with the second option, it would give more information overall, but the first option would illustrate adoption of the methodology more naturally. Either way, useful future work can be done with others applying SPECS.[31]

9.2 How SPECS deals with inadequacies in existing evaluation approaches and criteria

The literature review in Chapter 2 and the survey of current evaluation practice in computational creativity (Chapter 2 Section 2.3) highlighted various inadequacies in how the computational creativity deals with creativity evaluation. Methods used in current creativity evaluation practice in computational creativity have been discussed throughout Chapter 2 Sections 2.1, 2.3.2 and Section 2.3.4 as

[31] Some computational creativity researchers have already been in contact wishing to apply SPECS to evaluate their creative systems. This and other avenues for future work, are explored in Chapter 10 Section 10.4.

well as being returned to from a different perspective in Chapter 5 Section 5.3.3. Of particular interest when reflecting on the SPECS methodology is Section 2.3.4 of Chapter 2, where various evaluation criteria were highlighted that were inadequate for creativity evaluation (or that were notably useful and could be learned from). This current Section considers how SPECS would deal with the adoption of inadequate evaluation criteria or approaches in creativity evaluation.

Colton (2008b) dictates a model to follow for creativity evaluation that is applicable to all types of creativity in all creative domains, which does not take into account the advances made in understanding how creativity is manifested in different domains (Chapter 3 Section 3.6.4).[32] SPECS instead gives the researcher the freedom - and the responsibility - to choose appropriate creativity evaluation criteria, guiding the evaluator through how to develop such criteria (Steps 1a, 1b and 2) based on what it means for that system to be creative, in its creative domain. In SPECS, evaluators are required to state their evaluation criteria clearly and base their tests upon these criteria, such that their evaluative approach can be learned from, critiqued and/or repeated by researchers working on systems operating in similar domains. Evaluators are guided towards adopting appropriate evaluation standards for computational creativity and are required to justify their choices; therefore evaluators are held to account for their chosen standards and their choices and justifications can be moderated by the wider research community, in a manner similar to peer review.[33] Hence it is possible for researchers to adopt inadequate evaluation criteria, but they are required by SPECS to show *how* they have derived these criteria from an understanding of creativity as it is demonstrated in the creative area they are working in. The requirement to state all parts of this process and employ transparent evaluation methods means that weaker criteria or weaker justifications for adopting criteria can easily be identified and critiqued, from a broader range of perspectives and contexts. It is also hoped that due to the greater rigour required in this process by SPECS, there is a reduced likelihood of researchers adopting inadequate criteria, and that researchers instead spend more time grounding what their research has achieved in a creative context. This should help to enhance the credibility of computational creativity research, for making impactful contributions on a wider scale for accumulation of knowledge about creativity.

Looking at specific points raised in Chapter 2, the adoption of a more rigorous, systematic, stan-

[32] The way that Ritchie's criteria have generally been implemented for creativity evaluation is another example where a set of criteria are advocated regardless of the domain. The use of parameters and thresholds in the formal presentation of the criteria in Ritchie (2007) and accompanying text shows Ritchie's intentions for the criteria to be customised to best reflect some model of creativity before application for creativity evaluation; however this point has been overlooked in their application. As Chapter 2 Section 2.1.2 reflects, this is probably at least partially due to the lack of example evaluations carried out by Ritchie. In this thesis, Chapters 6 and 7 offer example evaluations of SPECS. Also, as some participants in ICCC'11 discussions commented (Chapter 2 Section 2.3.5), an issue in constructing formal models is how to best balance a broader coverage with more specific details; the formal and abstract nature of Ritchie's model makes it less obvious how the model's aspects should be related more directly to instances of creativity.

[33] As found in the opinion survey for Case Study 1 (Chapter 8 Section 8.1.1, even if relying on people's subjective opinion to evaluate the creativity of a system or systems, people can find it difficult to articulate an evaluation of creativity without having some definition of creativity to refer to for assistance.

dardised and transparent evaluation procedure across the field will help avoid situations where evaluation of creativity is not carried out at all, or is carried out in an ad-hoc, individualistic way that has not been justified. A requirement of SPECS is to justify the standards being used for evaluation in a context of creativity, which should help prevent the future adoption of criteria or methods purportedly to evaluate creativity that in fact do not evaluate the creativity of the system, but instead evaluate its ability to generate good quality artefacts (for example artefacts that are appropriate, correct and/or aesthetically pleasing).[34] In particular, SPECS requires that the researcher states how their creative system would manifest creativity by first reflecting upon creativity in general and then reviewing specific requirements for creativity as manifested in their particular domain, following the lead of relevant creativity research (see Chapter 3 Section 3.6.4). Using this approach will discourage the adoption of standards for creativity evaluation which are overly specific to that creative domain, or to the system in question, without considering more general implications of what it means to be creative. Similarly, the adoption of overly general evaluation standards that do not take into account specific areas of importance for creativity in that domain is also discouraged. SPECS ensures that both domain-general and domain-specific concerns are taken into account.

The emphasis of SPECS on evaluating the creativity of computational creativity systems should also highlight the importance of actually performing such evaluation. It is hoped that the proposal of SPECS, the highlighting of previous poor practice at performing evaluation (both within this thesis and in Jordanous (2011a)), and the demonstration of what can be learned from the SPECS creativity evaluation case studies (Chapter 6 and 7) will all encourage a more pro-active and rigorous approach to evaluation. This falls in line with the recent move of computational creativity research events to place more emphasis on creativity evaluation (see Chapter 1 Section 1.5.1). Hence the field can progress in this direction, moving away from scenarios where it is acceptable to label a system as creative without justification or evidence of how the system demonstrates creativity, or to propose such evaluation plans for the future without any requirement to actually perform that evaluation of their system, or to cite a creativity evaluation methodology as justification of why a system might be creative without implementing that methodology as originally intended.

All this is not to say that an evaluation practice pursuing a better understanding of creativity is currently non-existent within computational creativity research. On the contrary, from the papers surveyed in Chapter 2 Section 2.3, 35% of papers reviewed performed some assessment of their system's creativity to better understand how their system is creative, most notably through treating creativity as the combination of novelty and value (and occasionally a third, variable factor) or using

[34] As discussed previously in Chapter 3 and in Chapter 5, creativity transcends the ability to generate good quality artefacts; an example that springs to mind from this discussion is Duchamp's *Fountain* (Chapter 5 Section 5.1.2), which was originally considered too low in quality to exhibit in any prominent way.

Colton's creative tripod (Colton, 2008b) or Ritchie's criteria for creativity (Ritchie, 2007) as the underlying model of creativity. SPECS, however, asks researchers to check how such models fit in within the context of wider research on creativity before adopting them for system evaluation, to ensure that computational creativity evaluation uses models of creativity that fully represent (and fully take advantage of) the knowledge and understanding we have about creativity (see Chapter 3). This encourages computational creativity research to be more deeply grounded within the concept being studied computationally: creativity.

9.3 Reactions from the Computational Creativity community

During the work reported in this thesis, preliminary results and earlier versions of the work have been presented (Jordanous, 2010a, 2011a, of which the latter is most relevant) at peer-reviewed computational creativity research events.[35] This Section reports reactions from the computational creativity research community to the most relevant preliminary version of this work (Jordanous, 2011a) and other personal communications with researchers regarding the work.

Sometimes the idea of evaluating the creativity of computational systems is seen as being too complex to attempt (Oliveira et al., 2007; Cardoso et al., 2009).[36] The evaluation survey in Chapter 2 Section 2.3 showed confusion and a lack of universal direction within the computational creativity research community, as to how to evaluate the creativity of their systems. This is not to say that the research community is not interested in how to evaluate the creativity of their systems. On the contrary, personal communications with various researchers have revealed interest in this work (Norton, 2009; Pease, 2009; Brown, 2010; Rauchas, 2010; Bown, 2011; Brock-Nannestad, 2011; Comajuncosas, 2011; Jansen, 2011; Ox, 2011; Patterson, 2011; Spector, 2011; Tigas, 2011; Krish, 2012; E. Lewis, 2012, as some examples, all personal communications). Most of these researchers were considering applying the resulting methodology to evaluate their own systems; for example, as mentioned in Section 9.1.1, Lee Spector wanted to see how a musical improvisation system he had worked on would compare other systems and sent me details of this system (Spector & Alpern, 1994).[37]

To date, a number of computational creativity researchers have referenced the work in this thesis or in publications related to this thesis work (Jordanous, 2011a, 2010a, 2010c) as evidence to support their discussions about improving computational creativity evaluation.

- Saunders (2012) and Cook, Colton, and Pease (2012) cite the thesis itself. Saunders (2012)

[35] Other papers were also presented at other peer-reviewed research events with different audiences, (e.g. Jordanous, 2009, 2010b; Jordanous & Keller, 2011), but this Section focuses specifically on reactions of the immediate target audience of this thesis, computational creativity researchers.

[36] A comment about evaluating creativity being too hard, made during a talk at Sussex by Geraint Wiggins (Wiggins, 2008), was in fact a motivating factor in pursuing this thesis's research.

[37] Sadly this correspondence happened too late for Spector's system to be included in this thesis.

refers to the proposals for the SPECS methodology and both Saunders and Cook et al. refer to the overview of computational creativity evaluation in Chapter 2, described by Saunders (2012) as 'the most comprehensive and rigorous review to date on the evaluation of computational creativity' (Saunders, 2012, p. 223).

- In her considerations of evaluating linguistic creativity systems, (Zhu, 2012a) comments on the lack of a standardised, directed approach to evaluation of computational creativity, citing the publication in Jordanous (2011a) of the findings about current evaluation practice in the Chapter 2 Section 2.3survey (Zhu, 2012a, 2012b). Zhu also mentions the proposals in Jordanous (2011a) for a standardised framework to evaluate creativity.
- Morris, Burton, Bodily, and Ventura (2012) makes further mention of the inadequacies in practice and associated difficulties, as found in the Chapter 2 Section 2.3 survey and as reported in Jordanous (2011a). Morris et al. (2012) also cite the GAmprovising system (Jordanous, 2010c) as example of a creative system producing musical artefacts.
- In their evaluative investigation of musical 'metacreation', Eigenfeldt et al. (2012) comment on the need for further computational creativity research to pay greater attention to 'validation studies', citing Jordanous (2011a) as an example of existing investigations of this type.
- Commenting on analysis of accounts of creativity, (Pease, Charnley, & Colton, 2012) refers to an earlier version of the work in Chapter 4 (Jordanous, 2010a).

9.4 Towards adopting SPECS as standard for computational creativity evaluation

It is hoped that the proposals in this thesis offer some real value for the computational creativity community, who are generally favourable towards the idea of a more standardised and systematic approach to evaluation. Unsurprisingly, given the scope of the issues surrounding creativity evaluation, some issues remain unresolved in the inner details of the methodology, such as:

- How exactly should creativity be defined (taking into account specific domain requirements)?
- What tests should be performed to evaluate the standards highlighted as important for creativity in a particular domain?

While various ways of dealing with these issues are suggested in Chapters 4, 5, 6 and 7, these issues are too wide-ranging to be resolved within the scope of a single doctoral thesis. Indeed, as Chapter 3 observes, some of these issues have remained open (for human and computational creativity) despite decades of creativity research, and may continue to remain unsolved indefinitely. Under these circumstances, what the computational creativity research field needs is a practical methodology that can be applied for evaluation, using a working definition of creativity if necessary. SPECS is proposed as a practical methodology to fulfil this need. This allows researchers to evaluate their

progress without being too hampered by wider questions and unresolved issues that would otherwise slow down research on philosophical grounds.

Indications from the feedback received on my proposals are that there is both demand and interest for my work, with some potential collaborations in discussion regarding the application of the SPECS methodology for evaluation of other researchers' creative systems. Regular publications on the SPECS methodology and promotion of it at research events will help advertise SPECS to those who it is primarily intended for. Only time will tell, though, whether the community adopts the methodology widely, to achieve the aim of SPECS becoming a useful and widely-applied tool for creativity evaluation in computational creativity.

9.5 Summary

Some issues arose during the implementation of SPECS in Case Studies 1 and 2, reported in Section 9.1. These issues mostly concerned decisions made on the application of SPECS with the Chapter 4 components, such as how well the judges understood the components and how straightforward it was to use the components in evaluation. Wider issues in creativity evaluation were also considered, such as the use of subjective ratings for evaluation feedback, the pressures associated with lack of time, resources and/or information for evaluation and the question of what baseline standards to use for creativity, as was how to deal with missing information at the level of detail required and the applicability of the methodology across different domains. Alongside issues specific to the implementation of SPECS in Case Studies 1 and 2, Section 9.1 also reflected on the SPECS methodology itself, independently of its implementation in the case studies.

Previously, Chapter 2, Section 2.3.4 highlighted various criteria used for creativity evaluation that were inadequate for the purpose (or, alternatively, criteria that were notably useful and could be learned from). Section 9.2 demonstrated how the SPECS approach helps to prevent researchers from adopting inadequate evaluation criteria or approaches in creativity evaluation, mainly through the mechanisms of enforcing a more transparent and informed evaluation grounded in creativity research.

Reactions so far to the thesis content were reported in Section 9.3. SPECS is intended as a tool for the computational research community; hence their reactions and responses to the work in this thesis and to earlier publications during the development of this work indicate interest in what the thesis delivers. In particular, the level of response has demonstrated the interest in and demand for improving computational creativity evaluation across the field. Section 9.4 considered how best to promote the SPECS methodology to the research community, to meet and fulfil this demand.

Chapter 10

Conclusions

Parts of this Chapter are published in a peer-reviewed journal article (Jordanous, 2012).

Figure 10.1: *Wordle* word cloud of this Chapter's content

Overview

The Standardised Procedure for Evaluating Creative Systems, a methodology for evaluating the creativity of computational systems, is reprised from Chapter 5 in Section 10.1. Section 10.2 looks at the contributions of this work to research knowledge. In increasing order of generalness, contributions are made through: the case study findings for individual systems and their relevant creative domains (Section 10.2.1), provision of a methodological tool to answer the question 'How creative is this computer system' (Section 10.2.2) and a clearer understanding of creativity itself, both for computational creativity(Section 10.2.3) and human creativity research (Section 10.2.4).

The SPECS methodology, incorporating the Chapter 4 components, performed well in evaluation and provided useful results. Having said this, there are a number of avenues for further work which would enhance the work reported in this thesis (Section 10.3). Section 10.3.1 considers various work for the future regarding the components from Chapter 4 such as exploring potential reductions in the number of components or presenting the information differently. Section 10.3.2 considers various ways of developing the Case Study evaluations, such as how empirical tests can be used, perhaps in automated form, to assist evaluation. Reflections are made on potential future work relating to the evaluation of musical improvisation systems (Section 10.3.3), extending Case Study 1. Specific reflections on developing computational creativity evaluation are also considered in Section 10.3.4.

Section 10.4 considers how SPECS could be used in the future, predicting the effects of different users and taking into account how creativity may vary over time. Section 10.4 also discusses how to improve the future chances of the SPECS methodology being adopted as a standard creativity evaluation tool for computational creativity research, including thoughts on how it fits with other foundational advances in computational creativity.

Section 10.5 considers the success of this project. Section 10.6 summarises the thesis contributions. To conclude the thesis, Section 10.7 offers some final short reflections on the study of creativity.

10.1 Summary of the SPECS methodology

The final version of SPECS (Standardised Procedure for Evaluating Creative Systems) is reported in Chapter 5. Here the SPECS methodology is reprised in a form that is more generalised, for application that can (but does not have to) include the 14 components from Chapter 4:

1. **Identify a definition of creativity that your system should satisfy to be considered creative:**
   (a) What does it mean to be creative in a general context, independent of any domain specifics?
      - Research and identify a definition of creativity that you feel offers the most suitable definition of creativity.

- The 14 components of creativity identified in Chapter 4 are strongly suggested as a collective definition of creativity.

(b) What aspects of creativity are particularly important in the domain your system works in (and what aspects of creativity are less important in that domain)?

- Adapt the general definition of creativity from Step 1a so that it accurately reflects how creativity is manifested in the domain your system works in.

2. **Using Step 1, clearly state what standards you use to evaluate the creativity of your system.**
   - Identify the criteria for creativity included in the definition from Step 1 (a and b) and extract them from the definition, expressing each criterion as a separate standard to be tested.
3. **Test your creative system against the standards stated in Step 2 and report the results.**
   - For each standard stated in Step 2, devise test(s) to evaluate the system's performance against that standard.
   - Consider the test results in terms of how important the associated aspect of creativity is in that domain, with more important aspects of creativity being given greater consideration than less important aspects.

10.2 Overall contributions of this work

This thesis contributes to research knowledge and progress in a number of different ways. The SPECS acronym stands for Standardised Procedure for Evaluating Creative Systems:

- *Evaluation* is important for comparing how well a system performs in the context of previous related work.[1] Perhaps more importantly, evaluation also collects formative feedback about the strengths and weaknesses of the evaluated system, to better understand and identify how the system contributes to progress on a wider scale and pinpoint which areas should be focused on for improvements (Chapter 1).
- Many systems are described as being '*creative systems*', 'computational creativity programs' or similar, with no justification of why this description is appropriate (Chapter 2). This thesis provides an empirically derived definition to help researchers avoid the pitfalls associated with defining what it entails to be creative (Chapters 3 and 4) and recommends the use of this definition within SPECS.
- Computational creativity research has been provided with a suitable *procedure* to follow for

[1] This perspective on external system evaluation is distinct from a creative system's ability to self-evaluate, as part of the creative processes employed by the system. Chapter 5 Section 5.1.6 gave more details of this distinction.

evaluating the creativity of computational creativity systems (Chapter 5), something which the research field is demonstrably in need of (Chapter 2).

- This procedure is *standardised* so that it is useful for researchers across the field and applicable in many different domains, as shown in the case studies (Chapter 6 and particularly in Chapter 7). This flexibility is not at the expense of specificity, which was a concern raised by researchers at ICCC'11 during discussions of standardisable tools for computational creativity (Chapter 8). SPECS can and should be customised to meet the requirements of the particular creative domain being tackled by the given system.

10.2.1 Summary of case studies: Findings and benefits

In Case Study 1 (Chapter 6), the creativity of three musical improvisation systems was evaluated:

- Voyager (Lewis, 2000).
- GenJam (Biles, 2007).
- GAmprovising (Jordanous, 2010c).

These systems were evaluated in detail, using information about musical improvisation and a set of expert judges. In contrast, Case Study 2 (Chapter 7) explored a different kind of evaluation, looking at the initial impressions formed about systems from short encounters with the systems and limited information. SPECS was applied to evaluate the creativity of five computational systems presented at the International Conference on Computational Creativity in 2011 (ICCC'11):

- A collage generation module for *The Painting Fool* artistic system (Cook & Colton, 2011).
- A poetry generator (Rahman & Manurung, 2011).
- *DARCI*, a system that generates images to fit supplied adjectives (Norton et al., 2011).
- A reconstruction of the story-telling system MINSTREL (Tearse et al., 2011).
- A system that generates soundtracks to narratives based on the emotional content of the narrative (Monteith et al., 2011).

These five systems were evaluated with SPECS based on information given about the systems in 7-minute talks at the ICCC'11 conference.

For both case studies, creativity was defined as the weighted combination of 14 key components identified in Chapter 4 (Step 1). Each of these components was adopted as an standard for evaluation (Step 2) and given a numeric rating out of 10 by judges (Step 3). All ratings were weighted before analysis according to a measure of their perceived contribution to creativity in the domain being addressed by the system being evaluated.

Comparisons could be made between systems in Case Study 1, and to a limited extent, in Case Study 2 (though the difference in domains meant that feedback from the comparisons in Case Study

2 was generally less useful). In Case Study 1, GenJam was found to be overall most creative and received the highest ratings for most components, though Voyager was perceived as being more spontaneous.

For Case Study 2, less was to be gained from comparisons, though *DARCI* (Norton et al., 2011) was generally perceived as one of the more creative systems. This case study's benefits were in showing how creativity evaluations could be performed quickly, dealing with limited/missing information and short time scales. Additionally the evaluations provided feedback on how informative the ICCC'11 presentations were about the systems' creativity, highlighting areas which the judges were not able to evaluate given the information in the talks.

As Section 10.2.2 discusses, the formative feedback from both case studies makes valuable contributions to research knowledge, particularly for the authors of the systems evaluated but also for researchers working in similar systems. This is a key contribution of the case study findings. On a wider scale, the definitional work in identifying what constitutes musical improvisational creativity can be helpful for music researchers, as partially demonstrated by the acceptance of this work and earlier versions (Jordanous & Keller, 2011; Jordanous, 2010b, respectively) for presentation at international musicology conferences.

Key aspects of creativity in musical improvisation have therefore been identified: the ability to communicate and interact socially, the possession of relevant musical and improvisational skills and knowledge, and the emotional engagement and intention to be creative. Conversely, the actual musical results produced during improvisation are relatively less important for creativity when compared with the process of improvising. Also, general intelligence is considered less important than having specific expertise and knowledge. Beyond identifying evaluation standards for the work in this thesis, such knowledge has greater application. For example, a detailed understanding of what makes musical improvisation creative helps improvisers and their educators identify what they should work on to improve their creativity (Gibbs, 2010).

To a more summative extent the information on priorities in the Case Study 2 domains is useful to researchers working in those domains, although this information is briefer due to the limits placed on how it was gathered. Case Study 2 demonstrates how the 14 components can be quickly processed in terms of relative contributions to creativity in a domain, without necessarily requiring extensive investigation if researchers' time is short.

10.2.2 Addressing the research question of how to evaluate computational creativity

The research question driving this research is:

> How should we evaluate the creativity of a computational creativity system?

This thesis investigates this question in detail and offers a specific, well-researched, practically applicable and useful answer in the form of the SPECS methodology.

- The SPECS methodology has arisen from the consideration of issues in Chapters 2 and 3, as described in Chapter 5.
- SPECS was demonstrated in use in two contrasting case studies (Chapters 6 and 7), utilising a working definition of creativity derived in Chapter 4 and further investigations.
- This use of SPECS was extensively evaluated, particularly in comparison to alternative methods of creativity evaluation using other methodologies and/or human judgement (Chapter 8).
- Generally the SPECS methodology performed well in terms of satisfying the demands identified for creativity evaluation and in comparison to other systems, though some points have been noted for improvements (Section 10.3 of this Chapter).

10.2.3 Contributions to computational creativity research

Evaluation can give vital information about what our systems contribute to knowledge and how they can be improved. Currently the balance between evaluation of quality and evaluation of creativity is inappropriately skewed towards evaluating quality (Chapter 2 Section 2.3), with terms such as 'creative systems' used descriptively rather than with any kind of justification. A more even balance can be struck; measures of quality can be incorporated within the application of the SPECS methodology, for example through tests for the *Value* component.

SPECS is presented as a solution to the 'methodological malaise' (Bundy, 1990) that has arisen in computational creativity, as evidenced in the survey in Chapter 2. A lack of systematic and rigorous evaluative practice has been shown to exist within computational creativity research, leading to missing information on how a system contributes to research in a wider context, an inability to track progress in any measurable way and a generally inconsistent and non-standardised approach to evaluation that could significantly stilt research progress in the field (Bundy, 1990; Pearce et al., 2002).

In SPECS, computational creativity researchers are given steps to follow to evaluate their systems systematically and meaningfully. To assist this process, a working definition of creativity is offered, in the collective form of the components identified in Chapter 4. Chapters 6 and 7 offer two contrasting examples of how SPECS can be practically applied.

Chapter 8 considers how SPECS compares to existing methodological contributions in computational creativity evaluation. Experiments were carried out implementing key existing methodologies on the systems in Case Studies 1 and 2, under similar conditions imposed for SPECS in each case study. Ritchie's criteria (Ritchie, 2007) and Colton's creative tripod (Colton, 2008b) were applied and surveys of human opinion were also conducted, to explore whether human opinion (taking consensus

from a larger sample size if necessary) could represent a 'ground truth' by which to measure the success of alternative creativity methodologies against.[2] In Case Study 1 the methodologies generated similar comparative rankings of the systems according to their creativity, but generated different types and amounts of feedback, using different models of creativity. In Case Study 2, on systems from different domains, the same observations apply except that the methodologies also generated different comparative rankings, with no one system emerging as more creative than the others.

In Sections 8.5 and 8.4, five criteria were used to critically analyse and compare the methodologies listed above and also the FACE model framework (Colton et al., 2011).[3] This had the advantage of generating comments and comparative feedback on each of these methodologies from a number of evaluators, both from within the current doctoral work and from external evaluators.

In analysis, SPECS encompasses most positive aspects associated with existing methodologies and improves upon weaker aspects of the other methodologies in various areas. The key advances that SPECS makes are: 'extremely useful' feedback (according to the external evaluators); the emphasis on detailed formative feedback to assist system development rather than cursory and somewhat artificial summaries; the requirements to base the evaluation on an informed and researched view of what it means to be creative; and the ability to customise the evaluation standards for creativity towards particular requirements of the domain being investigated. The latter point in particular distinguishes SPECS from existing creativity evaluation contributions to date; as Chapter 3 Section 3.6.4 has shown, creativity can be manifested quite differently across different creative domains and SPECS is the first evaluation methodology to enable and actively encourage evaluators to take these variances into account.

An additional contribution to computational creativity research from the above analyses is the set of five criteria derived in Section 8.4.2, to be used for meta-evaluation of computational creativity evaluation methodologies, and the feedback regarding each of the above listed creativity evaluation methodology. The application of several different methodologies to evaluate the creativity of various computational systems also generates helpful examples of these methodologies in practical application. It has been noted in this thesis (particularly during Chapter 2) that example applications help to illustrate the use of an evaluation methodology; this was a key motivation for carrying out the two Case Studies.

As well as being a methodological contribution, the SPECS approach to evaluation has generated both comparative feedback on how creative various computational improvisers are and, perhaps more importantly, detailed formative feedback on how to improve each system's creativity. This feedback

[2]This assumption of human opinion as a 'ground truth' or right answer was found to be unreliable; variances in human opinion lead to differing feedback and comparisons, even in studies with a larger group of respondents.

[3]Reports of the FACE model were published only after the main experiments in Case Study 1, hence were not included for these larger experiments, but could be constructed and included for this later meta-evaluation work.

has been described as 'extremely useful' by researchers involved in developing the musical improvisation systems evaluated. Understanding why a computational system is seen as creative, or why one system is deemed more creative than another, gives vital information in the task of modelling creativity computationally:

> 'we are aiming, through the study of machine creativity, to (i) further our understanding of creativity (human and other), and/or (ii) build programs which are useful in achieving practical goals.'(Pease et al., 2001, p. 8)

For the authors of the nine systems evaluated in Case Studies 1 and 2, this thesis provides detailed evaluative analysis and comparison of the creativity of their systems, possibly in greater detail than they themselves have collected for evaluation. Chapters 6, 7 and 8 are rich in information on the strengths and weaknesses of these nine systems, both as individual systems and in comparison to the achievements and shortfalls of other systems. The system authors can see where their system excels, particularly when compared to other systems, to be fully aware of how their work contributes to research knowledge. Identifying weaknesses in their system is perhaps even more useful should the researchers wish to understand and develop their system's perceived creativity. The way SPECS has been implemented in Case Studies 1 and 2 makes it clear what work will be most productive in terms of creativity improvements.

The usefulness of the collected qualitative and quantitative evaluation data also extends further past those whose systems have been evaluated; those conducting research in modelling creativity in any of the covered areas would do well to look at the feedback collected for the appropriate systems, as there is much to learn from here. A wide variety of creative domains have been covered: artistic creativity, linguistic creativity, incorporating poetry and narrative generation and musical creativity with a particularly detailed focus on musical improvisation systems.

One area that is missing from the above list is scientific or mathematical creativity; it is unfortunate that no systems of this type were included in case studies. It is to be hoped, however, that it takes little effort to see how the SPECS methodology could be applied to systems such as those discussed in Colton et al. (2000), for example. The key to the SPECS methodology is that it can easily be customised to various creative domains; in fact this is highly advocated in SPECS, especially Step 1b.

10.2.4 Contributions to human creativity research

The benefits of a greater and more in-depth understanding of creativity are not solely restricted to computational creativity researchers, but also to creativity research in general. Problems with defining and understanding creativity are widely documented and investigated, often without satisfactory resolution (Chapter 3). Although this thesis focuses on how this has hindered research progress in computational creativity, there is a plethora of research on what constitutes creativity, from the early

to mid 20^{th} century (e.g. Poincaré, 1929; Guilford, 1950) to far more recent investigations (e.g. Plucker et al., 2004; Hennessey & Amabile, 2010); clearly this is an ongoing issue.

As Chapter 3 has reported, creativity research spans several different disciplines, each with their individual priorities and foci. The work in Chapter 4 brings together key contributions to research from a range of disciplines including psychology, education, management and artificial intelligence. Using empirical methods from computational linguistics, common themes across these contributions are identified to form the collection of components presented in Chapter 4. Such an overview is of interest to researchers in human creativity as well as computational creativity.

This thesis offers another contribution to human creativity that is not immediately apparent but which could have many potential benefits. Returning to the use of the descriptor 'Creative Systems' in the SPECS acronym, it is not specified that SPECS should just be applicable to computational creativity systems; the word 'computational' is not included in this acronym. It is not a great stretch of the imagination to see *people* as *creative systems*; hence the SPECS methodology could be used to evaluate the creativity of people as well as computational systems. The implementation of this type of evaluation shall not be attempted within the scope of this thesis; however this posits an intriguing use of the SPECS methodology.

10.3 Future development of this work

In Chapter 8 the SPECS methodology was shown to be an improvement on other methods for evaluation of creativity in several ways, including definitional clarification on what should be evaluated, the ability to take account of all Four Ps (Chapter 3 Section 3.4.2) rather than just product, as well as matching evaluation priorities to priorities of creativity in the relevant domain. As noted in Chapter 9 Section 9.1, though, there are a number of areas in which further work on refining SPECS, and the Chapter 4 componential definition of creativity, would be very useful for increasing the usability and applicability of the methodology. As Eigenfeldt et al. (2012) point out, 'while the computational creativity literature has started investigating [evaluation studies] ... a great deal remains to be done' (Eigenfeldt et al., 2012, p. 141).

10.3.1 Improvements related to the use of the Chapter 4 components model of creativity

Reduce the number of components used to define creativity 14 components of creativity were identified in Chapter 4. Although efforts were made during this work to keep this number of components as small as possible, judges in both case studies noted the quantity of information they were expected to provide. The size of the set of components was also reflected in the length of evaluation experiments in Case Study 1, the amount of data left unrated by judges in Case Study 2 where it might be argued

that this data had been provided during presentations (especially in cases where one judge was able to provide a rating but the other was not), as well as the time taken for analysis of the evaluation data. In external evaluation of the methodologies (Chapter 8 Section 8.4.2, Biles and Keller both commented on the volume of data, with Keller noting that 'if anything could be done to reduce the number of categories, that might it more attractive', echoing a previous similar comment made by Biles on the volume of data generated by SPECS.

There would be value in future projects addressing the composition of the set of components for creativity and how this set might be reduced in size. Although the computational linguistics and clustering methods have identified key aspects of creativity, techniques such as principal components analysis (PCA) or factor analysis could potentially minimise the set of components in different ways, by using the fewest components to represent as much of the data as possible (PCA) or by identifying any underlying patterns and common themes within the components (factor analysis) (Sauro & Kindlund, 2005). Another option would be to remove components that were shown not to contribute much to creativity in a particular domain, acknowledging that these components may offer some information but are relatively unimportant compared to those components making a greater contribution. This last option is simplest to implement but ignores the fact that every component identified in Chapter 4 represents words (or clusters of words) used significantly more often than expected in connection with creativity.[4] As long as important evaluative data is not lost, though, some form of reduction such as the suggestions outlined above would make the set of components from Chapter 4 more manageable to apply within SPECS, speeding up the evaluation process.

Improve the presentation of the component evaluation results Reducing the number of components used may sacrifice useful detail. It was noted that each of the 14 components were mentioned in responses to the questionnaire research investigating what musical improvisation creativity is, for Step 1b of SPECS (Chapter 6 Section 6.3.2). As seen in the various systems in Case Studies 1 and 2, some components become more important than others across different domains; it is useful to have the flexibility to represent as much detail as desired for a particular study.

An alternative to reducing the amount of data presented is to investigate how best to present large collections of data so that the volume of data is not 'imposing', (as Biles described the feedback for SPECS in Chapter 8 Section 8.4) Many fields related to computational creativity have techniques for presenting large volumes of detailed data in a manageable and accessible way, without sacrificing the detail, such as machine learning and other branches of Informatics, for example, or Psychology. Reviewing how these fields present large detailed sets of data and findings would be useful in improving the presentation of results obtained with SPECS and the components, making SPECS' results more

[4] Additionally, in the context of one specific domain (musical improvisation), it was noted that each component was mentioned (to varying degrees) in the questionnaire about creativity in musical improvisation in Chapter 6 Section 6.3.2.

user-friendly.

Make the components easier to understand and more clearly defined Many of the names attached to the components derived in Chapter 4 could be interpreted in multiple ways, as was shown in judges' interpretations in Case Study 1 and in feedback from the Chapter 8 external evaluation study. The components are presented with some explanatory definitions in Chapter 4, but could be defined more explicitly (Pérez y Pérez, 2012, personal communications). Similarly, in the Chapter 8 external evaluation study, Biles commented that there was a steep learning curve associated with applying SPECS as recommended (i.e. with the components): 'this is a tool that requires a lot of practice before it can be used productively.' Biles also thought, however, that such learning and practice would be productive: 'On the other hand, if this instrument is used by knowledgeable and motivated evaluators, it could yield a lot of useful information.'

Further work could therefore be devoted to making the meaning of the components clearer. This goal ties in with recent collaborative work with Bill Keller, where the components were recently published on the Semantic Web in the form of an ontology of creativity (Jordanous & Keller, 2012).[5] In this ontology, each component is linked both to the definitions already provided and back to the original 'creativity words' (see Chapter 4) that were grouped together in clusters to form that component. The creativity words collectively make a more precise contribution to the meaning of each component as they show the exact word groupings that each component was intended to describe. It would be helpful if this ontological representation was accessible in a more accessible, reader-friendly presentation than the OWL code it is currently published as; this would make available the word groupings to an audience who are not currently so familiar with Semantic Web technologies. Therefore one area for further work would be to explore alternative ways of viewing and/or publishing the creativity ontology, with suitable visualisations of the definitions and linked creativity words.

10.3.2 Further development and alternative approaches in the Case Studies

Incorporate empirical and automated tests The devising of empirical tests was not attempted in the case studies in this thesis. If adoption of such tests had led to flawed evaluation results in the case studies, it would have been unclear as to whether this was the result of problems within the methodological steps itself or within the specific empirical tests not measuring what they were intended to. The use of human evaluation of each standard removed this issue.[6]

Chapter 9 Section 9.1.4 points out some issues related to using subjective ratings provided through people's opinions as opposed to performing more objective tests. Alongside this discussion, Chapter

[5] The creativity ontology is published in OWL format (a standard format for representing Semantic Web ontologies) at *http://purl.org/creativity/ontology*, last accessed December 2012.

[6] Perhaps this issue was replaced with the issue of whether the judges understood each component well enough - a possibility despite careful attention paid to describing the components to judges.

9 Section 9.1.4 discusses how empirical tests could be used. This is relevant whether the Chapter 4 components are used to define creativity, or if another set of standards are adopted from an alternative definition of creativity during the SPECS steps. In discussing the application of SPECS for application, Pérez y Pérez suggests that 'we leave at least open the possibility of combining human judges and computer models' (Pérez y Pérez, 2012, personal communications).

The value of human judgement should not be overlooked, if the computational creativity system is intended to be perceived as creative by people. Human standards of aesthetics and success tend to change as domains develop. For those tests which can be automated, however, performing such evaluative tests manually can be time-consuming and perhaps tedious, particularly if tests must be repeated for comparison of several systems. An automated tool to explore different evaluation strategies would be helpful, reducing the reliance on human judges when evaluating creative systems and helping to avoid some potential bias creeping into this evaluation. Automated evaluation of creative systems where possible would also remove the requirement for the researcher to carry out mechanical and time-consuming evaluative tasks that could instead be delegated to an artificial critic agent, freeing up research time and effort for other tasks.

The devising of objective measurement tests would be a fascinating area in which to pursue further research, though the scale of this should not be under-estimated. The task of identifying empirical tests of components within the area of musical improvisation (the focus domain of Case Study 1) may be assisted by the plentiful research in related areas (Ames, 1992; Huron, 2001; Temperley, 2001; Goto & Hirata, 2004; Temperley, 2004; Huron, 2006; MacDonald, Byrne, & Carlton, 2006; Downie, 2008; Stowell, Robertson, Bryan-Kinns, & Plumbley, 2009, for a small subset of such work).

Obtain external evaluation for Case Study 2 results In Chapter 8 Section 8.4, the researchers behind the systems in Case Study 1 were invited to evaluate the results obtained by SPECS and by other creativity evaluation methodologies, in terms of five criteria that were presented in Section 8.4.2 of Chapter 8. It would also be very useful to invite the authors of the five Case Study 2 systems to evaluate the methodologies, to find out their opinions. This next step in gathering external further evaluation of the various methodologies should also yield interesting information on how the time and resource limitations have affected the quality of the various evaluations.

Compare evaluative findings in this thesis with other system evaluations Following on from the previous point mentioned, on how Case Study system authors would provide useful feedback on the relative performance of methodologies, it may also be the case that these authors conduct their own evaluation to their systems. For example, Norton et al. (2012) have a forthcoming publication on the DARCI system in Case Study 2 that includes their own evaluations of DARCI. Norton et al. evaluate DARCI on aspects other than creativity but include an evaluation of DARCI's creativity through

application of Colton's creative tripod (Colton, 2008b). As well as providing material for comparisons between the two creativity evaluation approaches on the same system, comparisons may also provide tangible evidence on impact of the limited resources and time in the Case Study 2 evaluations, so will be useful to investigate. Other Case Study system authors may also have evaluated the creativity of their system or be planning to do so, which would generate further material for comparison.

Conduct full FACE/IDEA evaluation In Case Study 1, FACE evaluation was carried out on the systems at a different time and in different circumstances to the other evaluation methodologies.[7] This may have biased comparisons between FACE and the other methodologies (although feedback from the external evaluators in the Chapter 8 Section 8.4 study suggested that the FACE evaluation data were a 'completely correct' reflection of their system along the four aspects of FACE). Another concern is that the FACE/IDEA models are still in active development, rather than being presented in a stable final form as the other methodologies have been.[8]

While some comparative conclusions could be drawn from the external evaluations conducted in Chapter 8 Section 8.4, one area for future work would be to conduct FACE (or IDEA) evaluation using comparable experiments to those used for the other methodologies in the case studies. In particular, FACE models would be constructed for Case Study 2 systems as well as Case Study 1, using similar resource and time constraints as for the other methodologies in the respective Case Studies.

Compare against other methodologies and creativity metrics Key methodologies in the existing computational creativity evaluation literature were selected for comparison to SPECS (Ritchie, 2007; Colton, 2008b; Colton et al., 2011, alongside the surveying of people's opinions). These methodologies were selected according to previous use within the community for evaluation or perceived future potential. Other methodologies and smaller-scale metrics have been suggested in the past for creativity evaluation (Chapter 2. Though the original set of methodologies was constrained to make evaluation and comparison more manageable, further methodologies and/or metrics could now also be added to future comparative investigations (e.g. Pease et al., 2001).

10.3.3 Specific reflections on evaluation of musical improvisation

Evaluation of musical improvisation: the influence of computerised output In the Chapter 8 Section 8.4 external evaluation study, one of the evaluators (Lewis) commented on the influence on judges of the ' "computerized" nature of the music', compared to hearing more acoustic performances. In performances with computational systems, Lewis has found that the use of Disklaviers is helpful

[7] This difference in timing was because details of the FACE model were published after the start of the Case Studies, when evaluative data had already been collected.

[8] For example, the 'E' of FACE has changed from 'Expression' in Colton et al. (2011) to 'Example' in Colton et al. (2012). The FACE/IDEA models are discussed in greater detail in Chapter 2 Section 2.1.5.

in disguising the computer players amongst the human players. Such comments bring to mind the findings in Moffat and Kelly (2006) about subconscious biases being introduced by the knowledge that a piece of music was computer generated. It would be useful to revisit the experiments in the context of the work of Moffat and Kelly (2006) and explore how experiments could incorporate tools such as Disklaviers and other methods of disguising computational performers. The research question of whether this would generate different evaluations or not is an intriguing one to explore; it is conjectured from the participant comments in the various surveys carried out in Chapters 6 and 8 that a more human-sounding performance may not attract the same number of negative opinions.[9]

Comparative evaluation across more musical improvisation systems As well as the Impro-Visor system (Gillick et al., 2010) and the jazz improvisation system by Spector and Alpern (1994), there are several other musical improvisation systems and models whose creativity could be evaluated (either externally, or by the system author(s)) (e.g. Thom, 2000; Johnson-Laird, 2002; Pachet, 2004; Hodgson, 2006c). Evaluation findings obtained from these further experiments would add to the knowledge contributions already obtained from the existing evaluations in Case Study 1. More data would therefore be provided, allowing researchers both to learn more about the creativity of musical improvisation systems from a wider range of sources and to potentially perform comparisons across more systems or across a selected subset of the most relevant systems with data available.

10.3.4 Specific reflections on evaluating computational creativity

Generating informed evidence to justify computational creativity In discussions of this thesis content with members of the Computer Science department at the University of Warwick, Meurig Beynon and Steve Russ posed a question on 'whether it even makes sense to speak of "computational creativity" ' (Beynon, 2012, personal communications). The SPECS evaluations in Case Study 1 meant that a response could be provided to this question based on an informed set of data and evidence, on how computer systems could be creative in improvising music, relative to how a human would be creative in improvising music. The ability of SPECS to generate evidence that can be used in response to this type of question arises as a direct result of both improving evaluation practice and in being clear and detailed in how a researcher targets the goal of creativity in their system. It is conjectured that more clear and grounded evaluative practice, such as that encouraged in SPECS gives computational creativity researchers tangible evidence to use to support justifications of why computational creativity is possible as a discipline. Experiments investigating the validity of this conjecture would potentially have useful consequences for the field as a whole.

[9] The points here lead onto another suggestion for final work that relates to the computational (or otherwise) origin of creative artefacts - see the final suggestion in this Section.

Conduct further investigations on the comparability of systems In external evaluation feedback, Keller commented on how useful it was to compare systems with different goals. A similar point was also made during the discussion sessions at the ICCC'11 conference on computational creativity (as reported in Chapter 2 Section 2.3.5).

As mentioned in Chapter 7 Section 7.1, researchers can learn from what has been done in the past.

> 'without cultural artifacts, civilization has no memory and no mechanism to learn from its successes and failures.'[10]

In the computational creativity literature, Pérez y Pérez and Sharples (2004) provide an isolated example of detailed comparisons between systems, in the context of linguistic creativity.[11] The review in Chapter 2 Section 2.3 found that only 11 out of 75 reviewed papers directly compared their systems to other systems or to people performing the same creative activity, with only 2 papers performing comparisons between their systems and systems from other researchers. In ICCC'11 discussions (Chapter 2 Section 2.3.5) a comment was made that creative systems are intended to be novel and therefore would not be comparable with other existing systems.[12]

Further investigation could be performed into what comparisons would be useful, both within computational creativity and in related disciplines. This could take the form of collecting opinions and studying attitudes within the context of computational creativity and examining situations where comparisons were performed and what was learnt. Such investigation could be situated within computational creativity and also within other related disciplines, to see how those other disciplines approach comparative evaluation. Other disciplines are likely to have more established practices compared to the relatively young computational creativity field (Chapter 1 Section 1.5).[13]

Follow up study to the Chapter 2 survey of current practice in creativity evaluation The survey of evaluative practice (Chapter 2 Section 2.3) covered the period up to 2011, mostly looking at papers published 2007-2011. Given the emphasis on evaluation seen in recent calls for research contributions at computational creativity events (Chapter 1 Section 1.5), it can be speculated that evaluative practice may change over the future years, hopefully influenced positively by this thesis and papers on evaluation such as those relating to this thesis and others (e.g. Eigenfeldt et al., 2012; Zhu, 2012a). Comparisons between this survey and a repeat of the survey in five or ten years time (or beyond) would help highlight any change in evaluative practice.

---

[10] Original source unattributed, quote taken from *http://archive.org/about* (last accessed November 2012).

[11] In personal communications (Pérez y Pérez, 2012), Pérez y Pérez re-emphasis his belief of the 'importance on analysing how systems develop and improve over the years'.

[12] The same requirements for novelty are however necessary in research practice generally; researchers are usually required to demonstrate that their research makes novel contributions.

[13] Evidence for the conjecture made about existing established evaluation practices could also be collected, to support this conjecture and to provide evidence on how evaluation is performed and treated in other similar fields.

Multi-agent simulations of a creative society Initial exploratory work has been conducted on a multi-agent system that simulates an interactive society consisting of creative musical improvisers and their critics. This simulation could host investigations into how existing creativity evaluation methods can be implemented in an 'artificial music critic' (Machado, Romero, Manaris, Santos, & Cardoso, 2003) to automate evaluation of musical creativity. In this musical society, agents that generate music interact with critic agents such that those agents deemed most creative by the critics become most productive over time, with less creative agents becoming less active.

Exploratory experiments have been carried out using Java.[14] Test simulations incorporated several Critic agents using a randomised number generator as a temporary evaluation score mechanism. The Critic agents interact with different Producer agents who use random but highly parameterised music generation strategies based on the improvisation methods in Jordanous (2010c). The intention is that the Producer agents would use different methods of improvisation to generate music. For the Critics, alongside automated implementation of SPECS tests, alternative methods for creativity evaluation could also be implemented (e.g. Colton et al., 2011; Colton, 2008b; Ritchie, 2007; Pease et al., 2001). Such a simulation can explore how amenable empirical evaluation methods are to automation in artificial agents and explore the evolution over time of a creative musical improvisation society.

Exploring the influence of perceived origins of creative artefacts The final avenue for further work described in this Section is potentially one of the most intriguing. The circumstances that have made this future work possible emerged through chance during the external evaluations for Case Study 1. As described in Chapter 6, Case Study 1 originally included four systems, of which the fourth was Impro-Visor (Gillick et al., 2010). Case Study 1 was reported Jordanous (2012) as including all four systems, believing that the musical examples used for evaluations were generated by each system and hence reporting the examples as being computer-generated in the various evaluation experiments. In external evaluation of the evaluative results by Keller for the Chapter 8 study, Keller noticed that the musical examples used for Impro-Visor were not generated by Impro-Visor's grammar-based processes, but were composed by Keller (occasionally with the assistance of the advice functionality built into Impro-Visor). The improvisations composed by Keller were published on Impro-Visor's website[15] to illustrate his assisted compositions as an example of Impro-Visor's educational functionality, helping people to learn how to improvise better. This situation has identified a key difference between the samples used for Impro-Visor and the system-generated samples used for the other three systems in Case Study 1.[16] While this meant that for the purposes of this thesis Impro-Visor had to be removed from Case Study 1, there are potentially some fascinating insights here for further research as

[14] Specifically the Java Agent DEvelopment Framework (JADE): http://jade.tilab.com/

[15] *http://www.cs.hmc.edu/keller/jazz/improvisor/*, last accessed December 2012.

[16] This thesis should be treated as a retraction of comments and findings made in Jordanous (2012) about Impro-Visor.

a human's creativity (assisted by a computational system in part) has been treated as the creativity of a computational system, due to the incorrect belief during evaluation that the musical examples were computer generated. The research questions opened up by this situation are somewhat tangential to the main practical focus of this thesis in performing evaluation of computational creativity, hence why Impro-Visor was removed from Case Study 1. The data obtained from this scenario, however, offer some intriguing insights on how beliefs about the generation method for a creative artefact influences its perceived creativity. These research questions will be pursued in future research.

10.4 Future use of the SPECS methodology

As Chapter 9 Section 9.1.6 discusses, it would be interesting to see how other people use the SPECS methodology. Although the two case studies in Chapters 6 and 7 take different approaches to evaluation and look at systems in different domains, some of the implementation aspects in each case study are deliberately kept similar for purposes of consistency and comparison, such as the use of judge-supplied subjective ratings and the use of the Chapter 4 components.

SPECS allows the researcher some freedom in how it is interpreted, whilst still ensuring the evaluator uses a clearly stated and transparent approach. Different perspectives on the methodology would be interesting to see, particularly in evaluation of more scientifically- or mathematically-orientated creative systems for which priorities will be different to the generally more artistic systems evaluated in this thesis. Another application it would be intriguing to see attempted is if SPECS is adopted for evaluating people's creativity, treating people as 'creative systems'.

Chapter 9 Section 9.1.6 discusses ways in which other researchers may come to use the methodology for evaluation, either by seeing how natural demand for the methodology develops or by carrying out controlled studies. Already there is some interest in applying SPECS to evaluate systems (Tigas, 2011, personal communications); it remains to be seen how demand will develop.

10.4.1 Promoting the SPECS methodology as a research tool

An important motivation of this work is to provide a tool for evaluation to assist computational creativity researchers in their research. In Chapter 9, Section 9.4 discussed how the SPECS methodology can be promoted to computational creativity researchers. The intention behind this promotion work, which would involve conference submissions and journal publications as well as circulation of this thesis, is that details of the SPECS methodology are made available to those who would most benefit: researchers who want a standardised procedure to evaluate creative systems.

Chapter 9 Section 9.3 describes reactions to a paper presenting the basis of this thesis (Jordanous, 2011a) at the International Computational Creativity Conference in 2011 (ICCC'11). One of the

points mentioned during debate was a comparison to evaluation in evolutionary computing, where the research community have collectively derived evaluative criteria over time. If the same process were to happen in computational creativity, this would not only strengthen evaluative practice in this research area but also engage the research community as a whole in evolving criteria for the steps in the SPECS methodology.[17] Such collective development of evaluative criteria would however hopefully motivate people towards evaluation, as they become involved in deciding what should be considered. A suitable arena is beginning to develop for discussions of this nature, through research proposals on the collective formulation of a *Computational Creativity Theory* (including the FACE model explored in Chapter 8 Section 8.4) as a formal description of creativity and creativity research, to underpin computational creativity research (Pease & Colton, 2011a; Colton et al., 2011; Pease & Colton, 2011b; Charnley et al., 2012; Colton et al., 2012).

10.4.2 Development of creativity over time

The previous Sections considered how the computational research community could become involved in developing the SPECS methodology, either through different applications of SPECS to their own systems or by constructing community-defined criteria for evaluation. Another aspect to consider is how creativity itself adapts over time. Perceptions do not necessarily remain constant over time but change according to background context and the influence of others. An example of this is Johann Sebastian Bach's music, which was considered outmoded and was largely ignored during and after Bach's lifetime, with popular interest only revived several decades later (from 1750 to around 1830), when musical compositional styles had moved on (Temperley & Wollny, 2011).

While it is strongly recommended in this thesis that the components in Chapter 4 are used as the basis for performing SPECS, these components are derived from literature taken from a fixed period of time (1950-2009) and may not continue to reflect key aspects of creativity as it evolves in future decades. SPECS, on the other hand, is customisable so that it can be adapted to different manifestations of creativity, without relying on any fixed interpretations of what creativity is. The methodology offered in Chapter 4 can be repeated with updated corpora as and when necessary, to update the componential representation of creativity. The SPECS methodology should therefore continue to be applicable in future decades, adapting to creativity as it exists in the future.

10.5 Gauging the success of this project

Returning to the original research aims of this thesis (Chapter 1 Section 1.6), this doctoral work aimed to make the following contributions:

[17] The components in Chapter 4 are offered as at least a starting point for such a process of evolution, drawing together several different strands of creativity research for a firm basis of understanding of creativity.

- A practically applicable, standardised, flexible methodology for evaluating how creative computational systems are. This methodological tool will generate constructive feedback as well as summative assessments.
  - *SPECS is derived in this thesis and is presented in Chapter 5, taking Evaluation Guidelines for good practice in evaluating the creativity of computational systems (Chapter 5 Section 5.2) and expressing the Evaluation Guidelines in a format that can be practically applied (Chapter 5 Sections 5.1 and 5.4). The Case Studies in Chapters 6 and 7 demonstrate evidence that the SPECS methodology generates formative feedback as well as summative assessments; this formative feedback has been deemed 'extremely useful' by external evaluators whose systems have been evaluated (Chapter 8 Section 8.4).*
- A clearer understanding of creativity research that crosses interdisciplinary divides and collects together findings from different academic perspectives.
  - *Chapter 3 investigates creativity research in detail, across a number of different academic perspectives on creativity are considered ranging from psychology research to legal interpretations, including how creativity has been treated in computational creativity research. The findings from this cross-disciplinary investigation are collected together and empirically analysed in Chapter 4.*
- A working definition of creativity to use for evaluation.
  - *The Chapter 4 components are derived as a working definition of creativity, to be used for evaluation with SPECS as criteria for testing. They represent the common elements obtained from empirically examining various different perspectives on creativity, over several decades.*
- Evaluative feedback for a number of systems in a variety of domains, as a result of practically applying the methodological evaluation tool in case studies.
  - *Chapters 6 and 7 report the practical application of the SPECS evaluation methodology to Case Studies 1 and 2 (respectively). These case studies collectively cover five different domains (or more depending on how systems are classified by domain) and generate evaluative feedback for the various systems that are investigated.*
- Criteria for meta-evaluation of creativity evaluation methodologies and the application of these criteria to critically compare key methodologies.
  - *Chapter 8 Section 8.4.2 derives five criteria for meta-evaluation of creativity evaluation methodologies from considerations of existing research in this area from computational creativity, theory verification/scientific method and good practice in research evaluation.*

In Chapter 8 Sections 8.4 and 8.5 these criteria are applied for critical comparison of the SPECS methodology with key existing creativity evaluation methodologies as identified in Chapter 2: Colton's creative tripod framework (Colton, 2008b); Ritchie's empirical criteria (Ritchie, 2007); the FACE model of creative acts (Colton et al., 2011); and also surveys of human opinion, to investigate if human opinion can be used to identify baselines or 'ground truths' for creativity evaluation.

Considering the overall success of this project, below several 'success-criteria' are listed that have emerged during this thesis for what this project must achieve, or what it would be desirable but not essential for the project to achieve. The essential criteria are listed with an asterisk; the other listed criteria are desirable but not essential. Next to each criterion, a response is given (italicised for easier identification) as to the project's performance on this criterion, with links to supporting evidence from relevant parts of the thesis.

For success, this project must(*)/should:

* Propose and produce a creativity evaluation methodology for computational creativity. - *Yes - the thesis proposes the Standardised Procedure for Evaluating Creative Systems (SPECS), produced during this thesis.*
* The above-mentioned methodology must(*)/should:
    * Be practically applicable, i.e. it can actually be used and when used, it generates evaluation feedback - *Yes - Case Studies 1 and 2 (Chapters 6 and 7 respectively) report how the SPECS methodology has been used to evaluate the creativity of various different systems, generating evaluation feedback in each case. Chapter 5 presents practical steps to follow to carry out SPECS evaluation, both in the presentation of SPECS and its accompanying commentary (Section 5.3) and in the decision tree style diagrams illustrating practical assistance for carrying out SPECS in various scenarios (Section 5.4).*
    * Address concerns highlighted by review of existing methodologies and current creativity evaluation practice - *Yes - Chapter 9 Section 9.2 describes how SPECS addresses the concerns raised by the reviews in Chapter 2, particularly with reference to the issues highlighted in Chapter 2 Section 2.3.4.*
    * Make contributions that surpass those of existing methodologies - *Yes - Chapter 8 investigates in detail how SPECS compares to other existing methodologies, through the comparative analysis of implementations of each methodology and through external evaluation. Overall comparisons are given at the end of this Chapter (Sections 8.4 and 8.5). To summarise the content referenced above as evidence for this current criterion, SPECS generates more evaluative feedback overall, in more detail than other methodologies. The*

feedback has been described as 'extremely useful' by the evaluators who considered it in comparison to the other methodologies considered in Section 8.4. Additionally, SPECS evaluation is based on adopting a model of creativity that is more informed and more faithful to creativity. Notably, SPECS allows (and encourages) definitions of creativity which can be customised to account for specific changes in requirements across different types of creativity, following the discussions in Chapter 3 Section 3.6.4 on how creativity is manifested differently in different creative domains.

  * Satisfy the five meta-evaluation criteria for computational creativity evaluation methodologies (Chapter 8 Section 8.4.2) of correctness, usefulness, faithfulness as a model of creativity, usability and generalisability - *Yes, with one caveat - Chapter 8 Sections 8.4 and 8.5 report how SPECS satisfies each of these criteria, according to external evaluators and further analysis of the thesis content within this current work, respectively. The one criterion on which SPECS does not perform quite as well in is 'usability', because of the large quantity of data generated and the learning curve perceived by some evaluators in learning the components. Although use of the components is not mandatory within SPECS, it is highly recommended as a base model of creativity. To avoid the situation where the components are not adopted because of objections to their perceived quantity or complexity, Section 10.3 in this current Chapter considers the components in terms of how they can be simplified, reduced or presented differently without sacrificing important information, for greater clarity.*
  - Be applicable for evaluation and generate results in scenarios where time is limited and/or some desired resources are unavailable - *Yes - Chapter 7 illustrates how SPECS has been successfully applied in these scenarios, generating results from evaluation. Chapter 9 Sections 9.1.1 and 9.1.3 also consider these issues in the context of how SPECS (but not necessarily all other methodologies) can be applied in such scenarios and why such applications are important within research.*

* [The project must(*)/should] impact upon on the computational creativity research community, the target audience for this thesis's offerings, through:

  * Recognition of and interest in the work by members of the research community - *Yes - as described in Chapter 9 Section 9.3, several members of the research community have enquired about the methodology and the thesis work as a whole, both within computational creativity and in other related research areas such as computer music.*
  - Acceptance of publications on the work in venues relevant to computational creativity - *Yes (Jordanous, 2012, 2011a, as the publications of particular relevance and impact, to*

date).

- Citations of the work by other researchers - *Yes - the work has been cited several times by computational creativity researchers, as described in Chapter 9 Section 9.3.*

* Application of the methodology by other researchers - *No (Not yet) - to date, although details of the SPECS methodology have been requested a number of times, the SPECS methodology has not yet been applied by other researchers, though current discussions on this subject may lead to future applications in the near future.*
* Long-term adoption of the methodology by the community as the standard option for creativity evaluation - *Unknown - at the time of writing, the work has only been made available in publications since 2010 and SPECS itself has only been available since 2011, so it is too early to predict any long-term adoption of the SPECS methodology.*
* Encourage emphasis on a more focused, systematic, standardised approach to evaluation of computational creativity within the research community - *Unknown - the citations of this thesis work (described in Chapter 9 Section 9.3) show that the findings and proposals made in this thesis have placed some more emphasis on creativity evaluation approaches. As for the previous point, though, it is too early to determine what longer-term effects this thesis will have on approaches to evaluation within computational creativity.*

Failure-criteria for this project For an alternative perspective to the criteria outlined above for investigating the success of this project, a slightly different question can also be considered: the 'Popperian' question (Pease, 2012, personal communications) of what it would mean for this project to have failed to achieve the desired research outcomes.

- If no methodology is offered at all
- If the methodology cannot be applied in practice
- If the evaluation does not generate useful feedback
- If the methodology is restricted to only evaluating certain types of creative system
- If the methodology actually evaluates something other than creativity, a relatively common problem within existing creativity evaluation practice (Chapter 2 Section 2.3)
- If the methodology is difficult to apply
- If the methodology and/or ideas are not adopted by the community
- If the results produced are incorrect[18]

[18] As has often been shown in this thesis, it is misleading to aim for a 'right answer' in creativity or a 'ground truth'. It is more realistic to aim for an evaluation that has an informed basis. Either the evaluation should usually agree to at least some extent with a common consensus (where there is one), or, if the evaluation radically differs from others, there is justification for the differing evaluations that is based within an understanding of what it means for that type of system to be creative. The discussions for this point indicate support for the view adopted in this thesis that creativity evaluation feedback is more useful as formative feedback than in imposing some summative score or decision on how creative a system is.

- If the methodology generates a single summative score or rating but does not offer any formative feedback to guide future development

Each of these failure-criteria are covered in the consideration of the above success-criteria, as applied to SPECS. The above discussion shows evidence that SPECS does not fall foul of any of these criteria, with one possible exception: the adoption of SPECS and the associated emphasis on more standardised creativity evaluation by the computational creativity community. Future adoption of SPECS for evaluation cannot be predicted at this point due to the relative youth of SPECS, although SPECS has already received some attention from the community. It shall be seen in the future whether SPECS passes or fails the test of long-term adoption.[19]

10.6 Summary of contributions

This thesis addresses the research question: *How should we evaluate the creativity of a computational creativity system?* The Standardised Procedure for Evaluating Creative Systems (SPECS) methodology is proposed, applied and analysed as a solution.

The key contributions of this thesis are:

- The SPECS methodology for a *Standardised Procedure for Evaluating Creative Systems* (Chapter 5), with suggestions for implementations (Chapter 5) and demonstrations of applications (Chapters 6 and 7).
- A working definition of creativity in the form of a collection of key components (Chapter 4), recommended for use within SPECS.
- Evaluation data, comparisons and formative feedback on nine creative systems (Chapters 6 and 7), including a detailed analysis of creativity in musical improvisation (Chapter 6).
- A review and comparison of existing creativity evaluation frameworks in computational research, in theory (Chapter 2) and in practice (Chapter 8), from the perspective of a number of evaluators, both internal and external to the current doctoral project.
- Criteria for meta-evaluation of computational creativity evaluation methodologies (Chapter 8).
- A survey of current evaluative practice trends in computational creativity (Chapter 2).
- The bringing together of several different research disciplines and academic backgrounds on creativity to inform the work in this thesis (Chapters 3 and 4).
- Review of attitudes to computational creativity as obtained through an opinion survey (Chapter 8) and an overview of relevant research literature (Chapter 1).

Chapter 1 introduces this work, including an overview of computational creativity research. It motivates the need to address evaluation in computational creativity research, seen as a 'Grand Chal-

[19] An old adage expresses this sentiment as 'the proof of the pudding is in the eating'.

lenge' in computational creativity research. It outlines the central assumption in this thesis that computers can be creative, though resistance to computational creativity is acknowledged and considered. This Chapter also outlines how the thesis can be read most effectively depending on the reader's motivations and interests in the topic.

Chapter 2 reviews previous work done in the area of evaluating computational creativity, examining existing frameworks (particularly Pease et al., 2001; Ritchie, 2007; Colton, 2008b). Evaluative practice in computational creativity is shown to be in danger of falling into a 'methodological malaise' (Bundy, 1990) through lack of rigour, standardisation and systematic approach, and often through lack of any creativity evaluation whatsoever to justify claims of systems being creative.

Chapter 3 looks at issues in defining creativity. Clarification of the meaning of creativity is needed, as shown in this Chapter, but several issues exist that hinder such clarification. Existing definitions of creativity are examined, including dictionary, research and legal definitions. Different perspectives on creativity are explored, resulting in proposals of how creativity should be approached.

Chapter 4 derives an empirical definition of creativity from key contributions to the academic literature on creativity, using computational linguistics and machine learning techniques. This work is conducted on the premise that if a word is used significantly more often than expected in discussing a particular topic, then it is linked to the meaning of that topic; this is a fundamental assumption in the usage-based approach used in cognitive linguistics. The log likelihood ratio statistic is used to detect 694 such *creativity words*. Clustering techniques and inspection of the data results in the identification of 14 key aspects or components of creativity. These components are presented as *building blocks* that collectively construct the meaning of creativity.

Chapter 5 derives the SPECS methodological steps from heuristics-based guidelines for evaluation. Upon examining several practical issues involved in evaluating computational creativity, recommendations are made for how a methodology should be implemented. Evaluation Guidelines are set out for good evaluative practice. From these Evaluation Guidelines, the three-step SPECS methodology is derived and presented with suggestions and guidance for implementation in different computational creativity evaluation scenarios. The components derived in Chapter 4 are strongly recommended for use when performing SPECS evaluation.

Chapter 6 applies the SPECS methodology to evaluate in detail the creativity of three musical improvisation systems: GAmprovising (Jordanous, 2010c), GenJam (Biles, 2007) and Voyager (Lewis, 2000). To inform its definition of what makes a musical improvisation creative, this case study uses a questionnaire collecting the opinions of several people of varying improvisational experience from professional improviser to listener. In evaluation, six expert judges collectively give ratings on the systems' performance on each of the Chapter 4 components and these ratings are weighted using the data obtained from the questionnaire to reflect that components' relative contribution to creativity in

musical improvisation. The results find GenJam to be most creative overall. Perhaps more importantly, formative feedback is gathered for each system to highlight their strengths and weaknesses in exhibiting creativity.

Chapter 7 explores a case study where evaluations must be made with limited time and information. This second case study is intended to simulate the forming of first impressions about how creative something is. Five systems presented at ICCC'11 are evaluated on the basis of the information given in the 7-minute conference presentations (Cook & Colton, 2011; Rahman & Manurung, 2011; Norton et al., 2011; Tearse et al., 2011; Monteith et al., 2011). Although *DARCI* is identified as being slightly more creative in general, this case study's contributions lie more in the individual formative feedback for each system and also in reporting how effective the presentations were at delivering the most relevant information about each system's creativity.

Chapter 8 compares and contrasts the findings of SPECS on the two case studies with findings from various other methods of evaluation: surveys of human evaluation of the systems' creativity (including reflection on people's attitudes to carrying out such a task) and the application of the evaluation methodologies by Colton (2008b) and Ritchie (2007). Comparative results were more consistent for Case Study 1 than Case Study 2, showing the limited benefits in comparing systems from different domains as opposed to similar systems. In both cases, the formative feedback obtained from SPECS outweighed the other methods. Additionally, the task of using SPECS for evaluation was perceived as easier than consulting human opinion directly on the systems' creativity, as people found it difficult to evaluate the creativity of systems without a definition of creativity being supplied.

Chapter 9 Several critical points of reflection on SPECS arose during its implementation and evaluation and are outlined in this Chapter. Taking an alternative perspective to Chapter 8 on evaluation of the SPECS methodology, this Chapter also examines the reactions of the computational creativity research community (the methodology's main target audience) to earlier presentations of this work.

Chapter 10 (this Chapter) reprises the final version of the SPECS methodology and reviews what contributions the methodology and related work in this thesis makes, through the case study evaluations, the addressing of the underlying research question and the contributions to research made in both computational and human creativity research. Whilst the SPECS methodology has been shown to offer many benefits, further development work could prove fruitful in a few areas outlined in this Chapter. As a tool which is offered to a research community for long term use, some reflections are made on the promotion of SPECS to its target audience and the longevity of the methodology.

Figure 10.2: A quote from *The Book of the Courtier* (Castiglione, 1528), prominently displayed on the wall outside the Nightingale Theatre in Brighton, UK.

10.7 Final reflections

Outside one of the theatres in Brighton, UK, where the University of Sussex is based, there is a quote from *The Book of the Courtier* (Castiglione, 1528), pictured in Figure 10.2:

> 'That which we consider to be true art is that which appears not to be art at all.' (Castiglione, 1528)

Conte Castiglione posits 'true art' as something which cannot be captured in a definition or preconceived notion of what art is. Creativity researchers face a similar battle; once something labelled as creative is understood or defined, it loses some of its mysticism. The pervasive influence of the creative 'muse' from centuries ago (Chapter 3 Section 3.2.3) still affects perception of creativity now:

> 'Even if this mysterious phenomenon can be isolated, quantified, and dissected, why bother? Wouldn't it make more sense to revel in the mystery and wonder of it all?' (Hennessey & Amabile, 2010, p. 570)

Hennessey and Amabile go on to give various reasons why to 'bother' studying the 'mysterious phenomenon' of creativity. As von Hentig (1998) puts it, high expectations surround the weak term 'creativity'. There is much to be gained from pursuing a better understanding.

For those of us who tackle this ill-defined and highly subjective topic, tools to progress this research and make creativity more tangible to work with are highly valuable. In computational creativity research, the adoption of a standard and systematic method of evaluating progress is sorely needed. The Standardised Procedure for Evaluating Creative Systems is offered to meet this need.

Bibliography

Aguilar, A., Hernandez, D., Pérez y Pérez, R., Rojas, M., & Zambrano, M. d. L. (2008). A computer model for novel arrangements of furniture. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 157–162 Madrid, Spain.

Albert, R. S., & Runco, M. A. (1999). A history of research on creativity. In Sternberg, R. J. (Ed.), *Handbook of Creativity*, chap. 2, pp. 16–31. Cambridge University Press, Cambridge, MA.

Alvarado Lopez, J., & Pérez y Pérez, R. (2008). A computer model for the generation of monophonic musical melodies. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 117–126 Madrid, Spain.

Alvarez Cos, M., Perez y Perez, R., & Aliseda, A. (2007). A generative grammar for pre-hispanic production: The case of El Tajin style. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 39–46 London, UK.

Amabile, T. M. (1996). *Creativity in context*. Westview Press, Boulder, Colorado.

Ames, C. (1992). Quantifying musical merit. *Interface*, *21*(1), 53–93.

Babcock Gove, P. (Ed.). (1969). *Webster's Third New International Dictionary of the English Language (Unabridged)* (3rd edition). G. Bell & Sons (for G. & C. Merriam Co.), London, UK.

Bacon, F. (1878). *Bacon's* Novum Organum */ edited with introduction, notes, etc. by Thomas Fowler*. Clarendon press, Oxford, UK. Originally published by Bacon in 1620.

Baer, J. (1998). The case for domain specificity of creativity. *Creativity Research Journal*, *11*(2), 173–177.

Baer, J. (2010). Is creativity domain-specific?. In Kaufman, J. C., & Sternberg, R. J. (Eds.), *The Cambridge Handbook of Creativity*, chap. 17, pp. 321–341. Cambridge University Press, New York, NY.

Bailey, D. (1993). *Improvisation: Its nature and practice in music*. Da Capo Press, New York.

Barnhart, C. L. (Ed.). (1963). *The American College Dictionary*. Random House, New York, NY.

Bentkowska-Kafel, A. (2009). The fix vs. the flux: Which digital heritage?. In Daniels, D., & Reisinger, G. (Eds.), *Netpioneers 1.0 - archiving, representing and contextualising early netbased art*, pp. 145–162. Sternberg Press in association with the Ludwig Boltzmann Institute, Berlin, Germany / New York, NY.

Berkowitz, A. L., & Ansari, D. (2010). Expertise-related deactivation of the right temporoparietal junction during musical improvisation. *NeuroImage*, *49*(1), 712–719.

Berliner, P. F. (1994). *Thinking in jazz: the infinite art of improvisation*. Chicago Studies in Ethnomusicology. The University of Chicago Press, Chicago, IL.

Bickerman, G., Bosley, S., Swire, P., & Keller, R. M. (2010). Learning to create jazz melodies using deep belief nets. In *Proceedings of the International Conference on Computational Creativity*, pp. 228–237 Lisbon, Portugal.

Biemann, C. (2006). Chinese Whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pp. 73–80 Morristown, NJ. Association for Computational Linguistics.

Biles, J. A. (1994). GenJam: A genetic algorithm for generating jazz solos. In *Proceedings of the International Computer Music Conference* Denmark.

Biles, J. A. (2007). Improvising with genetic algorithms: *GenJam*. In Miranda, E. R., & Biles, J. A. (Eds.), *Evolutionary Computer Music*, chap. 7, pp. 137–169. Springer-Verlag, London, UK.

Binsted, K., Pain, H., & Ritchie, G. (1997). Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition*, *5*(2), 309–358.

Bird, A. (1998). *Philosophy of science*. UCL Press, London.

Bird, J., & Stokes, D. (2007). Minimal creativity, evaluation and fractal pattern discrimination. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 121–128 London, UK.

Bix, B. (1991). H . L . A . Hart and the "open texture" of language. *Law and Philosophy*, *10*(1), 51–72.

Blanke, T. (2011). *Using Situation Theory to evaluate XML retrieval*. Dissertations in Database and Information Systems. IOS Press, Heidelberg, Germany.

Blitstein, R. (2010). Triumph of the cyborg composer. *Miller-McCune Magazine: http://www.miller-mccune.com/culture-society/triumph-of-the-cyborg-composer-*8507, *last accessed March 2011*.

BNC Consortium (2007). The British National Corpus, v.3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: *http://www.natcorp.ox.ac.uk/*, last accessed September 2011.

Boden, M. A. (1990). *The creative mind: Myths and mechanisms*. Basic Books, Inc, New York.

Boden, M. A. (Ed.). (1994a). *Dimensions of creativity*. MIT Press, Cambridge, MA.

Boden, M. A. (1994b). What is creativity?. In Boden, M. A. (Ed.), *Dimensions of Creativity*, chap. 4, pp. 75–117. MIT Press, Cambridge, MA.

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence*, *103*, 347–356.

Boden, M. A. (1999). Introduction [summary of Boden's keynote address to AISB'99]. In *AISB Quarterly - Special issue on AISB99: Creativity in the arts and sciences*, Vol. 102, p. 11.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd edition). Routledge, London, UK.

Boulal, A., Iordanidis, M., & Quast, A. (2012). Remember, restructure, reuse adding value to Compound Scholarly Publications in a digital networked environment. In Delve, J., Anderson, D., Dobreva, M., Baker, D., Billenness, C., & Konstantelos, L. (Eds.), *The Preservation of Complex Objects*, Vol. 1. Visualisations and Simulations, pp. 151–163. University of Portsmouth, Portsmouth, UK.

Bown, O., & Wiggins, G. A. (2007). On the meaning of life (in artificial life approaches to music). In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 65–72 London, UK.

Bringsjord, S. (2000). *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS*. Lawrence Erlbaum Associates, London, UK.

Briscoe, E., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* Sydney, Australia.

Brown, A. R. (2010). Generation in context: Musical enquiry through algorithmic music making. COGS research seminar, School of Informatics, University of Sussex (January 2010).

Brown, D. (2009a). Computational artistic creativity and its evaluation. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Brown, P. (2009b). Autonomy, signature and creativity. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Bryan-Kinns, N. (2009). Everyday creativity. In *Proceedings of the 7th ACM conference on Creativity and Cognition* Berkeley, California.

Bundy, A. (1990). What kind of field is AI?. In Partridge, D., & Wilks, Y. (Eds.), *The Foundations of Artificial Intelligence*, pp. 215–222. Cambridge University Press, Cambridge, UK.

Bundy, A. (1994). What is the difference between real creativity and mere novelty?. *Behavioural and Brain Sciences*, *17*(3), 533–534.

Bushinsky, S. (2009). Deux ex machina - a higher creative species in the game of chess. *AI Magazine*, *30*(3), 63–70.

Butler, C. (1985). *Statistics in Linguistics*. Blackwell, Oxford, UK.

Byrne, W., Schnier, T., & Hendley, R. J. (2008). Computational intelligence and case-based creativity in design. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 31–40 Madrid, Spain.

Cardoso, A., Veale, T., & Wiggins, G. A. (2009). Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI Magazine*, *30*(3), 15–22.

Cardoso, A., & Wiggins, G. A. (Eds.). (2007). *Proceedings of the 4th International Joint Workshop on Computational Creativity*, London, UK. IJWCC, Goldsmiths, University of London.

Carroll, C. (2011). Robots get real. *National Geographic*, *220*(2), 66–85.

Castiglione, c. B. (1528). *Il cortegiano (tr. The book of the courtier)*. Aldine Press, Venice, Italy.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press, New York, NY.

Chan, H., & Ventura, D. (2008). Automatic composition of themed mood pieces. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 109–115 Madrid, Spain.

Chandrasekaran, B., Josephson, J., & Benjamins, V. (1999). What are ontologies, and why do we need them?. *IEEE Intelligent Systems and Their Applications*, *14*(1), 20–26.

Charnley, J., Pease, A., & Colton, S. (2012). On the notion of framing in computational creativity. In *Proceedings of the International Conference on Computational Creativity*, pp. 77–81 Dublin, Ireland.

Chordia, P., & Rae, A. (2010). Tabla gyan: An artificial tabla improviser. In *Proceedings of the International Conference on Computational Creativity*, pp. 155–164 Lisbon, Portugal.

Chuan, C.-H., & Chew, E. (2007). A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 57–64 London, UK.

Clifford, R. D. (2004). Random numbers, chaos theory, and cogitation: A search for the minimal creativity standard in copyright law. *Denver University Law Review*, *82*(2), 259–299.

Cohen, H. (1999). Colouring without seeing: A problem in machine creativity. In *AISB Quarterly - Special issue on AISB99: Creativity in the arts and sciences*, Vol. 102, pp. 26–35.

Cohen, H. (2009a). The art of self-assembly: the self-assembly of art. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Cohen, L. M. (2009b). A review of: "expanding visions of creative intelligence: An interdisciplinary exploration by Don Ambrose". *Creativity Research Journal*, *21*(2-3), 307–308.

Collins, N. (2008). The analysis of generative music programs. *Organised Sound*, *13*(3), 237–248.

Collins, N. (2011). Personal communications. In conversation.

Collins, N. (2012). Automatic composition of electroacoustic art music utilizing machine listening. *Computer Music Journal*, *36*(3), 8–23.

Collins, T., Laney, R., Willis, A., & Garthwaite, P. (2010). Music: Patterns and harmony using discovered, polyphonic patterns to filter computer-generated music. In *Proceedings of the International Conference on Computational Creativity*, pp. 1–10 Lisbon, Portugal.

Colton, S. (1999). Refactorable numbers-a machine invention. *Journal of Integer Sequences*, *2*(99.1), 2.

Colton, S. (2002). *Automated theory formation in pure mathematics*. Distinguished dissertations. Springer, London, UK.

Colton, S. (2008a). Computational creativity. *AISB Quarterly*, *126*, 6–7.

Colton, S. (2008b). Creativity versus the perception of creativity in computational systems. In *Proceedings of AAAI Symposium on Creative Systems*, pp. 14–20.

Colton, S. (2008c). Experiments in constraint-based automated scene generation. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 127–136 Madrid, Spain.

Colton, S., Bundy, A., & Walsh, T. (2000). On the notion of interestingness in automated mathematical discovery. *International Journal of Human-Computer Studies*, *53*, 351–375.

Colton, S., Charnley, J., & Pease, A. (2011). Computational Creativity Theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 90–95 Mexico City, Mexico.

Colton, S., de Mataras, R. L., & Stock, O. (2009). Computational creativity: Coming of age. *AI Magazine*, *30*(3), 11–14.

Colton, S., Goodwin, J., & Veale, T. (2012). Full-face poetry generation. In *Proceedings of the International Conference on Computational Creativity*, pp. 95–102 Dublin, Ireland.

Colton, S., Gow, J., Torres, P., & Cairns, P. (2010). Experiments in objet trouvé browsing. In *Proceedings of the International Conference on Computational Creativity* Lisbon, Portugal.

Colton, S., Pease, A., & Ritchie, G. (2001). The effect of input knowledge on creativity. In *Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science*.

Colton, S., & Steel, G. (1999). Artificial intelligence and scientific creativity. *Artificial Intelligence and the Study of Behaviour Quarterly*, *102*.

Cook, M., Colton, S., & Pease, A. (2012). Aesthetic considerations for automated platformer design. In *Eighth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)* Stanford, CA. AAAI.

Cook, M., & Colton, S. (2011). Automated collage generation - with more intent. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 1–3 Mexico City, Mexico.

Cope, D. (2005). *Computer Models of Musical Creativity*. MIT Press, Cambridge, MA.

Copyright, Designs and Patents Act (1988). UK Government Legislation. Ch. 48/1988.

Csikszentmihalyi, M. (1988). Society, culture, and person: a systems view of creativity. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, chap. 13, pp. 325–339. Cambridge University Press, Cambridge, UK.

Csikszentmihalyi, M. (1996). *Creativity: Flow and the psychology of discovery and invention*. Harper Perennial, New York.

Csikszentmihalyi, M. (2009). The creative person and the creative system (keynote address). In *Proceeding of the seventh ACM conference on Creativity and cognition*, pp. 5–6 Berkeley, California.

Darwin, C. (2004). Obituary: Christopher Longuet-Higgins. *The Guardian*, June 10th, 2004. Available at *http://www.guardian.co.uk/news/2004/jun/10/guardianobituaries.highereducation*, last accessed November 2012.

de Barros, D. P., Primi, R., Miguel, F. K., Almeida, L. S., & Oliveira, E. P. (2010). Metaphor creation: A measure of creativity or intelligence?. *European Journal of Educational Psychology*, *3*(1), 103–115.

de Melo, C. M., & Gratch, J. (2010). Evolving expression of emotions through color in virtual humans using genetic algorithms. In *Proceedings of the International Conference on Computational Creativity*, pp. 248–257 Lisbon, Portugal.

de Silva Garza, A. G., & Gero, J. (2010). Elementary social interactions and their effects on creativity: A computational simulation. In *Proceedings of the International Conference on Computational Creativity*, pp. 110–119 Lisbon, Portugal.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. Penguin Books, London, UK.

Dietrich, A., & Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin*, *136*(5), 822–848.

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*(4), 247–255.

Dreyfus, H. L. (1979). *What computers can't do: The limits of Artificial Intelligence* (Revised edition). Harper & Row, London, UK.

Dunning, A. (2006). The tasks of the AHDS: Ten years on. *Ariadne*, *July*(48), 1–6.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Edmonds, D., & Eidinow, J. (2001). *Wittgenstein's Poker*. Faber and Faber.

Edmonds, E. (2009a). Statement. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Edmonds, E. (2009b). Words on the creativity and cognition studios contribution. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Eigenfeldt, A., Burnett, A., & Pasquier, P. (2012). Evaluating musical metacreation in a live performance context. In *International Conference on Computational Creativity*, p. 140 Dublin, Ireland.

Eigenfeldt, A., & Pasquier, P. (2010). Realtime generation of harmonic progressions using constrained markov selection. In *Proceedings of the International Conference on Computational Creativity*, pp. 16–25 Lisbon, Portugal.

Elliott, R. K. (1971). Versions of creativity. *Journal of Philosophy of Education*, *5*(2), 139–152.

Erickson, B. H., & Nosanchuk, T. A. (1992). *Understanding Data* (2nd edition). Open University Press, Buckingham, UK.

Evans, V., & Green, M. (2006). *Cognitive linguistics: An introduction*. Edinburgh University Press, Edinburgh, UK.

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities* (1st edition). Basic Books, Inc, New York.

Feist (1991). Feist Publications, Inc. v. Rural Telephone Service Co. *499 US 340*, *111 S. Ct 1282, 113 L. Ed. 2d 358*(Supreme Court).

Feyerabend, P. K. (1993). *Against method* (3rd edition). Verso, London, UK.

Finke, R., Ward, T., & Smith, S. (1992). *Creative Cognition: Theory, research and applications*. MIT Press, Cambridge, MA.

Forth, J., McLean, A., & Wiggins, G. (2008). Musical creativity on the conceptual level. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 21–30 Madrid, Spain.

Forth, J., Wiggins, G. A., & McLean, A. (2010). Unifying conceptual spaces: Concept formation in musical creative systems. *Minds and Machines*, *20*(4), 503–532.

Friis-Olivarius, M., Wallentin, M., & Vuust, P. (2009). Improvisation - the neural foundation for creativity (poster). In *Proceedings of the 7th ACM Creativity and Cognition Conference*, pp. 411–412 Berkeley, California.

Gabora, L. (2011). If experts converge on the same answer are they less creative than beginners? redefining creativity in terms of adaptive landscapes. *arXiv.org*, *arXiv preprint*(1106.2265).

Gallie, W. B. (1956). Essentially contested concepts. *Proceedings of the Aristotelian Society*, *56*, 167–198.

Gardner, H. (1993). *Creating minds*. Basic Books, New York, NY.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(60), 471–479.

Garmonsway, G. N., & Simpson, J. (Eds.). (1969). *The Penguin English dictionary* (2nd edition). Penguin Books, London, UK.

Gero, J. (2010). Novel associations. In *Proceedings of the International Conference on Computational Creativity* Lisbon, Portugal.

Gervás, P. (2000). WASP: Evaluation of different strategies for the automatic generation of Spanish verse. In *Proceedings of AISB Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science*, pp. 93–100 Birmingham, UK.

Gervás, P. (2002). Exploring quantitative evaluations of the creativity of automatic poets. In *Proceedings of the 2nd. Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science (ECAI 2002)*.

Gervas, P. (2009). Computational approaches to storytelling and creativity. *AI Magazine*, *30*(3), 49–62.

Gervás, P., & León, C. (2010). Story generation driven by system-modified evaluation validated by human judges. In *Proceedings of the International Conference on Computational Creativity*, pp. 85–89 Lisbon, Portugal.

Gervás, P., Perez y Perez, R., Sosa, R., & Lemaitre, C. (2007). On the fly collaborative story-telling: Revising contributions to match a shared partial story line. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 13–20 London, UK.

Getzels, J., & Jackson, P. (1962). *Creativity and intelligence*. Wiley, New York, NY.

Gibbs, L. (2010). Evaluating creative (jazz) improvisation: Distinguishing invention and creativity. In *Proceedings of Leeds International Jazz Conference 2010: Improvisation - jazz in the creative moment* Leeds, UK.

Gillick, J., Tang, K., & Keller, R. M. (2010). Machine learning of jazz grammars. *Computer Music Journal*, *34*(3), 56–66.

Goldman, R. J. (1964). The Minnesota tests of creative thinking. *Educational Research*, *7*(1), 3–14.

Gomes, P., Seco, N., Pereira, F. C., Paiva, P., Carreiro, P., Ferreira, J. L., & Bento, C. (2006). The importance of retrieval in creative design analogies. *Knowledge-Based Systems*, *19*(7), 480 – 488.

Goto, M., & Hirata, K. (2004). Recent studies on music information processing. *Acoust Sci & Tech*, *25*(6), 419–425.

Grace, K., Saunders, R., & Gero, J. (2008). A computational model for constructing novel associations. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 91–100 Madrid, Spain.

Grace, K., Saunders, R., & Gero, J. (2010). Constructing conceptual spaces for novel associations. In *Proceedings of the International Conference on Computational Creativity*, pp. 120–129 Lisbon, Portugal.

Guilford, J. P. (1950). Creativity. *American Psychologist*, *5*, 444–454.

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill, New York, NY.

Hacking, I. (1981a). Lakatos's philosophy of science. In Hacking, I. (Ed.), *Scientific revolutions*, chap. VI, pp. 128–143. Oxford University Press, Oxford, UK.

Hacking, I. (Ed.). (1981b). *Scientific revolutions*. Oxford University Press, Oxford, UK.

Hadamard, J. (1945). *An Essay on the Psychology of Invention in the Mathematical Field*. Princeton University Press, Princeton, NJ.

Haenen, J., & Rauchas, S. (2006). Investigating artificial creativity by generating melodies, using connectionist knowledge representation. In *Proceedings of the 3rd International Joint Workshop on Computational Creativity (ECAI06 Workshop)* Riva del Garda, Italy.

Haiman, J. (1980). Dictionaries and encyclopedias. *Lingua*, *4*, 329–357.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Hassan, S., Gervás, P., León, C., & Hervas, R. (2007). A computer model that generates biography-like narratives. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 5–12 London, UK.

Heilman, K. M. (2005). *Creativity and the Brain*. Psychology Press, Hove, UK.

Hennessey, B. A., & Amabile, T. M. (1988). The conditions of creativity. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, chap. 1, pp. 11–38. Cambridge University Press, Cambridge, UK.

Hennessey, B. A., & Amabile, T. M. (2010). Creativity. *Annual Review of Psychology*, *61*, 569–598.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Hodgson, P. (1998). Music is all in the mind sounding off. *Sound on Sound online: http://www.soundonsound.com/sos/oct98/articles/soundoff.htm, last accessed March 2011.*

Hodgson, P. (2006a). The evolution of melodic complexity in the music of Charles Parker. In Baroni, M., Addessi, A. R., Caterina, R., & Costa, M. (Eds.), *Proceedings of the 9th International Conference on Music Perception and Cognition (ICMPC9)*, pp. 997–1002 Bologna, Italy.

Hodgson, P. (2006b). Learning and the evolution of melodic complexisty in virtuoso jazz improvisation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*, pp. 1506–1510 Vancouver, Canada.

Hodgson, P. (2006c). *Modelling cognition in creative musical improvisation*. Ph.D. thesis, Department of Informatics, University of Sussex, Brighton, UK.

Holmes, N. (2009). The automation of originality: When originality is automated, what becomes of personality?. *Computer*, *March 2009*, 98–100.

Hong, N. C., Crouch, S., Hettrick, S., Parkinson, T., & Shreeve, M. (2010). Software preservation: Benefits framework. Tech. rep., Software Sustainability Institute and Curtis & Cartwright Consulting Ltd.

Hull, M., & Colton, S. (2007). Towards a general framework for program generation in creative domains. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 137–144 London, UK.

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, *19*(1), 1–64.

Huron, D. (2006). *Sweet Anticipation*. MIT Press, Cambridge, MA.

ICCC'10 (2010). Proceedings of the International Conference for Computational Creativity, Lisbon, Portugal. *http://eden.dei.uc.pt/ amilcar/ftp/e-Proceedings_ICCC-X.pdf*, last accessed May 2012.

ICCC'11 (2011). Proceedings of the International Conference for Computational Creativity, Mexico City, Mexico. *http://iccc11.cua.uam.mx/proceedings/*, last accessed May 2012.

IJWCC'07 (2007). Proceedings of the International Joint Workshop for Computational Creativity, London, UK. *http://doc.gold.ac.uk/isms/CC07/CC07Proceedings.pdf*, last accessed May 2012.

Intellectual Property Office (2010). Intellectual Property Office Website. *http://www.ipo.gov.uk*, last accessed September 2011.

Ivcevic, Z. (2009). Creativity map: Toward the next generation of theories of creativity.. *Psychology of Aesthetics, Creativity, and the Arts*, *3*(1), 17–21.

Jackson, P. W., & Messick, S. (1967). The person, the product, and the response: Conceptual problems in the assessment of creativity. In Kagan, J. (Ed.), *Creativity and Learning*, pp. 1–19. Beacon Press, Boston.

Jennings, K. E. (2010a). Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines*, *20*(4), 489–501.

Jennings, K. E. (2010b). Search strategies and the creative process. In *Proceedings of the International Conference on Computational Creativity*, pp. 130–139 Lisbon, Portugal.

Jick, T. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, *24*(4), 602–611.

Johnson-Laird, P. (2002). How jazz musicians improvise. *Music Perception*, *19*(3), 415–442.

Jordanous, A., & Keller, B. (2012). Weaving creativity into the semantic web: a language-processing approach. In *International Conference on Computational Creativity*, pp. 216–220.

Jordanous, A. (2009). Evaluating machine creativity. In *Proceedings of the 7th ACM Creativity and Cognition Conference*, pp. 331–332 Berkeley, California.

Jordanous, A. (2010a). Defining creativity: Finding keywords for creativity using corpus linguistics techniques. In *Proceedings of the International Conference on Computational Creativity*, pp. 278–287 Lisbon, Portugal.

Jordanous, A. (2010b). Defining creativity in music improvisation. In *Proceedings of the Empirical Musicology II Conference* Leeds, UK.

Jordanous, A. (2010c). A fitness function for creativity in jazz improvisation and beyond. In *Proceedings of the International Conference on Computational Creativity*, pp. 223–227 Lisbon, Portugal.

Jordanous, A. (2011a). Evaluating evaluation: Assessing progress in computational creativity research. In *Proceedings of the Second International Conference on Computational Creativity (ICCC-11)* Mexico City, Mexico.

Jordanous, A. (2011b). International Conference for Computational Creativity: A review. *AISB Quarterly*, *128*.

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, *4*(3), 246–279.

Jordanous, A., & Keller, B. (2011). What makes a musical improvisation creative? (extended abstract). In *Proceedings of the 7th Conference on Interdisciplinary Musicology* Glasgow, UK.

Kagan, J. (Ed.). (1967). *Creativity and learning*. Beacon Press Press, Boston, MA.

Karjala, D. S. (2008). Copyright and creativity. *UCLA Entertainment Law Review*, *15*, 169–201.

Kaufman, G. (2003). What to measure? A new look at the concept of creativity. *Scandinavian Journal of Educational Research*, *47*(3), 235–251.

Kaufman, J. C. (2009). *Creativity 101*. The Psych 101 series. Springer, New York.

Kaufman, J. C., Kaufman, S. B., & Lichtenberger, E. O. (2011). Finding creative potential on intelligence tests via divergent production finding creative potential on intelligence tests via divergent production finding creative potential on intelligence tests via divergent production. *Canadian Journal of School Psychology*, *26*(2), 83–106.

Kazai, G., & Lalmas, M. (2005). Notes on what to measure in INEX. In *INEX 2005 Workshop on Element Retrieval Methodology* Glasgow, UK.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133.

Kilgarriff, A. (2006). Where to go if you would like to find out more about a word than the dictionary tells you. *Macmillan English Dictionary Magazine*, *Issue 35 (Jan-Feb)*.

Koestler, A. (1964). *The act of creation*. Danube Books, New York.

Kraft, U. (2005). Unleashing creativity. *Scientific American Mind*, *April*.

Krzeczkowska, A., El-Hage, J., Colton, S., & Clark, S. (2010). Automated collage generation - with intent. In *Proceedings of the International Conference on Computational Creativity*, pp. 36–40 Lisbon, Portugal.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL.

Lakatos, I. (1978). *The methodology of scientific research programmes*, Vol. 1 of Philosophical Papers (edited by John Worrall and Gregory Currie). Cambridge University Press, Cambridge, UK.

Lakoff, G. (1987). *Women, Fire and Dangerous things: What Categories reveal about the mind*. University of Chicago Press, Chicago, IL.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago, IL.

León, C., de Albornoz, J. C., & Gervás, P. (2008). A framework for building creative objects from heterogenous generation systems. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 71–80 Madrid, Spain.

León, C., & Gervás, P. (2008). Creative storytelling based on exploration and transformation of constraint rules. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 51–60 Madrid, Spain.

León, C., & Gervás, P. (2010). The role of evaluation-driven rejection in the successful exploration of a conceptual space of stories. *Minds and Machines*, *20*(4), 615–634.

Lewis, G. E. (2000). Too many notes: Computers, complexity and culture in Voyager. *Leonardo Music Journal*, *10*, 33–39.

Lewis, G. E. (2011). Improvising with creative machines: Reflections on human-machine interaction (keynote talk). In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. xii–xiii Mexico City, Mexico.

Lewis, M. R. (2009). Casually evolving creative technology systems. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

LeWitt, S. (1967). Paragraphs on conceptual art. *Artforum International Magazine*, *June 1967*.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304 Madison, WI.

López, A. R., Oliveira, A. P., & Cardoso, A. (2010). Real-time emotion-driven music engine. In *Proceedings of the International Conference on Computational Creativity*, pp. 150–154 Lisbon, Portugal.

Lovelace, A. (1953). Notes on manabrea's sketch of the analytical engine invented by charles babbage. In Bowden, B. (Ed.), *Faster than thought : a symposium on digital computing machines*. Pitman, London.

MacDonald, R., Byrne, C., & Carlton, L. (2006). Creativity and flow in musical composition: An empirical investigation. *Psychology of Music*, *34*, 292–306.

Macedo, L., & Cardoso, A. (2002). Assessing creativity: the importance of unexpected novelty. In *Proceedings of the 2nd. Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science, 15th European Conference on Artificial Intelligence (ECAI 2002)*, pp. 31–38.

Machado, P., & Cardoso, A. (2002). All the truth about NEvAr. *Applied Intelligence*, *16*(2), 101–118.

Machado, P., & Nunes, H. (2010). A step towards the evolution of visual languages. In *Proceedings of the International Conference on Computational Creativity*, pp. 41–50 Lisbon, Portugal.

Machado, P., Romero, J., Manaris, B., Santos, A., & Cardoso, A. (2003). Power to the critics. In *Proceedings of the 3rd Workshop on Creative Systems: Approaches to Creativity in Artificial Intelligence and Cognitive Science*, pp. 55–64 Acapulco, Mexico. IJCAI.

MacKinnon, D. W. (1970). Creativity: a multi-faceted phenomenon. In Roslansky, J. D. (Ed.), *Creativity: A Discussion at the Nobel Conference*, pp. 17–32. North-Holland Publishing Company, Amsterdam, The Netherlands.

Mandel, G. N. (2011). To promote the creative process: Intellectual property law and the psychology of creativity. *Notre Dame Law Review (online)*, *86*(1).

Manurung, H. M. (2003). *An Evolutionary Algorithm Approach to Poetry Generation*. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

Masters, J. (2011). Vocal improvisation - a tool to access the most fundamental levels of individual musicality. In *The Improvised Space: Techniques, Traditions and Technologies* London, UK.

Mayer, R. E. (1999). Fifty years of creativity research. In Sternberg, R. J. (Ed.), *Handbook of Creativity*, chap. 22, pp. 449–460. Cambridge University Press, Cambridge, UK.

McCarthy, J. (1979). Ascribing mental qualities to machines. In Ringle, M. (Ed.), *Philosophical Perspectives in Artificial Intelligence*. Humanities Press, Atlantic Highlands, NJ.

McCorduck, P. (1991). *Aaron's code: Meta-art, artificial intelligence, and the work of Harold Cohen*. WH Freeman, New York, NY.

McCormack, J. (2007). Creative ecosystems. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 129–136 London, UK.

McCormack, J. (2009). Creative ecosystems. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

McCrae, R. R., & Costa Jr, P. T. (1999). A five-factor theory of personality. In Pervin, L. A., & John, O. P. (Eds.), *Handbook of personality: theory and research* (2nd edition)., chap. 5, pp. 139–153. The Guilford Press, New York.

McGraw, G., & Hofstadter, D. (1993). Perception and creation of diverse alphabetic styles. *AISB Quarterly*, *85*, 42–49.

McGreggor, K., Kunda, M., & Goel, A. (2010). A fractal approach towards visual analogy. In *Proceedings of the International Conference on Computational Creativity*, pp. 65–74 Lisbon, Portugal.

McLean, A. (2009). Embodied creativity. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

McLean, A., & Wiggins, G. (2010). Live coding towards computational creativity. In *Proceedings of the International Conference on Computational Creativity*, pp. 175–179 Lisbon, Portugal.

Mednick, S. A. (1967). *The Remote Associates Test*. Houghton Mifflin Company, Boston.

Meehan, J. (1981). Tale-Spin. In Schank, R. C., & Riesbeck, C. K. (Eds.), *Inside computer understanding: five programs plus minatures*. Lawrence Erlbaum Associates, Hillside, NJ.

Meehan, J. R. (1976). *The metanovel: writing stories by computer.* Ph.D. thesis, Yale University, New Haven, CT, USA.

Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176.

Minsky, M. L. (1982). Why people think computers can't. *AI Magazine*, *3*(4), 3.

Miranda, E. R., & Biles, J. A. (Eds.). (2007). *Evolutionary computer music*. Springer-Verlag, London, UK.

Moffat, D. C., & Kelly, M. (2006). An investigation into people's bias against computational creativity in music composition. In *Proceedings of the 3rd International Joint Workshop on Computational Creativity (ECAI06 Workshop)* Riva del Garda, Italy.

Monteith, K., Francisco, V., Martinez, T., Gervás, P., & Ventura, D. (2011). Automatic generation of emotionally-targeted soundtracks. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 60–62 Mexico City, Mexico.

Monteith, K., Martinez, T., & Ventura, D. (2010). Automatic generation of music for inducing emotive response. In *Proceedings of the International Conference on Computational Creativity*, pp. 140–149 Lisbon, Portugal.

Montfort, N., & Pérez y Pérez, R. (2008). Integrating a plot generator and an automatic narrator to create and tell stories. In *Proceedings of the 5th International Joint Workshop on Computational Creativity* Madrid, Spain.

Mooney, R. L. (1963). A conceptual model for integrating four approaches to the identification of creative talent. In Taylor, C. W., & Barron, F. (Eds.), *Scientific Creativity: Its Recognition and Development*, chap. 27, pp. 331–340. John Wiley & Sons, New York.

Mornoi, A., Zuben, F., & Manzolli, J. (2002). ArTbitration: Human-machine interaction in artistic domains. *Leonardo*, *35*(2), 185–188.

Morris, R., Burton, S., Bodily, P., & Ventura, D. (2012). Soup over bean of pure joy: Culinary ruminations of an artificial chef. In *International Conference on Computational Creativity*, p. 119.

Mozetič, I., Lavrač, N., Podpečan, V., Novak, P. K., Motaln, H., Petek, M., Gruden, K., Toivonen, H., & Kulovesi, K. (2010). Bisociative knowledge discovery for microarray data analysis. In *Proceedings of the International Conference on Computational Creativity*, pp. 190–199 Lisbon, Portugal.

Nake, F. (2009). Creativity in algorithmic art. In *Proceeding of the 7th ACM conference on Creativity and Cognition*, pp. 97–106 Berkeley, California.

Nelson, C., Brummel, B., Grove, F., Jorgenson, N., Sen, S., & Gamble, R. (2010). Measuring creativity in software development. In *Proceedings of the International Conference on Computational Creativity*, pp. 205–214 Lisbon, Portugal.

Newell, A., Shaw, J. G., & Simon, H. A. (1963). The process of creative thinking. In Gruber, H. E., Terrell, G., & Wertheimer, E. (Eds.), *Contemporary Approaches to Creative Thinking*, pp. 63–119. Atherton, New York.

Nichols, S. (2009). Time to change our thinking: Dismantling the silo model of digital scholarship. *Ariadne*, *58*.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, pp. 152–158 Boston, MA.

Noll, A. M. (1994). The beginnings of computer art in the United States: A memoir. *Leonardo*, *27*(1), 39–44.

Norton, D., Heath, D., & Ventura, D. (2010). Establishing appreciation in a creative system. In *Proceedings of the International Conference on Computational Creativity*, pp. 26–35 Lisbon, Portugal.

Norton, D., Heath, D., & Ventura, D. (2011). Autonomously creating quality images. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 10–15 Mexico City, Mexico.

Norton, D., Heath, D., & Ventura, D. (2012). Finding creativity in an artificial artist. *Journal of Creative Behaviour*, *forthcoming (2012)*.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh, UK.

Odena, O., & Welch, G. (2009). A generative model of teachers' thinking on musical creativity. *Psychology of Music*, *37*(4), 416–442.

O'Donoghue, D. P. (2007). Evaluating computer-generated analogies. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 31–38 London, UK.

O'Donoghue, D. P., Bohan, A., & Keane, M. T. (2006). Seeing things: Inventive reasoning with geometric analogies and topographic maps. *New Generation Computing*, *24*(3), 267–288.

Oliveira, H., Cardoso, A., & Pereira, F. C. (2007). Tra-la-lyrics: An approach to generate text based on rhythm. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 47–54 London, UK.

Pachet, F. (2004). Beyond the cybernetic jam fantasy: the continuator. *Computer Graphics and Applications, IEEE*, *24*(1), 31 –35.

Parker, E. (2011). Drifting on a reed (keynote presentation). In *The Improvised Space: Techniques, Traditions and Technologies* London, UK.

Pearce, M., & Wiggins, G. (2001). Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science* York, UK.

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. thesis, Department of Computing, City University, London, UK.

Pearce, M. T., Meredith, D., & Wiggins, G. A. (2002). Motivations and methodologies for automation of the compositional process. *Musicae Scientae*, *6*(2), 119–147.

Pearce, M. T., & Wiggins, G. A. (2007). Evaluating cognitive models of musical composition. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 73–80 London, UK.

Pease, A., Charnley, J., & Colton, S. (2012). Using grounded theory to suggest types of framing information for computational creativity. In Besold, T. R., Kühnberger, K.-U., Schorlemmer, M., & Smaill, A. (Eds.), *International Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI) at ECAI 2012*, pp. 7–13 Montpellier, France. Publications of the Institute of Cognitive Science, Osnabrück, Germany.

Pease, A., & Colton, S. (2011a). Computational Creativity Theory: Inspirations behind the FACE and the IDEA models. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 72–77 Mexico City, Mexico.

Pease, A., & Colton, S. (2011b). On impact and evaluation in computational creativity: A discussion of the Turing Test and an alternative proposal. In *Proceedings of the AISB'11 Convention* York, UK. AISB.

Pease, A., Guhe, M., & Smaill, A. (2010). Some aspects of analogical reasoning in mathematical creativity. In *Proceedings of the International Conference on Computational Creativity*, pp. 60–64 Lisbon, Portugal.

Pease, A., Winterstein, D., & Colton, S. (2001). Evaluating machine creativity. In *Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science*, pp. 129–137.

Peinado, F., Francisco, V., Hervás, R., & Gervás, P. (2010). Assessing the novelty of computer-generated narratives using empirical metrics. *Minds and Machines*, *20*(4), 565–588.

Peinado, F., & Gervas, P. (2006). Evaluation of automatic generation of basic stories. *New Generation Computing*, *24*(3), 289–302.

Pereira, F. C., & Cardoso, A. (2006). Experiments with free concept generation in Divago. *Knowledge-Based Systems*, *19*(7), 459 – 470.

Pereira, F. C., Mendes, M., Gervás, P., & Cardoso, A. (2005). Experiments with assessment of creative systems: An application of Ritchie's criteria. In *Proceedings of the Workshop on Computational Creativity (IJCAI 05)*.

Pérez y Pérez, R., Aguilar, A., & Negrete, S. (2010). The ERI-Designer: A computer model for the arrangement of furniture. *Minds and Machines*, *20*(4), 533–564.

Pérez y Pérez, R. (1999). *MEXICA: A Computer Model of Creativity in Writing*. Ph.D. thesis, University of Sussex, Brighton, UK.

Pérez y Pérez, R., Negrete, S., Penãlosa, E., Aṽila, R., Castellanos, V., & Lemaitre, C. (2010). MEXICA-Impro: A computational model for narrative improvisation. In *Proceedings of the International Conference on Computational Creativity*, pp. 90–99 Lisbon, Portugal.

Pérez y Pérez, R., & Sharples, M. (2004). Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-Based Systems*, *17*(1), 15–29.

Perkins, D. N. (1994). Creativity: Beyond the Darwinian paradigm. In Boden, M. A. (Ed.), *Dimensions of Creativity*, chap. 5, pp. 119–142. MIT Press, Cambridge, MA.

Plucker, J. A. (1998). Beware of simple conclusions: The case for content generality of creativity. *Creativity Research Journal*, *11*(2), 179–182.

Plucker, J. A., & Beghetto, R. A. (2004). Why creativity is domain general, why it looks domain specific, and why the distinction doesn't matter. In Sternberg, R. J., Grigorenko, E. L., & Singer, J. L. (Eds.), *Creativity: From Potential to Realization*, chap. 9, pp. 153–167. American Psychological Association, Washington, DC.

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, *39*(2), 83–96.

Poincaré, H. (1929). Mathematical creation. In *The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method.*, Vol. Science and Method [Original French version published 1908, Authorized translation by George Bruce Halsted], chap. III of Book I. Science and the Scientist, pp. 383–394. The Science Press, New York.

Pollack, I., & Pickett, J. (1957). Cocktail party effect. *The Journal of the Acoustical Society of America*, *29*, 1262.

Poole, S. (1998). Silicon swing. *http://stevenpoole.net/articles/silicon-swing/*, *last accessed September 2011*.

Popper, K. (1972). *The Logic of Scientific Discovery* (3rd edition). Hutchinson, London.

Porter, B. (2009). Simulating morphogenesis. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Potts, H. E. (1944). The definition of invention in patent law. *The Modern Law Review*, *7*(3), 113–123.

Pressing, J. (1987). Improvisation: Methods and models. In Sloboda, J. A. (Ed.), *Generative Processes in Music: The Psychology of Performance, Improvisation and Composition*, chap. 7, pp. 129–178. Oxford University Press, Oxford, UK.

Rahman, F., & Manurung, R. (2011). Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 4–9 Mexico City, Mexico.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of ACL Workshop on Comparing Corpora* Hong Kong.

Reeves, B., & Nass, C. (2002). *The media equation: how people treat computers, television, and new media like real people and places*. CSLI Publications, Stanford, CA.

Reffin-Smith, B. (2010). 43 dodgy statements on computer art. http://zombiepataphysics.blogspot.com/2010/03/43-dodgy-statements-on-computer-art.html, last accessed July 2010.

Resnick, M. (2007). Sowing the seeds for a more creative society. *Learning and Leading with Technology*, *35*(4).

Rhodes, M. (1961). An analysis of creativity. *Phi Delta Kappan*, *42*(7), 305–310.

Riedl, M. O. (2010). Story planning: Creativity through exploration, retrieval, and analogical transformation. *Minds and Machines*, *20*(4), 589–614.

Riedl, M. O. (2008). Vignette-based story planning: Creativity through exploration and retrieval. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 41–50 Madrid, Spain.

Riedl, M. O., & Young, R. M. (2006). Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing*, *24*(3), 303–323.

Ritchie, G. (2001). Assessing creativity. In *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*, pp. 3–11 York, UK.

Ritchie, G. (2006). The transformational creativity hypothesis. *New Generation Computing*, *24*(3), 241–266.

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, *17*, 67–99.

Ritchie, G. (2008). Uninformed resource creation for humour simulation. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 147–150 Madrid, Spain.

Ritchie, G., Manurung, R., Pain, H., Waller, A., Black, R., & O'Mara, D. (2007). A practical application of computational humour. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 91–98 London, UK.

Robey, D. (2011). Introduction to Digital Humanities. Digital Humanities workshop, University of Reading, UK (September 2011).

Robinson, K. (2006). TED2006 talk: Ken Robinson says schools kill creativity. *http://www.ted.com/talks/ken_robinson_says_schools_kill_creativity.html*, last accessed September 2011.

Robinson, M. (Ed.). (1999). *Chambers 21st century dictionary*. Chambers Harrap, Edinburgh, UK.

Romero, P., & Calvillo-Gamez, E. (2011). Towards an embodied view of flow. In *Proceedings of the 2nd International Workshop on User Models for Motivational Systems: the affective and the rational routes to persuasion (UMMS 2011)*, pp. 100–105 Girona, Spain.

Rothenburg, J. (1999). Ensuring the longevity of digital information. Available at *http://www.clir.org/pubs/archives/ensuring.pdf* (last accessed November 2012). Expanded version of the article *Ensuring the Longevity of Digital Documents*, that appeared in the January 1995 edition of Scientific American (Vol. 272, Number 1, pp. 42-7).

Rowlands, S. (2011). Discussion article: Disciplinary boundaries for creativity. *Creative Education*, *2*(1), 47–55.

Runco, M. (2003). Creativity, cognition, and their educational implications. In Houtz, J. (Ed.), *The educational psychology of creativity*, pp. 25–56. Hampton Press, New York, NY.

Runco, M., Dow, G., & Smith, W. (2006). Information, Experience, and Divergent Thinking: An Empirical Test.. *Creativity research journal*, *18*(3), 267–277.

Russell, B. (1912). *The Problems of Philosophy*. Williams & Northgate, London, UK.

Saunders, R. (2012). Towards autonomous creative systems: A computational approach. *Cognitive Computation*, *4*(3), 216–225.

Saunders, R. (2009). Artificial creative systems: Completing the creative cycle. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Saunders, R., Gemeinboeck, P., Lombard, A., Bourke, D., & Kocaballi, B. (2010). Curious whispers: An embodied artificial creative system. In *Proceedings of the International Conference on Computational Creativity*, pp. 100–109 Lisbon, Portugal.

Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the CHI'05 conference on Human Factors in Computing Systems* Portland, OR.

Schmid, K. (1996). Making AI systems more creative: The IPC-model. *Knowledge-Based Systems*, *9*(6), 385–397.

Searle, J. R. (1980). Minds, brains, and programs. In Haugeland, J. (Ed.), *Mind design II: Philosophy, Psychology, Artificial Intelligence* (2nd edition)., Vol. 3, pp. 417–457. MIT Press, Cambridge, MA.

Seth, A. K. (2009). Explanatory correlates of consciousness: Theoretical and computational challenges. *Cognitive Computation*, *1*, 50–63.

Simonton, D. K. (1988). Creativity, leadership, and chance. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, chap. 16, pp. 386–426. Cambridge University Press, Cambridge, UK.

Simpson, J. A., & Weiner, E. S. C. (Eds.). (1989). *The Oxford English Dictionary* (2nd edition)., Vol. III. Oxford University Press, Oxford, UK.

Sinclair, J. M. (Ed.). (1992). *B.B.C. English Dictionary*. Harper Collins, Glasgow, UK.

Sloman, A. (1978). *The computer revolution in philosophy*. Harvester Press, Hassocks, Sussex.

Snyder, A., Mitchell, J., Bossomaier, T., & Pallier, G. (2004). The Creativity Quotient: An objective scoring of ideational fluency. *Creativity Research Journal*, *16*(4), 415–420.

Sorenson, N., & Pasquier, P. (2010). The evolution of fun: Automatic level design through challenge modeling. In *Proceedings of the International Conference on Computational Creativity*, pp. 258–267 Lisbon, Portugal.

Sosa, R., Gero, J., & Jennings, K. (2009). Growing and destroying the worth of ideas. In *Proceedings of the 7th ACM Creativity and Cognition conference*, pp. 295–304 Berkeley, California.

Spector, L., & Alpern, A. (1994). Criticism, culture, and the automatic generation of artworks. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, pp. 3–8 Menlo Park, CA and Cambridge, MA. AAAI Press/The MIT Press.

Spivak, J. (1996). Who needs a band when you have software?. *Democrat and Chronicle, Rochester NY, reproduced at http://www.ist.rit.edu/ jab/Spevak.html,, Last accessed September 2011*(2).

Stein, M. I. (1963). A transactional approach to creativity. In Taylor, C. W., & Barron, F. (Eds.), *Scientific Creativity: Its Recognition and Development*, chap. 18, pp. 217–227. John Wiley & Sons, New York.

Sternberg, R. J. (1988). A three-facet model of creativity. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, chap. 5, pp. 125–147. Cambridge University Press, Cambridge, UK.

Sternberg, R. J. (Ed.). (1999). *Handbook of Creativity*. Cambridge University Press, Cambridge, UK.

Sternberg, R. J., Grigorenko, E. L., & Singer, J. L. (Eds.). (2004). *Creativity: From potential to realization*. American Psychological Association, Washington, DC.

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. In Sternberg, R. J. (Ed.), *Handbook of Creativity*, chap. 1, pp. 3–15. Cambridge University Press, Cambridge, UK.

Stowell, D., Robertson, A., Bryan-Kinns, N., & Plumbley, M. D. (2009). Evaluation of live human-computer music-making: quantitative and qualitative approaches. *International Journal of Human-Computer Studies*, *67*, 960–975.

Strapparava, C., Valitutti, A., & Stock, O. (2007). Automatizing two creative functions for advertising. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 99–105 London, UK.

Swartjes, I., & Vromen, J. (2007). Narrative inspiration: Using case based problem solving to support emergent story generation. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 21–28 London, UK.

Tardif, T. Z., & Sternberg, R. J. (1988). What do we know about creativity?. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, chap. 17, pp. 429–440. Cambridge University Press, Cambridge, UK.

Taylor, C. W. (1988). Various approaches to and definitions of creativity. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, chap. 4, pp. 99–121. Cambridge University Press, Cambridge, UK.

Tearse, B., Mawhorter, P., Mateas, M., & Wardrip-Fonin, N. (2011). Experimental results from a rational reconstruction of MINSTREL. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 54–59 Mexico City, Mexico.

Temperley, D. (2001). *The Cognition Of Basic Musical Structures*. MIT Press, Cambridge, MA.

Temperley, D. (2004). An evaluation system for metrical models. *Computer Music Journal*, *28*(3), 28–44.

Temperley, N., & Wollny, P. (2011). Bach revival. Available at Grove Music Online, part of Oxford Music Online (subscription required): *http://www.oxfordmusiconline.com/subscriber/article/grove/music/* 01708, last accessed May 2012.

Thagard, P. (1988). *Computational Philosophy of Science*. MIT Press, Cambridge, MA.

Thom, B. (2000). BoB: an interactive improvisational music companion. In *Proceedings of the Fourth International Conference on Autonomous Agents*, pp. 309–316. ACM.

Thornton, C. (2007). How thinking inside the box can become thinking outside the box. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 113–120 London, UK.

Thornton, C. (2009). Self-redundancy in music. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Torrance, E. P. (1967). Scientific views of creativity and factors affecting its growth. In Kagan, J. (Ed.), *Creativity and Learning*, pp. 73–91. Beacon Press, Boston.

Torrance, E. P. (1974). *Torrance Tests of Creative Thinking*. Scholastic Testing Service, Bensenville, IL.

Torrance, E. P. (1988). The nature of creativity as manifest in its testing. In Sternberg, R. J. (Ed.), *The Nature of Creativity*, chap. 2, pp. 43–75. Cambridge University Press, Cambridge, UK.

Treffry, D. (Ed.). (1998). *Collins English Dictionary* (4th edition). Harper Collins, Glasgow, UK.

Treffry, D. (Ed.). (2000). *The Times English Dictionary*. Harper Collins, Glasgow, UK.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Turner, S. R. (1994). *The creative process: a computer model of storytelling and creativity*. Erlbaum, Hillside, NJ.

Van den Eynden, V., Corti, L., Woollard, M., Bishop, L., & Horton, L. (2011). Managing and sharing data. Tech. rep., UK Data Archive. Revised 3rd edition.

Veale, T. (2006a). An analogy-oriented type hierarchy for linguistic creativity. *Knowledge-Based Systems*, *19*(7), 471 – 479.

Veale, T. (2006b). Re-representation and creative analogy: A lexico-semantic perspective. *New Generation Computing*, *24*(3), 223–240.

Veale, T. (2012). *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. Bloomsbury Academic.

Veale, T., Gervás, P., & Pease, A. (2006). Understanding creativity: A computational perspective. *New Generation Computing*, *24*(3), 203–207.

Veale, T., & Hao, Y. (2008). Slip-sliding along in linguistic creativity: Building a fluid space for connecting disparate ideas. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 101–108 Madrid, Spain.

Venour, C., Ritchie, G., & Mellish, C. (2010). Quantifying humorous lexical incongruity. In *Proceedings of the International Conference on Computational Creativity*, pp. 268–277 Lisbon, Portugal.

Ventura, D. (2008). A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 11–19 Madrid, Spain.

Ventura, D. (2011). No free lunch in the search for creativity. In *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 108–110 Mexico City, Mexico.

von Hentig, H. (1998). *Kreativität: Hohe Erwartungen an einen schwachen Begriff (tr. Creativity: A high expectation of a weak term)*. Beltz Taschenbuch, Munich, Germany.

Waismann, F. (1946). The many-level-structure of language. *Synthese*, *5*(5-6), 221–229.

Waismann, F. (1965). *The principles of linguistic philosophy*. Macmillan, London, UK.

Waismann, F. (1968). Verifiability. In *Essays on Logic and Language*, pp. 117–144. Basil Blackwell, Oxford, UK.

Wallach, M., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity-intelligence distinction*. Holt, Rinehart and Winston, New York, NY.

Wallas, G. (1926). *The Art of Thought* (1st edition). Jonathan Cape, London, UK.

Wallas, G. (1945). *The Art of Thought* (abridged edition). C. A. Watts & Co, London, UK.

Ward, J., Thompson-Lake, D., Ely, R., & Kaminski, F. (2008). Synaesthesia, creativity and art: What is the link?. *British Journal of Psychology*, *99*(1), 127–141.

Warner, J. (2010). The absence of creativity in *Feist* and the computational process. *Journal of the American Society for Information Science and Technology*, *n/a*.

Weeds, J. E. (2003). *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, Informatics, University of Sussex, Brighton, UK.

Weiley, V. (2009). Remixing realities: Distributed studios for collaborative creativity. In *Proceedings of the 7th ACM Creativity and Cognition Conference*, pp. 345–346 Berkeley, California.

Weisberg, R. W. (1988). Problem solving and creativity. In Sternberg, R. J. (Ed.), *The Nature of Creativity*. Cambridge University Press, Cambridge, UK.

Weizenbaum, J. (1976). *Computer Power and Human Reason*. W. H. Freeman and Co., San Fransisco, CA.

Whorley, R., Wiggins, G., Rhodes, C., & Pearce, M. (2010). Development of techniques for the computational modelling of harmony. In *Proceedings of the International Conference on Computational Creativity*, pp. 11–15 Lisbon, Portugal.

Whorley, R. P., Wiggins, G. A., & Pearce, M. T. (2007). Systematic evaluation and improvement of statistical models of harmony. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 81–88 London, UK.

Widmer, G., Flossmann, S., & Grachten, M. (2009). YQX plays Chopin. *AI Magazine*, *30*(3), 35–48.

Wiggins, G., Miranda, E., Smaill, A., & Harris, M. (1993). A framework for the evaluation of music representation systems. *Computer Music Journal*, *17*(3), 31–42.

Wiggins, G. A. (2000). Preface. In *Proceedings of AISB Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science*, p. iii Birmingham, UK.

Wiggins, G. A. (2001). Towards a more precise characterisation of creativity in AI. In *Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science*.

Wiggins, G. A. (2003). Categorising creative systems. In *Proceedings of the 3rd Workshop on Creative Systems: Approaches to Creativity in Artificial Intelligence and Cognitive Science* Acapulco, Mexico. IJCAI.

Wiggins, G. A. (2006a). A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems*, *19*(7), 449–458.

Wiggins, G. A. (2006b). Searching for computational creativity. *New Generation Computing*, *24*(3), 209–222.

Wiggins, G. A. (2008). Closing the loop: Computational creativity from a model of music cognition. COGS research seminar, School of Informatics, University of Sussex (October 2008).

Williams, F. (1967). The mystique of unconscious creation. In Kagan, J. (Ed.), *Creativity and Learning*, pp. 142–152. Beacon Press, Boston.

Williams, R. (1976). *Keywords: a vocabulary of culture and society*. Fontana/Croom Helm, Glasgow, UK.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus, tr. CK Ogden*. Routledge Kegan Paul, London, UK.

Wittgenstein, L. (1958). *Philosophical Investigations, eds. Anscombe, G. E. M. and Rhees, R. and Von Wright, G. H.* (2nd edition). Basil Blackwell, Oxford, UK.

Yeap, W., Opas, T., & Mahyar, N. (2010). On two desiderata for creativity support tools. In *Proceedings of the International Conference on Computational Creativity*, pp. 180–189 Lisbon, Portugal.

Young, M. (2009). Creative computers, improvisation and intimacy. In *Computational Creativity: An Interdisciplinary Approach*, No. 09291 in Dagstuhl Seminar Proceedings Dagstuhl, Germany.

Young, M., & Bown, O. (2010). Clap-along: A negotiation strategy for creative musical interaction with computational systems. In *Proceedings of the International Conference on Computational Creativity*, pp. 215–222 Lisbon, Portugal.

Zhu, J. (2012a). Towards a mixed evaluation approach for computational narrative systems. In *International Conference on Computational Creativity*.

Zhu, J. (2012b). Designing an interdisciplinary user evaluation for the riu computational narrative system. In Oyarzun, D., Peinado, F., Young, R., Elizalde, A., & Mndez, G. (Eds.), *Interactive Storytelling*, Lecture Notes in Computer Science, pp. 126–131. Springer Berlin Heidelberg.

Zhu, J., & Ontañón, S. (2010). Towards analogy-based story generation. In *Proceedings of the International Conference on Computational Creativity*, pp. 75–84 Lisbon, Portugal.

Zhu, X., Xu, Z., & Khot, T. (2009). How creative is your writing? a linguistic creativity measure from computer science and cognitive psychology perspectives. In *Proceedings of NAACL HLT Workshop on Computational Approaches to Linguistic Creativity (ACL)*, pp. 87–93 Boulder, Colorado.

Zongker, D. (2006). Chicken chicken chicken: Chicken chicken. *Annals of Improbable Research*, *12*(5), 16–21.

Appendix A

Papers surveyed in the review of current practice (Chapter 2 Section 2.3)

The following 75 papers document the computational creativity systems included in the survey on current evaluation practice in computational creativity, in Chapter 2 Section 2.3:

Journals

Applied Intelligence 16(2), 2002

1. Machado and Cardoso (2002) All the truth about NEvAr. *Applied Intelligence*, 16(2):101-118.

Leonardo 35(2), 2002

1. Mornoi, Zuben, and Manzolli (2002) ArTbitration: Human-machine interaction in artistic domains. *Leonardo*, 35(2):185-188.

Knowledge-Based Systems 19(7), 2006

1. Gomes, Seco, Pereira, Paiva, Carreiro, Ferreira, and Bento (2006) The importance of retrieval in creative design analogies. *Knowledge-Based Systems*, 19(7):480-488.
2. Pereira and Cardoso (2006) Experiments with free concept generation in Divago. *Knowledge-Based Systems*, 19(7):459-470.
3. Veale (2006a) An analogy-oriented type hierarchy for linguistic creativity. *Knowledge-Based Systems*, 19(7):471-479.

New Generation Computing 24(3), 2006

1. O'Donoghue, Bohan, and Keane (2006) Seeing things: Inventive reasoning with geometric analogies and topographic maps. *New Generation Computing*, 24(3):267-288.
2. Peinado and Gervas (2006) Evaluation of automatic generation of basic stories. *New Generation Computing*, 24(3):289-302.
3. Riedl and Young (2006) Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing*, 24(3):303-323.
4. Veale (2006b) Re-representation and creative analogy: A lexico-semantic perspective. *New Generation Computing*, 24(3):223-240.

A.I. Magazine 30(3), 2009

1. Bushinsky (2009) Deux ex machina - a higher creative species in the game of chess. *AI Magazine*, 30(3), 63-70.
2. Widmer, Flossmann, and Grachten (2009) YQX plays Chopin. *AI Magazine*, 30(3):35-48.

Minds and Machines 20(4), 2010

1. León and Gervás (2010) The role of evaluation-driven rejection in the successful exploration of a conceptual space of stories. *Minds and Machines*, 20(4):615-634.
2. Pérez y Pérez, Aguilar, and Negrete (2010) The ERI-Designer: A computer model for the arrangement of furniture. *Minds and Machines*, 20(4):533-564.
3. Riedl (2010) Story planning: Creativity through exploration, retrieval, and analogical transformation. *Minds and Machines*, 20(4):589-614.

Conferences, Workshops and Seminars

IJWCC'07: International Joint Workshop for Computational Creativity, London, UK, 2007

1. Alvarez Cos, Perez y Perez, and Aliseda (2007) A generative grammar for pre-hispanic production: The case of El Tajin style.
2. Bird and Stokes (2007) Minimal creativity, evaluation and fractal pattern discrimination.
3. Chuan and Chew (2007) A hybrid system for automatic generation of style-specific accompaniment.
4. Gervás, Perez y Perez, Sosa, and Lemaitre (2007) On the fly collaborative story-telling: Revising contributions to match a shared partial story line.
5. Hassan, Gervás, León, and Hervas (2007) A computer model that generates biography-like narratives.
6. Hull and Colton (2007) Towards a general framework for program generation in creative domains.
7. O'Donoghue (2007) Evaluating computer-generated analogies.
8. Oliveira, Cardoso, and Pereira (2007) Tra-la-lyrics: An approach to generate text based on rhythm.
9. Pearce and Wiggins (2007) Evaluating cognitive models of musical composition.
10. Ritchie, Manurung, Pain, Waller, Black, and O'Mara (2007) A practical application of computational humour.
11. McCormack (2007) Creative ecosystems.
12. Strapparava, Valitutti, and Stock (2007) Automatizing two creative functions for advertising.
13. Swartjes and Vromen (2007) Narrative inspiration: Using case based problem solving to support emergent story generation.
14. Whorley, Wiggins, and Pearce (2007) Systematic evaluation and improvement of statistical models of harmony.

IJWCC'08: International Joint Workshop for Computational Creativity, Madrid, Spain, 2008

1. Aguilar, Hernandez, Pérez y Pérez, Rojas, and Zambrano (2008) A computer model for novel arrangements of furniture.
2. Chan and Ventura (2008) Automatic composition of themed mood pieces.
3. Colton (2008c) Experiments in constraint-based automated scene generation.
4. Forth, McLean, and Wiggins (2008) Musical creativity on the conceptual level.
5. Grace, Saunders, and Gero (2008) A computational model for constructing novel associations.
6. León, de Albornoz, and Gervás (2008) A framework for building creative objects from heterogeneous generation systems.
7. León and Gervás (2008) Creative storytelling based on exploration and transformation of constraint rules.
8. Alvarado Lopez and Pérez y Pérez (2008) A computer model for the generation of monophonic musical melodies.
9. Montfort and Pérez y Pérez (2008) Integrating a plot generator and an automatic narrator to create and tell stories.
10. Riedl (2008) Vignette-based story planning: Creativity through exploration and retrieval.
11. Veale and Hao (2008) Slip-sliding along in linguistic creativity: Building a fluid space for connecting disparate ideas.
12. Ventura (2008) A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems.

Computational Creativity: An Interdisciplinary Approach, Dagstuhl, Germany, 2009

1. Brown (2009b) Autonomy, signature and creativity.
2. Cohen (2009a) The art of self-assembly: The self-assembly of art.

3. Edmonds (2009a) Statement.
4. Edmonds (2009b) Words on the Creativity and Cognition Studios contribution.
5. Lewis (2009) Casually evolving creative technology systems.
6. McCormack (2009) Creative ecosystems.
7. Porter (2009) Simulating morphogenesis.
8. Saunders (2009) Artificial creative systems: Completing the creative cycle.
9. Thornton (2009) Self-redundancy in music.
10. Young (2009) Creative computers, improvisation and intimacy.

ICCCX: International Conference on Computational Creativity, Lisbon, Portugal, 2010

1. Bickerman, Bosley, Swire, and Keller (2010) Learning to create jazz melodies using deep belief nets.
2. Chordia and Rae (2010) Tabla Gyan: An artificial tabla improviser.
3. Collins, Laney, Willis, and Garthwaite (2010) Music: Patterns and harmony using discovered, polyphonic patterns to filter computer-generated music.
4. Colton, Gow, Torres, and Cairns (2010) Experiments in objet trouvé browsing.
5. de Melo and Gratch (2010) Evolving expression of emotions through color in virtual humans using genetic algorithms.
6. Eigenfeldt and Pasquier (2010) Realtime generation of harmonic progressions using constrained Markov selection.
7. de Silva Garza and Gero (2010) Elementary social interactions and their effects on creativity: A computational simulation.
8. Grace, Saunders, and Gero (2010) Constructing conceptual spaces for novel associations.
9. Jordanous (2010c) A fitness function for creativity in jazz improvisation and beyond.
10. Krzeczkowska, El-Hage, Colton, and Clark (2010) Automated collage generation - with intent.
11. López, Oliveira, and Cardoso (2010) Real-time emotion-driven music engine.
12. Machado and Nunes (2010) A step towards the evolution of visual languages.
13. McGreggor, Kunda, and Goel (2010) A fractal approach towards visual analogy.
14. Monteith, Martinez, and Ventura (2010) Automatic generation of music for inducing emotive response.
15. Mozetič, Lavrač, Podpečan, Novak, Motaln, Petek, Gruden, Toivonen, and Kulovesi (2010) Bisociative knowledge discovery for microarray data analysis.
16. Nelson, Brummel, Grove, Jorgenson, Sen, and Gamble (2010) Measuring creativity in software development.
17. Norton, Heath, and Ventura (2010) Establishing Appreciation in a Creative System.
18. Pérez y Pérez, Negrete, Penãlosa, Aṽila, Castellanos, and Lemaitre (2010) MEXICA-Impro: A computational model for narrative improvisation.
19. Saunders, Gemeinboeck, Lombard, Bourke, and Kocaballi (2010) Curious Whispers: An embodied artificial creative system.
20. Sorenson and Pasquier (2010) The evolution of fun: Automatic level design through challenge modeling.
21. Venour, Ritchie, and Mellish (2010) Quantifying humorous lexical incongruity.
22. Whorley, Wiggins, Rhodes, and Pearce (2010) Development of techniques for the computational modelling of harmony.
23. Yeap, Opas, and Mahyar (2010) On two desiderata for creativity support tools.
24. Young and Bown (2010) Clap-along: A negotiation strategy for creative musical interaction with computational systems.
25. Zhu and Ontañón (2010) Towards analogy-based story generation.

Appendix B

Papers used to derive a definition of creativity (Chapter 4)

Creativity Corpus

The following 30 papers were used as the *creativity corpus* for the computational linguistics work in Chapter 4 deriving a definition of creativity:

1. T. M. Amabile. The social psychology of creativity: A componential conceptualization. Journal of Personality and Social Psychology, 45(2):357-376, 1983.
2. M. A. Boden. Precis of The Creative Mind: Myths and mechanisms. Behavioural and Brain Sciences, 17(3):519-570, 1994.
3. D. T. Campbell. Blind variation and selective retentions in creative thought as in other knowledge processes. Psychological Review, 67(7):380-400, 1960.
4. S. Colton, A. Pease, and G. Ritchie. The effect of input knowledge on creativity. In Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science, 2001.
5. M. Csikszentmihalyi. Motivation and creativity: Toward a synthesis of structural and energistic approaches to cognition. New Ideas in Psychology, 6(2):159-176, 1988.
6. M. Dellas and E. L. Gaier. Identification of creativity: The individual. Psychological Bulletin, 73(1):55- 73, 1970.
7. A. Dietrich. The cognitive neuroscience of creativity. Psychonomic Bulletin & Review, 11(6):1011-1026, 2004.
8. G. Domino. Identification of potentially creative persons from the adjective check list. Journal of Consulting and Clinical Psychology, 35(1):48-51, 1970.
9. W. Duch. Intuition, insight, imagination and creativity. IEEE Computational Intelligence Magazine, 2(3):40-52, 2007.
10. C. S. Findlay and C. J. Lumsden. The creative mind: Toward an evolutionary theory of discovery and innovation. Journal of Social and Biological Systems, 11(1):3-55, 1988.
11. C. M. Ford. A theory of individual creative action in multiple social domains. The Academy of Management Review, 21(4):1112-1142, 1996.
12. J. Gero. Creativity, emergence and evolution in design. Knowledge-Based Systems, 9(7):435-448, 1996.
13. H. G. Gough. A creative personality scale for the adjective checklist. Journal of Personality and Social Psychology, 37(8):1398-1405, 1979.
14. J. P. Guilford. Creativity. American Psychologist, 5:444-454, 1950.
15. Z. Ivcevic. Creativity map: Toward the next generation of theories of creativity. Psychology of Aesthetics, Creativity, and the Arts, 3(1):17-21, 2009.
16. K. H. Kim. Can we trust creativity tests? A review of the Torrance tests of creative thinking (TTCT). Creativity Research Journal, 18(1):3-14, 2006.
17. L. A. King, L. McKee Walker, and S. J. Broyles. Creativity and the five-factor model. Journal of Research in Personality, 30(2):189-203, 1996.
18. R. R. McCrae. Creativity, divergent thinking, and openness to experience. Journal of Personality and Social Psychology, 52(6):1258-1265, 1987.

19. S. A. Mednick. The associative basis of the creative process. Psychological Review, 69(3):220-232, 1962.
20. M. D. Mumford and S. B. Gustafson. Creativity syndrome: Integration, application, and innovation. Psychological Bulletin, 103(1):27-43, 1988.
21. M. T. Pearce, D. Meredith, and G. A. Wiggins. Motivations and methodologies for automation of the compositional process. Musicae Scientae, 6(2):119-147, 2002.
22. J. A. Plucker, R. A. Beghetto, and G. T. Dow. Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. Educational Psychologist, 39(2):83-96, 2004.
23. R. Richards, D. K. Kinney, M. Benet, and A. P. C. Merzel. Assessing everyday creativity: Characteristics of the lifetime creativity scales and validation with three large samples. Journal of Personality and Social Psychology, 54(3):476-485, 1988.
24. G. Ritchie. The transformational creativity hypothesis. New Generation Computing, 24(3):241-266, 2006.
25. G. Ritchie. Some empirical criteria for attributing creativity to a computer program. Minds and Machines, 17:67-99, 2007.
26. D. L. Rubenson and M. A. Runco. The psychoeconomic approach to creativity. New Ideas in Psychology, 10(2):131-147, 1992.
27. M. A. Runco and I. Chand. Cognition and creativity. Educational Psychology Review, 7(3):243-267, 1995.
28. D. K. Simonton. Creativity: Cognitive, personal, developmental, and social aspects. American Psychologist, 55(1):151-158, 2000.
29. J. R. Suler. Primary process thinking and creativity. Psychological Bulletin, 88(1):144-165, 1980.
30. G. A. Wiggins. A preliminary framework for description, analysis and comparison of creative systems. Knowledge-Based Systems, 19(7):449-458, 2006.

Non-Creativity Corpus

The following 60 papers were used as the *non-creativity corpus* for this work:

1. C. Ames and J. Archer. Achievement goals in the classroom: Students' learning strategies and motivation processes. Journal of Educational Psychology, 80(3):260-267, 1988.
2. J. Anderson and D. Gerbing. Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103(3):411-423, 1988.
3. J. Arnett. Emerging adulthood: A theory of development from the late teens through the twenties. American Psychologist, 55(5):469-480, 2000.
4. M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing, 50(2):174-188, 2002.
5. A. Baddeley. Exploring the central executive. Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 49(1):5-28, 1996.
6. T. Baker, M. Piper, D. McCarthy, M. Majeskie, and M. Fiore. Addiction motivation reformulated: An affective processing model of negative reinforcement. Psychological Review, 111(1):33-51, 2004.
7. P. Barnett and I. Gotlib. Psychosocial functioning and depression: Distinguishing among antecedents, concomitants, and consequences. Psychological Bulletin, 104(1):97-126, 1988.
8. J. Baron. Nonconsequentialist decisions. Behavioral and Brain Sciences, 17(1):1-42, 1994.
9. F. Beach. The snark was a boojum. American Psychologist, 5(4):115-124, 1950.
10. M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7:2399-2434, 2006.
11. G. Bonanno. Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely aversive events? American Psychologist, 59(1):20-28, 2004.
12. T. Chan and L. Vese. Active contours without edges. IEEE Transactions on Image Processing, 10(2):266-277, 2001.
13. H. Cheng and J. Schweitzer. Cultural values reflected in Chinese and U.S. television commercials. Journal of Advertising Research, 36(3):27-45, 1996.
14. C. Coello Coello. Evolutionary multi-objective optimization: A historical view of the field. IEEE Computational Intelligence Magazine, 1(1):28-36, 2006.
15. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):603-619, 2002.
16. L. Cronbach and L. Furby. How we should measure 'change': Or should we? Psychological Bulletin, 74(1):68-80, 1970.

17. M. Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. Journal of Personality and Social Psychology, 44(1):113-126, 1983.
18. J. Demšar. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7:1-30, 2006.
19. M. Dorigo, M. Birattari, and T. Stützle. Ant colony optimization artificial ants as a computational intelligence technique. IEEE Computational Intelligence Magazine, 1(4):28-39, 2006.
20. E. Fischer and J. Turner. Orientations to seeking professional help: Development and research utility of an attitude scale. Journal of Consulting and Clinical Psychology, 35(1 PART 1):79-90, 1970.
21. J. Gibson. Observations on active touch. Psychological Review, 69(6):477-491, 1962.
22. A. Gopnik and J. Astington. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. Child development, 59(1):26-37, 1988.
23. S. Gosling, S. Vazire, S. Srivastava, and O. John. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. American Psychologist, 59(2):93-104, 2004.
24. J. Gray. The psychophysiological basis of introversion-extraversion. Behaviour Research and Therapy, 8(3):249-266, 1970.
25. B. Grosz and S. Kraus. Collaborative plans for complex group action. Artificial Intelligence, 86(2):269- 357, 1996.
26. P. Groves and R. Thompson. Habituation: A dual-process theory. Psychological Review, 77(5):419-450, 1970.
27. M. Hall, J. Anderson, S. Amarasinghe, B. Murphy, S.-W. Liao, E. Bugnion, and M. Lam. Maximizing multiprocessor performance with the SUIF compiler. Computer, 29(12):84-89, 1996.
28. F. Happé. The role of age and verbal ability in the theory of mind task performance of subjects with autism. Child development, 66(3):843-855, 1995.
29. C. Harland. Supply chain management: Relationships, chains and networks. British Journal of Management, 7(SPEC. ISS.):S63-S80, 1996.
30. S. Hayes, K. Strosahl, K. Wilson, R. Bissett, J. Pistorello, D. Toarmino, M. Polusny, T. Dykstra, S. Batten, J. Bergan, S. Stewart, M. Zvolensky, G. Eifert, F. Bond, J. Forsyth, M. Karekla, and S. McCurry. Measuring experiential avoidance: A preliminary test of a working model. Psychological Record, 54(4):553-578, 2004.
31. J. Hirsch and K. Lücke. Overview no. 76. mechanism of deformation and development of rolling textures in polycrystalline f.c.c. metals-i. description of rolling texture development in homogeneous cuzn alloys. Acta Metallurgica, 36(11):2863-2882, 1988.
32. P. Killeen. Mathematical principles of reinforcement. Behavioral and Brain Sciences, 17(1):105-172, 1994.
33. P. Kirschner, J. Sweller, and R. Clark. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41(2):75-86, 2006.
34. S. Liao. Notes on the homotopy analysis method: Some definitions and theorems. Communications in Nonlinear Science and Numerical Simulation, 14(4):983-997, 2009.
35. C. Lord, L. Ross, and M. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37(11):2098- 2109, 1979.
36. S. Luthar, D. Cicchetti, and B. Becker. The construct of resilience: A critical evaluation and guidelines for future work. Child Development, 71(3):543-562, 2000.
37. G. Mandler. Recognizing: The judgment of previous occurrence. Psychological Review, 87(3):252-271, 1980.
38. G. Miller and J. Selfridge. Verbal context and the recall of meaningful material. The American journal of psychology, 63(2):176-185, 1950.
39. S. Miller. Monitoring and blunting: Validation of a questionnaire to assess styles of information seeking under threat. Journal of Personality and Social Psychology, 52(2):345-353, 1987.
40. G. Navarro and V. Mäkinen. Compressed full-text indexes. ACM Computing Surveys, 39(1), 2007.
41. M. Nissen and P. Bullemer. Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19(1):1-32, 1987.
42. J. Payne, J. Bettman, and E. Johnson. Adaptive strategy selection in decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3):534-552, 1988.
43. J. Prochaska, C. DiClemente, and J. Norcross. In search of how people change: Applications to addictive behaviors. American Psychologist, 47(9):1102-1114, 1992.
44. T. Richardson, M. Shokrollahi, and R. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. IEEE Transactions on Information Theory, 47(2):619-637, 2001.
45. W. Rozeboom. The fallacy of the null-hypothesis significance test. Psychological Bulletin, 57(5):416-428, 1960.
46. C. Rusbult. A longitudinal test of the investment model: The development (and deterioration) of satisfaction and commitment in heterosexual involvements. Journal of Personality and Social Psychology, 45(1):101-117, 1983.

47. T. Ryan. Significance tests for multiple comparison of proportion, variance, and other statistics. Psychological Bulletin, 57(4):318-328, 1960.
48. W. Schultz. Behavioral theories and the neurophysiology of reward. Annual Review of Psychology, 57:87-115, 2006.
49. E. Sirin, B. Parsia, B. Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical OWL-DL reasoner. Web Semantics, 5(2):51-53, 2007.
50. T. Srull and R. Wyer. The role of category accessibility in the interpretation of information about persons: Some determinants and implications. Journal of Personality and Social Psychology, 37(10):1660- 1672, 1979.
51. J. Steiger. Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87(2):245-251, 1980.
52. L. Steinberg, S. Lamborn, S. Dornbusch, and N. Darling. Impact of parenting practices on adolescent achievement: authoritative parenting, school involvement, and encouragement to succeed. Child development, 63(5):1266-1281, 1992.
53. D. Tao, X. Li, X. Wu, W. Hu, and S. Maybank. Supervised tensor learning. Knowledge and Information Systems, 13(1):1-42, 2007.
54. A. Tellegen, D. Lykken, T. Bouchard Jr., K. Wilcox, N. Segal, and S. Rich. Personality similarity in twins reared apart and together. Journal of Personality and Social Psychology, 54(6):1031-1039, 1988.
55. L. Thomas and D. Ganster. Impact of family-supportive work variables on work-family conflict and strain: A control perspective. Journal of Applied Psychology, 80(1):6-15, 1995.
56. I. Thompson. Coupled reaction channels calculations in nuclear physics. Computer Physics Reports, 7(4):167-212, 1988.
57. E. Tulving. Subjective organization in free recall of "unrelated" words. Psychological Review, 69(4):344- 354, 1962.
58. U. Von Luxburg. A tutorial on spectral clustering. Statistics and Computing, 17(4):395-416, 2007.
59. J. Williams, A. Mathews, and C. MacLeod. The emotional stroop task and psychopathology. Psychological Bulletin, 122(1):3-24, 1996.
60. J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):210-227, 2009.

Appendix C

Creativity words identified in Chapter 4

The following words were identified in Chapter 4 as *creativity words*, words which were found to occur significantly more often than expected when discussing the nature of creativity. There are 389 nouns (indicated with the suffix _N), 72 verbs (suffixed _V), 205 adjectives (suffixed _J) and 28 adverbs (suffixed _R). Words are listed in descending order of LLR score (see Chapter 4 Section 4.2), with a word's LLR score given in brackets after the word.

- thinking_N (834.55)
- process_N (612.05)
- innovation_N (546.2)
- artefact_N (514.33)
- idea_N (475.74)
- program_N (474.41)
- domain_N (436.58)
- cognitive_J (393.79)
- divergent_J (357.43)
- accomplishment_N (355.35)
- openness_N (328.57)
- discovery_N (327.38)
- primary_J (326.65)
- originality_N (315.6)
- criterion_N (312.61)
- intelligence_N (309.31)
- ability_N (299.27)
- knowledge_N (290.48)
- individual_N (243.34)
- human_J (234.41)
- novelty_N (232.72)
- conceptual_J (232.58)
- art_N (232.52)
- new_J (227.61)
- production_N (216.24)
- composition_N (206.58)
- musical_J (206.18)
- artistic_J (205.1)
- thought_N (202.08)
- activity_N (197.17)
- concept_N (189.9)
- artist_N (188.4)
- personality_N (175.19)
- transformational_J (174.1)
- skill_N (167.98)
- contribution_N (162.4)
- talent_N (162.17)
- motivation_N (159.51)
- scientific_J (157.51)
- genre_N (152.63)
- intellectual_J (149.37)
- typicality_N (145.48)
- prefrontal_J (140.77)
- insight_N (139.65)
- vocational_J (138.32)
- field_N (137.17)
- potential_N (136.14)
- sociocultural_J (135.94)
- rating_N (134.79)
- formal_J (133.73)
- computational_J (133.6)
- composer_N (131.17)
- psychic_J (131.17)
- associative_J (121.53)
- brain_N (118.04)
- novel_J (117.68)
- fluency_N (117.42)
- inspire_V (116.06)
- facilitate_V (116.04)
- generate_V (115.89)
- chapter_N (109.8)
- conscientiousness_N (109.72)
- gene-culture_N (109.7)
- novel_N (108.39)
- quality_N (106.85)
- flexibility_N (106.34)
- scientist_N (101.92)
- produce_V (101.73)
- unconscious_J (100.36)
- psychology_N (99.91)
- science_N (99.65)
- understanding_N (99.49)
- poem_N (99.13)
- remote_J (98.09)
- painting_N (97.78)
- productivity_N (96.09)
- element_N (94.42)
- endeavor_N (93.82)
- minor_J (93.23)
- primitive_J (91.56)
- innovative_J (91.39)
- output_N (91.1)
- music_N (90.79)
- structure_N (90.77)
- gift_V (90.62)
- market_N (89.94)
- product_N (89.44)
- faculty_N (89.05)
- perhaps_R (83.61)
- barren_N (83.47)
- transformation_N (81.62)
- artefact_J (81.09)
- ideation_N (81.09)
- melody_N (81.09)
- phenotype_N (81.09)
- capacity_N (79.1)
- aesthetic_J (78.7)
- avocational_J (78.7)
- association_N (78.08)
- semantic_J (76.73)
- circuit_N (75.88)
- emergence_N (75.88)
- organisational_J (74.12)
- epigenetic_J (73.93)
- characteristic_N (72.64)
- achievement_N (72.3)
- analogy_N (72.25)
- ego_N (71.93)
- agreeableness_N (71.55)
- am_R (71.55)
- compositional_J (71.55)
- domain-relevant_J (71.55)
- framework_N (69.95)
- consciousness_N (69.76)
- combination_N (69.76)
- interest_N (69.62)
- influence_N (68.39)
- evolutionary_J (68.14)
- imagination_N (65.75)
- environment_N (65.56)
- secondary_J (65.5)
- extrinsic_J (64.46)
- danish_J (64.39)
- invention_N (62.43)
- ideational_J (62.01)
- perceptual_J (61.59)
- appropriateness_N (61.19)
- unusual_J (60.9)
- deliberate_J (60.29)
- ai_N (59.95)
- synthesis_N (59.62)
- transmission_N (59.14)
- notion_N (58.9)
- mathematician_N (58.8)
- abstract_J (58.36)
- imagery_N (58.01)
- productive_J (57.83)
- hierarchy_N (57.33)
- heterarchy_N (57.24)
- listener_N (57.24)
- assessment_N (56.27)
- membership_N (55.42)
- inspiration_N (54.85)
- myth_N (54.83)

- mutation_N (54.18)
- organic_J (52.47)
- iq_N (51.9)
- rater_N (51.87)
- perspective_N (51.4)
- logical_J (51.26)
- validity_N (51.2)
- manifest_V (50.29)
- possess_V (50.29)
- genius_N (50.08)
- empirical_J (49.68)
- emergent_J (49.04)
- spontaneous_J (48.67)
- rate_V (48.35)
- developmental_J (48.08)
- welsh_J (47.7)
- deem_V (47.68)
- interest_V (47.04)
- influence_V (46.96)
- poetry_N (46.81)
- quantity_N (46.78)
- intrinsic_J (46.71)
- career_N (46.67)
- conceptualisation_N (46.67)
- variation_N (46.65)
- value_V (46.6)
- drive_N (45.97)
- repertoire_N (45.97)
- blind_J (45.47)
- habitual_J (45.4)
- highly_R (45.32)
- architect_N (45.31)
- componential_J (45.31)
- fine-tuned_J (45.31)
- cortex_N (45.16)
- psychoanalytic_J (44.89)
- adjective_N (44.52)
- peer_N (44.52)
- schema_N (43.67)
- lack_V (43.54)
- genetic_J (43.27)
- artificial_J (43.03)
- locomotion_N (42.93)
- pine_N (42.93)
- heuristic_N (42.82)
- keyword_N (42.67)
- provincial_J (42.67)
- judge_N (42.07)
- receptivity_N (40.54)
- contribute_V (40.16)
- generative_J (40.15)
- human_N (39.94)
- implicit_J (39)
- occupational_J (38.56)
- rational_J (38.42)
- possibility_N (38.34)
- biological_J (38.26)
- incubation_N (38.16)
- reorganisation_N (38.16)
- marginal_J (37.16)
- compose_V (36.69)
- story_N (36.55)
- cognition_N (36.3)
- external_J (36.25)
- retention_N (36.2)
- clarify_V (35.92)
- hemisphere_N (35.77)
- high-valued_J (35.77)
- imaginative_J (35.77)
- origence_N (35.77)
- space-definition_N (35.77)
- environmental_J (35.65)
- recognise_V (35.62)
- explicit_J (35.55)
- evaluation_N (34.89)
- observable_J (34.88)
- culture_N (34.67)
- discover_V (34.51)
- conscious_J (34.51)
- ambiguity_N (34.42)
- society_N (34.23)
- enable_V (34.16)
- writer_N (34.07)
- joke_N (33.55)
- routine_J (33.54)
- configuration_N (33.5)
- consequences_N (33.39)
- examinee_N (33.39)
- intellectence_N (33.39)
- neo-pi_J (33.39)
- psychoeconomic_J (33.39)
- subnetwork_N (33.39)
- uninspiration_N (33.39)
- content_J (33.33)
- economic_J (33.26)
- protocol_N (33.04)
- benefit_N (32.69)
- selective_J (32.33)
- valuable_J (31.99)
- claim_N (31.8)
- associate_N (31.64)
- atom_N (31.55)
- scoring_N (31.55)
- appreciation_N (31.37)
- medium_J (31.37)
- allele_N (31)
- divergent-thinking_J (31)
- energistic_J (31)
- interplay_N (31)
- tat_N (31)
- thinker_N (31)
- uncreative_J (31)
- workings_N (31)
- language_N (30.99)
- suitable_J (30.82)
- psychologist_N (30.57)
- link_N (30.37)
- aptitude_N (29.97)
- societal_J (29.96)
- educational_J (29.94)
- teacher_N (29.94)
- generation_N (29.6)
- gestalt_N (29.2)
- literary_J (29.2)
- prototype_N (29.2)
- stochastic_J (29.01)
- certainly_R (28.96)
- collage_N (28.62)
- fine-tuning_N (28.62)
- innovator_N (28.62)
- molecule_N (28.62)
- node-link_N (28.62)
- essential_J (28.5)
- extraversion_N (28.46)
- usefulness_N (28.25)
- expert_J (28.24)
- score_V (28.07)
- enhance_V (27.91)
- and/or_N (27.68)
- direct_V (27.5)
- linguistic_J (27.05)
- prerequisite_N (27.05)
- functional_J (26.98)
- operational_J (26.83)
- absorptive_J (26.76)
- fuzzy_J (26.76)
- genetics_N (26.76)
- surprise_N (26.29)
- aberration_N (26.23)
- brainstorming_N (26.23)
- buffer_N (26.23)
- commonplace_J (26.23)
- h-creativity_N (26.23)
- historian_N (26.23)
- innovativeness_N (26.23)
- interrater_N (26.23)
- intrapersonal_J (26.23)
- noncomputational_J (26.23)
- refrigerator_N (26.23)
- stakeholder_N (26.23)
- synonym_N (26.23)
- intuition_N (26.19)
- institutional_J (25.8)
- wide_J (25.46)
- abstraction_N (25.24)
- merely_R (25.22)
- conformity_N (24.96)
- lifetime_N (24.93)
- illogical_J (24.9)
- dissociate_V (24.52)
- interviewer_N (24.52)
- neuroscience_N (24.52)
- preference_N (24.34)
- capable_J (24.32)
- meaning_N (23.95)
- associational_J (23.85)
- basal_J (23.85)
- disciplinary_N (23.85)
- fuster_N (23.85)
- genotype_N (23.85)
- h-creative_J (23.85)
- informally_R (23.85)
- inheritance_N (23.85)
- lifespan_N (23.85)
- morpheme_N (23.85)
- multi-dimensional_J (23.85)
- musician_N (23.85)
- neurocognitive_J (23.85)
- nominate_V (23.85)
- nomination_N (23.85)
- p-creativity_N (23.85)
- tonal_J (23.85)
- untypical_J (23.85)
- harmony_N (23.75)
- solver_N (23.75)
- subsystem_N (23.75)
- hierarchical_J (23.35)
- logically_R (23.28)
- informal_J (23.26)
- rely_V (23.11)
- chess_N (22.99)
- testable_J (22.99)
- male_J (22.97)
- eminent_J (22.78)
- generator_N (22.78)
- mysterious_J (22.78)
- transform_V (22.59)
- judge_V (22.54)
- gene_N (22.46)
- structure_V (22.45)
- curiosity_N (22.29)
- domain-specific_N (22.29)
- manifestation_N (22.16)
- graduate_N (22.01)
- logic_N (21.92)
- cite_V (21.76)
- loose_J (21.72)
- triangle_N (21.71)
- biocultural_J (21.46)
- coevolution_N (21.46)
- coevolutionary_J (21.46)
- discoverer_N (21.46)
- exceptional_J (21.46)
- firstly_R (21.46)
- five-factor_N (21.46)
- fragmentation_N (21.46)
- heterarchical_J (21.46)
- hunch_N (21.46)
- mentor_N (21.46)
- metalevel_N (21.46)
- o-node_N (21.46)
- sensemaking_N (21.46)
- superspace_N (21.46)
- organise_V (21.45)
- obvious_J (21.19)
- proposal_N (20.86)
- abstract_N (20.67)
- internalise_V (20.67)
- biology_N (20.58)
- political_J (20.34)
- acknowledge_V (20.28)
- battery_N (20.28)
- game_N (20.28)
- neglect_V (20.28)
- foundation_N (20.2)
- corpus_N (20.07)
- grader_N (20.07)
- universality_N (20.07)
- tentative_J (19.95)
- disposition_N (19.94)
- metaphor_N (19.71)
- everyday_J (19.62)
- detrimental_J (19.11)
- anthropologist_N (19.08)
- ativity_N (19.08)
- avocational_N (19.08)
- conspecific_N (19.08)
- deterministic_J (19.08)
- factual_J (19.08)
- fascination_N (19.08)
- feminine_J (19.08)
- hoverfly_R (19.08)
- hypnosis_N (19.08)
- hypnotic_J (19.08)
- image_V (19.08)
- imposition_N (19.08)
- infant_N (19.08)
- innately_R (19.08)
- interpreter_N (19.08)
- melodic_J (19.08)
- metaphorical_J (19.08)
- mutant_N (19.08)
- neo_N (19.08)
- p-creative_J (19.08)

- portrait_N (19.08)
- psi_N (19.08)
- sensibility_N (19.08)
- stylistic_J (19.08)
- well-formed_J (19.08)
- positively_R (19)
- guideline_N (18.72)
- pitch_N (18.72)
- peak_N (18.6)
- grammar_N (18.58)
- history_N (18.53)
- break_V (18.53)
- audience_N (18.48)
- stereotype_N (18.48)
- reality_N (18.46)
- potentially_R (18.46)
- conform_V (18.41)
- expert_N (18.41)
- mathematical_J (18.39)
- designer_N (18.34)
- pertain_V (18.33)
- probably_R (18.12)
- historical_J (17.91)
- conducive_J (17.87)
- dream_V (17.87)
- insightful_J (17.87)
- narrative_N (17.87)
- synthesise_V (17.87)
- apparent_J (17.81)
- factorial_J (17.73)
- title_N (17.73)
- invest_V (17.55)
- apparently_R (17.48)
- dream_N (17.21)
- realistic_J (17.16)
- problem-solving_N (17.1)
- educator_N (17)
- inherent_J (17)
- occupation_N (16.74)
- survival_N (16.72)
- benzene_N (16.69)
- c.f._N (16.69)
- chapters_N (16.69)
- chemist_N (16.69)
- circularity_N (16.69)
- eminence_N (16.69)
- enquiry_N (16.69)
- fertile_J (16.69)
- focussing_N (16.69)
- historic_J (16.69)
- impossibilist_J (16.69)
- intellect_N (16.69)
- interestingness_N (16.69)
- marvelous_J (16.69)
- nominee_N (16.69)
- non-involved_J (16.69)
- officer_N (16.69)
- patent_N (16.69)
- poetic_J (16.69)
- reality-oriented_J (16.69)
- reentry_N (16.69)
- results/results_N (16.69)
- serendipity_N (16.69)
- social-psychological_J (16.69)
- tacitly_R (16.69)
- tier_N (16.69)
- tremendous_J (16.69)
- universal_N (16.69)
- warmup_N (16.69)
- atypical_J (16.63)
- supply_V (16.63)
- phonological_J (16.62)
- play_N (16.62)
- progress_N (16.61)
- open_J (16.53)
- enhancement_N (16.52)
- king_N (16.52)
- radical_J (16.52)
- real-life_N (16.52)
- law_N (16.24)
- heuristic_J (16.14)
- actor_N (16.12)
- ordinary_J (15.9)
- exemplar_N (15.78)
- perseverance_N (15.78)
- blind_N (15.69)
- criteria_N (15.69)
- programmer_N (15.69)
- relativity_N (15.69)
- sudden_J (15.69)
- syntax_N (15.69)
- construction_N (15.47)
- ball_N (15.35)
- conjecture_N (15.35)
- unconventional_J (15.35)
- universe_N (15.35)
- impose_V (15.3)
- constrain_V (15.14)
- articulate_V (15.1)
- demand_V (14.97)
- deny_V (14.48)
- innate_J (14.48)
- revision_N (14.48)
- temporarily_R (14.48)
- requisite_J (14.4)
- archival_J (14.31)
- artefact-set_N (14.31)
- blindly_R (14.31)
- blind-variation-and-selective-retention_N (14.31)
- canalisation_N (14.31)
- combinational_J (14.31)
- concrete_N (14.31)
- cough_N (14.31)
- cross-cultural_J (14.31)
- daydream_N (14.31)
- deduction_N (14.31)
- drive-related_J (14.31)
- edition_N (14.31)
- flexibly_R (14.31)
- grade_V (14.31)
- historiometric_J (14.31)
- home-key_N (14.31)
- imitate_V (14.31)
- inflexible_J (14.31)
- ingenuity_N (14.31)
- intrapopulation_N (14.31)
- jape_N (14.31)
- mach_N (14.31)
- mechanistic_J (14.31)
- morphological_J (14.31)
- psychoticism_N (14.31)
- r.s._N (14.31)
- reputation_N (14.31)
- script_N (14.31)
- sims_N (14.31)
- subjectivity_N (14.31)
- submarket_N (14.31)
- symphony_N (14.31)
- talented_J (14.31)
- tests_N (14.31)
- trial-and-error_N (14.31)
- verdict_N (14.31)
- consumer_N (14.1)
- constantly_R (13.86)
- algorithmic_J (13.84)
- claim_V (13.84)
- overt_J (13.66)
- biochemical_J (13.52)
- camp_N (13.52)
- funny_J (13.52)
- inventive_J (13.52)
- landscape_N (13.52)
- meta-level_N (13.52)
- nurture_V (13.52)
- phenotypic_J (13.52)
- redefinition_N (13.52)
- roadblock_N (13.52)
- senior_N (13.52)
- substantiate_V (13.52)
- transcend_V (13.52)
- thesis_N (13.52)
- aim_V (13.51)
- climate_N (13.51)
- conception_N (13.49)
- criticise_V (13.39)
- mathematics_N (13.36)
- purely_R (13.36)
- fundamentally_R (13.35)
- whereby_R (13.35)
- writing_N (13.35)
- entity_N (13.31)
- undertake_V (13.31)
- field_V (13.29)
- master_V (13.29)
- preconscious_J (13.29)
- old_J (13.21)
- exploratory_J (13.09)
- topic_N (13.01)
- devise_V (12.85)
- largely_R (12.76)
- conceive_V (12.61)
- pose_V (12.61)
- integrative_J (12.6)
- engine_N (12.48)
- masculine_J (12.48)
- debate_N (12.39)
- leisure_N (12.39)
- linkage_N (12.18)
- independence_N (12.09)
- appraise_V (11.97)
- closure_N (11.97)
- deliberately_R (11.97)
- drawing_N (11.97)
- self-confidence_N (11.97)
- abstractly_R (11.92)
- achiever_N (11.92)
- acrobat_N (11.92)
- aesthetics_N (11.92)
- ai-model_N (11.92)
- allude_V (11.92)
- and-selective-retention_N (11.92)
- artistic_N (11.92)
- associative_N (11.92)
- big_N (11.92)
- boredom_N (11.92)
- canalise_V (11.92)
- chorale_N (11.92)
- chord_N (11.92)
- coevolutionary_N (11.92)
- conformist_J (11.92)
- consensual_J (11.92)
- consequent_N (11.92)
- copycat_N (11.92)
- curious_J (11.92)
- curvilinear_J (11.92)
- defocused_J (11.92)
- divergent_N (11.92)
- drosophila_N (11.92)
- falsify_V (11.92)
- fixedness_N (11.92)
- freshman_N (11.92)
- hemispheric_N (11.92)
- hood_N (11.92)
- hypothesised_J (11.92)
- ideational_N (11.92)
- intrapsychic_J (11.92)
- inventor_N (11.92)
- judgemental_J (11.92)
- kindergarten_V (11.92)
- knowledge-based_J (11.92)
- macroscopic_J (11.92)
- neuroscientific_J (11.92)
- one-armed_J (11.92)
- painter_N (11.92)
- patent_V (11.92)
- planetary_J (11.92)
- poet_N (11.92)
- problem-finding_N (11.92)
- punctuation_N (11.92)
- re-invent_V (11.92)
- selectional_J (11.92)
- serendipitous_J (11.92)
- shortcut_N (11.92)
- sonnet_N (11.92)
- substitutive_J (11.92)
- tire_N (11.92)
- unregulated_J (11.92)
- valuation_N (11.92)
- viability_N (11.92)
- wild_N (11.92)
- map_N (11.75)
- advance_V (11.75)
- assemble_V (11.73)
- loosely_R (11.73)
- invent_V (11.72)
- revise_V (11.72)
- elementary_J (11.49)
- happen_V (11.42)
- aberrant_J (11.39)
- aspiration_N (11.39)
- broad-based_J (11.39)
- cellular_J (11.39)
- chase_N (11.39)
- clue_N (11.39)
- dynamical_J (11.39)
- gas_N (11.39)
- intellectually_R (11.39)
- nobel_N (11.39)
- obvious_N (11.39)
- propensity_N (11.39)
- richness_N (11.39)
- sociological_J (11.39)
- synonymous_J (11.39)
- elaboration_N (11.23)
- flexible_J (11.07)
- empirically_R (10.9)

Appendix D

Statements illustrating the Chapter 4 components in a musical improvisation context

---

Using the annotated participant data from the questionnaires about musical improvisation (Chapter 6 Section 6.3.2), statements were extracted to illustrate how each component is relevant to improvisation. These statements were used as test statements for each component in Case Study 1 (Chapter 6, to help the judges analyse the three musical improvisation systems using the components by contextualising each component within musical improvisation. The statements are deliberately repetitive to some degree, illustrating different ways in which related points were made by different participants.

Social Communication and Interaction

Can the system give live performances?
Has the system done gigs/performances to an audience?
Does the system attract people who want it to give gigs?
Can the system transmit material to others (live performance, distribution of media/products, etc)?
Can the system be influenced by others? and choose what contributions influence it?
Does the system challenge prevailing norms?
How is the system perceived by an audience? (positive and/or negative perceptions) And its peers? Its critics?
Can the system communicate, exchange information and interact on a live basis? If so, how?
Does the system synchronise with what is going on around it?
Can other musicians play with the system? Does the system influence the creativity of the other musicians?
How does the system receive and respond to feedback?
How do context and environmental factors affect what the system does?
What is the system aimed at communicating?
What is the system aware of, outside of its own system boundary?
Does the system listen to what is going on around it?
How does the system deal with mistakes or unexpected events from other performers or in the environment?
How does the system help other musicians around it? Does it hinder them at all?
How is the system helped by other musicians around it? Can it be hindered by them at all?
How does the system function as one musician forming part of a whole performance?
What other information does the system display for others, besides the sound (e.g. visuals)?
Do the audience ‘connect’ with the system? and vice versa?

Domain Competence

What domain knowledge does the system have?
How does it use its domain knowledge?
Can the system build upon simpler domain knowledge e.g. single facts, to form more complex interpretations?
What ‘tried and trusted’ patterns and ‘licks’ does the system use?
Does the system practice, train or learn new knowledge?
Does the system have sufficient domain knowledge to use?
Does the system have technical ability and mechanical domain skills such as scales, chord inversions?
Is the system aware of any repertoire?

What musical ideas does the system have?
How does the system express the information it has?
How does the system make use of the contextual information it has?
What domain skills and abilities does the system have?
How reliant is the system on domain heuristics and existing knowledge?
How competent a performer is the system in real-time?
How does the improvisation relate to a given theme of improvisation?
Does the system use phrasing and interpretation appropriately?
How aware is the system of previous history of performances/recordings?
What stylistic constraints are placed on the system? Are the constraints beneficial or a hindrance?
Does the system make use of style, scales, structure, tempo, mood, keys, notes, rhythm, shapes, lines, silence?
What genres and styles can the system use?
In what ways is the system conditioned to a particular style or approach?
Are the system's operations prescribed or described?
Does the system have its own personal style?
Does the system rely on musical clichés?
Can the system recover from mistakes or unintentional events?

Intention and Emotional Involvement

Is the system 'satisfied' in some way by what it is doing?
What is built into the system ('inherent', fundamental) for creative improvisation?
What motivates the system to improvise?
Can it choose to improvise?
Does the system get some reward from doing improvisation? Does it enjoy it?
Is there any element of pain and/or pleasure for the system?
What impulses and non-built in behaviour can the system follow?
What human emotions and behaviours does the system simulate/'experience'/'feel'? Fear? Self-discovery? Inhibition? Attitude? Willingness? Intuition? Spiritual?
How does the system express a personality and individual style?
How sincere and transparent is the system in reporting how it works?
To what extent is the system displaying an inner process of creative improvisation?
Is the system 'in the moment'?
How emotionally involved is the system in the process?
Does the system trust its own results?
Is there a sense of play, playfulness, fun, childish discovery, adventure in the way the system operates?
Can the system display different moods at different times?
How confident is the system in what it does? Does this change with context/results/reception?
Does the system display passion and desire for the process?
What attracts the system to improvise? and to make music?
How interested is the system in what it is doing?
Can the system express feelings, emotions, moods?
Is the system stimulated into improvising?
How intensely is the system involved in what it is doing? Or does it abstract from the process?
Can the system work with loosely specified information such as feel or emotion?
How sensitive is the system to outside interference/feedback?
What fuels and directs the system's process?

Active Involvement and Persistence

How is the system embodied in the improvisation process?
What is the physical appearance and make-up of the system?
Does the system have a number of options and steps open to it?
Can the system follow different paths to try different things out?
How does the system keep going over the creative process?
What practical steps does the system take to come up with, enhance and develop the improvisation?
How does the system recover from problems in the process?
Can the system stagnate?
Does the system actually do anything?
How pragmatic and functional is the system?
How active and busy is the system? Is the system continually active or can it occasionally stop?
How persistent is the system process?

How actively involved is the system in the process?
How does the system cope with external distractions and problems that may affect it?
Can the system compensate for problems in the process?
Can the system get lost? How does it recover?
What does the system find difficult? How does it compensate for this?
Is the system aware of any shortcomings it has, and can it compensate?
Can the system cope if it has been given poorly presented/bad format input?

Variety, Divergence and Experimentation

What experimentation can the system do?
How does the system use existing knowledge to experiment?
How open is the system to new input? What new input can the system access?
Can the system come up with new ideas?
Can the system produce results using an open approach?
How does the system come up with a variety of different things?
Can the system demonstrate 'imagination'?
Can the system diverge from its own program/instructions and experiment/explore?
How does the system demonstrate 'divergent thinking' (going in new directions)?
How inventive is the system?
Is the system biased towards a certain perspective?
How flexible is the system?
Can the system change its approach or come up with new ideas if needed?
Can the system diverge from the theme of the music, and play 'out' of the harmonic structure?
Can the system behave in an unexpected way?
Can the system break rules or transcend constraints?
Can the system test different things out?
Can the system detect constraints and try to break them?
What constraints does the system work under?
What different styles, structures and forms can the system work with? Can it generate its own style?
Can the system interpret information in different ways?
How does the system explore unknown territory?
How does the system allow for domain transformation?
Is the system limited to a prescribed style?
Is there variety in how the system approaches improvisation (or subtasks)?

Dealing with Uncertainty

How does the system deal with new things?
To what extent does the system use unfamiliar methods or material?
What risks does the system take (e.g. trying things that may or may not work)?
Can the system develop and try out new heuristics or new methods?
Can the system employ methods for which there is no guaranteed results?
How does the system deal with irrational or contradictory information?
How does the system deal with missing information?
How does the system cope when standard methods or material is not appropriate?
Can the system cope with poorly specified/ambiguous input or instructions?
Can the system cope with external distractions or mistakes by co-performers?
If performing with other musicians, does the system attempt predictions of what they will do?
How does the system use incoming data in real-time (e.g. with interactive performance)?
Does the system encounter situations where it has not been given pre-defined rules to follow? If so, how does it deal with this?

Originality

Can the system come up with new ideas?
Can the system generate musical material it has not been exposed to before?
How original is the end result of the system?
How unexpected are the resulting improvisations?
Can the system generate different material on different attempts?
Is the system deterministic or non-deterministic?
Can the system output be predicted if you know the program it follows and the input it is given?
Can the system express known information in new ways?

Can the system produce improvisations which transform domain rules and constraints?
Can the system find new ways of improvising which have not been done before?
Can the system produce surprising results? (If so, are these results surprising to the system implementers?)
Can the system produce unique results?
Is there potential for the system output to be re-interpreted by others in a new way?
To what extent does the system use unfamiliar patterns and non-prescribed heuristics?

Spontaneity and Subconscious Processing

Does the system wait for creative inspiration?
What creative impulses can the system generate/receive?
Can the system generate improvisations in real-time?
Can the system enter a 'flow'-like state where improvisation is generated fluidly and smoothly?
How much latency is there in the system?
Can the system take chances (in real time) and recover from any risks if needed?
How spontaneous can the system be?
How much processing time does the system need?
Can the system operate without much pre-planning?
To what extent is the system in control of all its processes and decisions?
Does anything occur in the system at a level where it is not centrally controlled?
How aware is the system of all that is happening during the creative process?

Independence and Freedom

Does the system have opportunities to express itself?
Can the system operate without conforming to a norm?
How much freedom does the system have to operate in?
How restricted is the system?
To what extent can the system make its own creative decisions?
Can the system break out of any pre-programmed rules it may have?
How much inspiring material does the system need?
To what extent is the system controlled/restricted to certain approaches by its programming?
How much structure does the system need?
What external moderation is done of the system's output (e.g. by external performance hardware)?
To what extent is an improvisation bound by the rules of a piece or structure it has been given?
To what extent can a system transcend constraints?
Does the system operate in a way that chance and risk can occur?
How stylistically free or varied can the system's improvisations be?
What constraints restrict the system's operations?
How much does the system rely on being given information on music theory or existing traditions?
Can the system operate independently, without assistance?
To what extent is the system limited by input from others? (e.g. does it need interactive input?)
How much space does the system get to develop ideas?

Progression and Development

How does the system supplement existing knowledge with new or learned knowledge?
Can the system learn new things? and use the new information in its creative process?
How does the system develop its improvisations during the creative process?
Can the system develop the way in which it operates?
Can the system develop/refine ideas?
Does the system progress during an improvisation, from initial ideas to a final improvisation?
Does the system make contributions towards domain progress, e.g. by identifying constraints which could be broken?
Does the system produce improvisations which move towards an end goal or have some global structure?
Can the system retain references/links to original basic structures while still developing an improvisation?
What directions do the system's improvisation take?
Is the system output influenced by music just preceding that improvisation?
Is the system output influenced by previous improvisations/interactive input?
Do the system's improvisations build upon what has already happened in that improvisation?
How do the system's improvisations develop over the course of that improvisation?

Thinking and Evaluation

Can the system evaluate its own output and be 'satisfied' with what it produces?
Can the system balance several relevant factors and/or considerations?
Can the system recognise what is appropriate?
Does the system recognise when an improvisation it is generating is nearing completion?
Can the system evaluate its output in the context of constraints which were imposed (e.g. structures, chord progressions)?
Can the system be logical about its music?
Can the system monitor how well it is doing?
What decisions does the system have to make?
How does the system make decisions?
Does the system have good judgement?
Does the system continually evaluate, contextualise and re-evaluate?
How does the system allow for different interpretations of the same improvisation by two different parties?
What information does the system require to make decisions?
How does the computer form opinions about and judge the quality of improvisations?
Can the system operate in a rational way?

Value

Do the produced improvisations satisfy the system in terms of having produced what it needs to?
Do the produced improvisations appear polished to an external listener?
Do the produced improvisations sound good to the audience?
Do the produced improvisations have some worth and value, as improvisations?
Can the system's output provoke a response from others such as 'now that IS good'?
Can the system hide any sophisticated logic or complex processes and produce improvisations that sound simple?
How does the system's output compare to what has been done before (e.g. historically good improvisation)?
Does the system provide anything to be learnt from, e.g. a new approach or transformation of domain constraints?
Does the system's output seem to 'work' and be correct according to what type of output is expected?
Have stylistic/auditory/other constraints been adhered to, where relevant?
Is the output of sufficient depth to be interpreted at different levels or in different ways?
Does the system output stand alone as a musical improvisation?

Generation of Results

Does the system produce musical improvisations?
Can the system perform its improvisations musically, by making sound?
Does the system create improvisations during the process (rather than replicating stored music)?
Can the system produce new things?
Does the system know when it has produced an end product?
What output does the system produce?
Can the system produce results in real-time?

General Intellect

Can the system employ lateral thinking (i.e. come up with innovative approaches) when needed?
How does the system organise its processes?
Generally, how intelligent is the system?
Does the system employ alternative modes of thought?

Appendix E

Evaluation form used to evaluate systems in Case Study 2 (Chapter 7)

EVALUATION for System:_____ Author(s):_____

Domain (art/maths/etc):_____ My domain expertise: Basic / Reasonable / Expert

___ 10	**Active Involvement and Persistence** • Being actively involved; reacting to and having a deliberate effect on the creative process. • The tenacity to persist with the creative process throughout, even during problematic points.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Creation of Results** • Working towards some end target, or goal, or result. • Producing something (tangible or intangible) that previously did not exist.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Dealing with Uncertainty** • Coping with incomplete, missing, inconsistent, contradictory, ambiguous and/or uncertain information. Element of risk and chance - no guarantee that information problems will be resolved. • Not relying on every step of the process to be specified in detail; perhaps even avoiding routine or pre-existing methods and solutions.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Domain Competence** • Domain-specific intelligence, knowledge, talent, skills, experience and expertise. • Knowing a domain well enough to be equipped to recognise gaps, needs or problems that need solving and to generate, validate, develop and promote new ideas in that domain.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**General Intellect** • General intelligence and IQ. • Good mental capacity.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Independence and Freedom** • Working independently with autonomy over actions and decisions. • Freedom to work without being bound to pre-existing solutions, processes or biases; perhaps challenging cultural or domain norms.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Intention and Emotional Involvement** • Personal and emotional investment, immersion, self-expression and involvement in the creative process. • The intention and desire to be creative; creativity is its own reward, a positive process giving fulfilment and enjoyment.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*

___ 10	**Originality** • Novelty and originality - a new product, or doing something in a new way, or seeing new links and relations between previously unassociated concepts. • Results that are unpredictable, unexpected, surprising, unusual, out of the ordinary.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Progression and Development** • Movement, advancement, evolution and development during the creative process. • Whilst progress may or may not be linear, and an actual end goal may be only loosely specified (if at all), the entire process should represent some progress in a particular domain or task.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Social Interaction and Communication** • Communicating and promoting creative work to others in a persuasive and positive manner. • Mutual influence, feedback, sharing and collaboration between society and individual.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Spontaneity and Subconscious Processing** • No need to be in control of the whole process - thoughts and activities may inform the creative process subconsciously without being inaccessible for conscious analysis, or may receive less attention than others. • Being able to react quickly and spontaneously during the creative process when appropriate, without needing to spend time thinking about options too much.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Thinking and Evaluation** • Consciously evaluating several options to recognise potential value in each and identify the best option, using reasoning and good judgment. • Proactively selecting a decided choice from possible options, without allowing the process to stagnate under indecision.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Value** • Making a useful contribution that is valued by others and recognized as an achievement and influential advancement; perceived as special, 'not just something anybody would have done'. • End product is relevant and appropriate to the domain being worked in.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*
___ 10	**Variety, Divergence and Experimentation** • Generating a variety of different ideas to compare and choose from, with the flexibility to be open to several perspectives and to experiment and try different options out without bias. • Multi-tasking during the creative process.	*Importance/Relevance for creativity:* □ *Crucial for creativity* □ *Quite important* □ *A little important* □ *Irrelevant to domain*

ANY OTHER COMMENTS?

Appendix F

Derivation of Ritchie's criteria for each of the case study systems

---

In Chapter 8, creativity evaluation results obtained using SPECS evaluation were compared to the results obtained using Ritchie's empirical criteria framework (Ritchie, 2007). The three musical improvisation systems in Case Study 1 were evaluated using Ritchie's criteria: GAmprovising (Jordanous, 2010c), GenJam (Biles, 2007) and Voyager (Lewis, 2000). Five systems from ICCC'11 were evaluated for Case Study 2. As Chapter 8 Section 8.2.3 has shown, there was insufficient information available to apply Ritchie's criteria to two systems: Tearse et al.'s reconstruction of the story generator MINSTREL (Tearse et al., 2011) and Monteith et al.'s musical soundtrack generator (Monteith et al., 2011). The criteria were applied successfully to the other Case Study 2 systems: Cook and Colton's collage generation module for *The Painting Fool* (Cook & Colton, 2011), Rahman and Manurung's poetry generator (Rahman & Manurung, 2011) and Norton et al.'s image generator *DARCI* (Norton et al., 2011).

Ritchie's criteria applied to the Case Study 1 musical improvisation systems

Ritchie's criteria framework is detailed in Chapter 2 Section 2.1.2; here a brief summary acts as reminder of key points. Working through the framework in Ritchie (2007) the key parts of the formalisation needing further definition before application are:[1]

- Results set $\mathcal{R}$: items output by the program. (See Chapter 8 Section 8.2.1.)
- Inspiring set I: items either used in the program as source material or in program construction.

  This is difficult to define for each system. Licks databases and stored melodies are used but not made available for GenJam (Biles, 2007). It would be challenging to identify which databases were used to train these two systems for the individual tracks selected, even by the programmers themselves, particularly as some of these tracks are now many years old (for example, GenJam's 'Analog Blues' was released on CD in 1996). Voyager and GenJam take interactive input to construct parts of their improvisations (Lewis, 2000; Biles, 2007) and GAmprovising uses random note generation (Jordanous, 2010c).

[1] Explanations are paraphrased from Ritchie (2007, pp. 76-77)

The inspiring set for these systems could be defined as all musical improvisations, or all jazz improvisations for GenJam and all avant-garde/free improvisations for Voyager and GAmprovising. The size of the inspiring set would best be represented as ∞ as this set would be both extremely large and uncountable. As Chapter 8 Section 8.3.4 discusses, though, an inspiring set of size ∞ would lead to criteria 9-18 being identical for each system.

Alternatively, the inspiring set could be defined as the empty set $\emptyset$ but this does not make use of what we know about the inspiring sets for some of the programs.

In the knowledge that GenJam uses licks databases and stored melodies, we can say that although the contents of the inspiring set are unknown, it is non-empty, i.e. $|I| > 0$. The products of this system are constructed from different members of I but no members of $\mathcal{R}$ are members of I (though *parts* of members of $\mathcal{R}$ may be members of I). Therefore for use in the criteria, $\mathcal{R} - I = \mathcal{R}$ and $\mathcal{R} \cap I = \emptyset$.

Voyager and GAmprovising do not use an inspiring set (Lewis, 2000; Jordanous, 2010c).[2] For Voyager and GAmprovising, $I = \emptyset$, $\mathcal{R} - I = \mathcal{R}$ and $\mathcal{R} \cap I = \emptyset$.

- *typ*: the typicality rating scheme. Each item in $\mathcal{R}$ was rated during the survey for typicality. The Likert ratings given (using the scale in Chapter 8 Figure 8.4) were converted into numeric form of [1, 2, 3, 4, 5]: Strongly disagree = 1, through to Strongly Agree = 5. As Ritchie's rating schemes should range between 0 and 1 (for the purposes of the T and V functions) the ratings were normalised such that Strongly disagree = 0.2 and Strongly Agree = 1.0.
- *val*: the value rating scheme. Similarly to typicality, each item in $\mathcal{R}$ was rated for value, and the Likert ratings converted to numeric form as above.
- $T_{\alpha,\beta}(\mathcal{R})$: the subset of items in $\mathcal{R}$ falling in an acceptable range of typicality. 'Acceptable typicality' was defined in this case as giving a more positive agreement response than 'neutral' to the question about typicality. As the response 'Neutral' corresponds to a rating of 0.6:
  - $T_{\alpha,1}(\mathcal{R})$, $\alpha = 0.6$: If the mean *typ* rating for the items in $\mathcal{R} > 0.6$ then an acceptable level of typicality is reached.
  - $T_{0,\beta}(\mathcal{R})$, $\beta = 0.6$: If the mean *typ* rating for the items in $\mathcal{R} < 0.6$ then an acceptable level of atypicality is reached.
  - 'Neutral' answers are discarded, for the purposes of this evaluation.
- $V_{\alpha,\beta}(\mathcal{R})$: the subset of items in $\mathcal{R}$ falling in an acceptable range of quality. 'Acceptable quality' was defined in this case as giving a more positive agreement response than 'neutral' to the question about value. As the response 'Neutral' corresponds to a rating of 0.6:
  - $V_{\gamma,1}(\mathcal{R})$, $\gamma = 0.6$: If the mean *val* rating for the items in $\mathcal{R} > 0.6$ then an acceptable level of quality is reached.
  - No criteria ask for the subset of low-valued items produced by the system.
  - 'Neutral' answers are discarded, for the purposes of this evaluation.
- $AV(F,X)$: taking the average value of function F over set X (here the mean average is used).
- $ratio(X,Y)$: the size of set X divided by the size of set Y (where $|Y| \neq 0$).
- θ, 'general comparison level in all criteria' (Ritchie, 2007, p. 79). A single acceptable threshold value for each criterion to pass if it can be said to be satisfied. In keeping with previous

[2] Although Voyager (and GenJam) take input from co-performers during run time it is difficult to see how to include that in I without making this set infinitely large.

applications of Ritchie's criteria (Gervás, 2002; Pereira et al., 2005), θ is set at 0.5.

Table F.1 lists the survey findings for each item in the three $\mathcal{R}$ sets and the corresponding *typ* and *val* ratings, indicating how results fit into the relevant T and V functions for each system.

Table F.1: Survey data for the *typ* and *val* rating schemes. The superscript ^{T} in the *typ* column indicates that an item is acceptably typical as it satisfies $T_{0.6,1}$. The superscript ^{A} in the *typ* column indicates that an item is acceptably atypical as it satisfies $T_{0,0.6}$. The superscript ^{V} in the *val* column indicates that an item is acceptably high quality as it satisfies $V_{0.6,1}$.

System	item $r \in \mathcal{R}$	Mean typicality	*typ*	Mean value	*val*
GAmprovising	Track 1	2.9	0.58^A	2.6	0.52
GAmprovising	Track 2	3.1	0.62^T	2.7	0.54
GAmprovising	Track 3	3.0	0.6	2.7	0.54
GenJam	Analog Blues	3.3	0.66^T	2.7	0.54
GenJam	Change Tranes	4.0	0.8^T	3.5	0.7^V
GenJam	The Rake	3.9	0.78^T	3.3	0.66^V
Voyager	Duo 1	3.2	0.64^T	2.8	0.56
Voyager	Duo 2	3.4	0.68^T	3.1	0.62^V
Voyager	Duo 3	3.1	0.62^T	3.0	0.6

The 8th and 10th criteria are listed as 8a and 10a respectively. This is in keeping with Ritchie (2007), where the original criterion 8 from Ritchie (2001) is replaced with a revised criterion 8a and similarly 10a replaces 10, due to formal flaws with the original criteria 8 and 10. There is an error in Ritchie (2007) in criterion 17: the second closing bracket should not be followed by a third closing bracket, as this unbalances the brackets. This is corrected in the criteria as used below.[3]

GAmprovising

1. $AV(typ, \mathcal{R}) > \theta \quad \longrightarrow \quad \frac{(0.58+0.62+0.6)}{3}) = 0.6 > 0.5 \therefore TRUE$
2. $ratio(T_{\alpha,1}(\mathcal{R}), \mathcal{R}) > \theta \quad \longrightarrow \quad \frac{1}{3} = 0.333 \not> 0.5 \therefore FALSE$
3. $AV(val, \mathcal{R}) > \theta \quad \longrightarrow \quad \frac{(0.52+0.54+0.54)}{3}) = 0.533 > 0.5 \therefore TRUE$
4. $ratio(V_{\gamma,1}(\mathcal{R}), \mathcal{R}) > \theta \quad \longrightarrow \quad \frac{0}{3} = 0 \not> 0.5 \therefore FALSE$
5. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{\alpha,1}(\mathcal{R}), T_{\alpha,1}(\mathcal{R})) > \theta \quad \longrightarrow \quad \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$
6. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), \mathcal{R}) > \theta \quad \longrightarrow \quad \frac{0}{3} = 0 \not> 0.5 \therefore FALSE$
7. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), T_{0,\beta}(\mathcal{R})) > \theta \quad \longrightarrow \quad \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$

8a. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), V_{\gamma,1}(\mathcal{R})) > \theta \quad \longrightarrow \quad \frac{0}{0} = undefined$[4] $\therefore$ not applicable

9. $ratio(I \cap \mathcal{R}, I) > \theta \quad \longrightarrow \quad \frac{0}{0} = undefined \therefore$ not applicable

10a. $(1 - ratio(I \cap \mathcal{R}, \mathcal{R})) > \theta \quad \longrightarrow \quad 1 - \frac{0}{3} = 1 > 0.5 \therefore TRUE$

11. $AV(typ, (\mathcal{R} - I)) > \theta \quad \equiv$ Criterion 1 $\therefore TRUE$
12. $AV(val, (\mathcal{R} - I)) > \theta \quad \equiv$ Criterion 3 $\therefore TRUE$
13. $ratio(T_{\alpha,1}(\mathcal{R} - I), \mathcal{R}) > \theta \quad \equiv$ Criterion 2 $\therefore FALSE$
14. $ratio(V_{\gamma,1}(\mathcal{R} - I), \mathcal{R}) > \theta \quad \equiv$ Criterion 4 $\therefore FALSE$
15. $ratio(T_{\alpha,1}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \quad \equiv$ Criteria 2, 13 $\therefore FALSE$

[3] Where necessary, all answers below are rounded to 3 significant figures.

[4] It is a point of controversy whether $\frac{0}{0} = 1$ (because any number divided by itself is 1), or $\frac{0}{0} = 0$ (because 0 divided by any number is 0), or if $\frac{0}{0}$ is undefined. The last option is consistently adopted here, because of the ambiguity of $\frac{0}{0}$.

16. $ratio(V_{\gamma,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criteria 4, 14 $\therefore FALSE$
17. $ratio(V_{\gamma,1}(\mathcal{R}-I)\cap T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\longrightarrow$ $\frac{0}{3}=0\not>0.5\therefore FALSE$
18. $ratio(V_{\gamma,1}(\mathcal{R}-I)\cap T_{0,\beta}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criterion 6 $\therefore FALSE$

GenJam

1. $AV(typ,\mathcal{R})>\theta$ $\longrightarrow$ $\frac{(0.66+0.8+0.78)}{3})=0.747>0.5\therefore TRUE$
2. $ratio(T_{\alpha,1}(\mathcal{R}),\mathcal{R})>\theta$ $\longrightarrow$ $\frac{3}{3}=1>0.5\therefore TRUE$
3. $AV(val,\mathcal{R})>\theta$ $\longrightarrow$ $\frac{(0.54+0.7+0.66)}{3})=0.633>0.5\therefore TRUE$
4. $ratio(V_{\gamma,1}(\mathcal{R}),\mathcal{R})>\theta$ $\longrightarrow$ $\frac{2}{3}=0.667>0.5\therefore TRUE$
5. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{\alpha,1}(\mathcal{R}),T_{\alpha,1}(\mathcal{R}))>\theta$ $\longrightarrow$ $\frac{2}{3}=0.667>0.5\therefore TRUE$
6. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{0,\beta}(\mathcal{R}),\mathcal{R})>\theta$ $\longrightarrow$ $\frac{0}{3}=0\not>0.5\therefore FALSE$
7. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{0,\beta}(\mathcal{R}),T_{0,\beta}(\mathcal{R}))>\theta$ $\longrightarrow$ $\frac{0}{0}=undefined\therefore$ not applicable

8a. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{0,\beta}(\mathcal{R}),V_{\gamma,1}(\mathcal{R}))>\theta$ $\longrightarrow$ $\frac{0}{2}=0\not>0.5\therefore FALSE$

9. $ratio(I\cap\mathcal{R},I)>\theta$ $\longrightarrow$ $\frac{0}{x}|x>0=0\not>0.5\therefore FALSE$

10a. $(1-ratio(I\cap\mathcal{R},\mathcal{R}))>\theta$ $\longrightarrow$ $1-\frac{0}{3}=1>0.5\therefore TRUE$

11. $AV(typ,(\mathcal{R}-I))>\theta$ $\equiv$ Criterion 1 $\therefore TRUE$
12. $AV(val,(\mathcal{R}-I))>\theta$ $\equiv$ Criterion 3 $\therefore TRUE$
13. $ratio(T_{\alpha,1}(\mathcal{R}-I),\mathcal{R})>\theta$ $\equiv$ Criterion 2 $\therefore TRUE$
14. $ratio(V_{\gamma,1}(\mathcal{R}-I),\mathcal{R})>\theta$ $\equiv$ Criterion 4 $\therefore TRUE$
15. $ratio(T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criteria 2, 13 $\therefore TRUE$
16. $ratio(V_{\gamma,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criteria 4, 14 $\therefore TRUE$
17. $ratio(V_{\gamma,1}(\mathcal{R}-I)\cap T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\longrightarrow$ $\frac{2}{3}=0.667>0.5\therefore TRUE$
18. $ratio(V_{\gamma,1}(\mathcal{R}-I)\cap T_{0,\beta}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criterion 6 $\therefore FALSE$

Voyager

1. $AV(typ,\mathcal{R})>\theta$ $\longrightarrow$ $\frac{(0.64+0.68+0.62)}{3})=0.647>0.5\therefore TRUE$
2. $ratio(T_{\alpha,1}(\mathcal{R}),\mathcal{R})>\theta$ $\longrightarrow$ $\frac{3}{3}=1>0.5\therefore TRUE$
3. $AV(val,\mathcal{R})>\theta$ $\longrightarrow$ $\frac{(0.56+0.62+0.6)}{3})=0.593>0.5\therefore TRUE$
4. $ratio(V_{\gamma,1}(\mathcal{R}),\mathcal{R})>\theta$ $\longrightarrow$ $\frac{1}{3}=0.333\not>0.5\therefore FALSE$
5. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{\alpha,1}(\mathcal{R}),T_{\alpha,1}(\mathcal{R}))>\theta$ $\longrightarrow$ $\frac{1}{3}=0.333\not>0.5\therefore FALSE$
6. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{0,\beta}(\mathcal{R}),\mathcal{R})>\theta$ $\longrightarrow$ $\frac{0}{3}=0\not>0.5\therefore FALSE$
7. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{0,\beta}(\mathcal{R}),T_{0,\beta}(\mathcal{R}))>\theta$ $\longrightarrow$ $\frac{0}{0}=undefined\therefore$ not applicable

8a. $ratio(V_{\gamma,1}(\mathcal{R})\cap T_{0,\beta}(\mathcal{R}),V_{\gamma,1}(\mathcal{R}))>\theta$ $\longrightarrow$ $\frac{0}{1}=0\not>0.5\therefore FALSE$

9. $ratio(I\cap\mathcal{R},I)>\theta$ $\longrightarrow$ $\frac{0}{0}=undefined\therefore$ not applicable

10a. $(1-ratio(I\cap\mathcal{R},\mathcal{R}))>\theta$ $\longrightarrow$ $1-\frac{0}{3}=1>0.5\therefore TRUE$

11. $AV(typ,(\mathcal{R}-I))>\theta$ $\equiv$ Criterion 1 $\therefore TRUE$
12. $AV(val,(\mathcal{R}-I))>\theta$ $\equiv$ Criterion 3 $\therefore TRUE$
13. $ratio(T_{\alpha,1}(\mathcal{R}-I),\mathcal{R})>\theta$ $\equiv$ Criterion 2 $\therefore TRUE$
14. $ratio(V_{\gamma,1}(\mathcal{R}-I),\mathcal{R})>\theta$ $\equiv$ Criterion 4 $\therefore FALSE$
15. $ratio(T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criteria 2, 13 $\therefore TRUE$
16. $ratio(V_{\gamma,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criteria 4, 14 $\therefore FALSE$
17. $ratio(V_{\gamma,1}(\mathcal{R}-I)\cap T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\longrightarrow$ $\frac{1}{3}=0.333\not>0.5\therefore FALSE$
18. $ratio(V_{\gamma,1}(\mathcal{R}-I)\cap T_{0,\beta}(\mathcal{R}-I),(\mathcal{R}-I))>\theta$ $\equiv$ Criterion 6 $\therefore FALSE$

Ritchie's criteria applied to the Case Study 2 ICCC'11 systems

Application of the criteria for evaluation

Working through the framework in Ritchie (2007) similarly to in Case Study 1:

- Results set $\mathcal{R}$: items output by the program. (See Chapter 8 Section 8.2.3.)

- Inspiring set I: items either used in the program as source material or in program construction. (See Chapter 8 Section 8.2.3.)
- *typ*: the typicality rating scheme. This was defined in Chapter 8 Section 8.2.3 for each system; the results are in Chapter 8 Table 8.9. As there are two judges, each item in $\mathcal{R}$ has two typicality ratings that are averaged to produce the overall typicality rating for an artefact in $\mathcal{R}$.
- *val*: the value rating scheme. This was defined in Chapter 8 Section 8.2.3 for each system; the results are in Chapter 8 Table 8.9. As there are two judges, each item in $\mathcal{R}$ has two value ratings which are averaged to produce the overall value rating for an artefact in $\mathcal{R}$.
- $T_{\alpha,\beta}(\mathcal{R})$: the subset of items in $\mathcal{R}$ falling in an acceptable range of typicality.
  - $T_{\alpha,1}(\mathcal{R})$: In all cases, an acceptable level of typicality is reached if at least one judge answers 'Strongly Agree' or 'Agree' to the question about typicality of a given artefact and the other judge answers 'Strongly Agree', 'Agree' or 'Neutral'. The threshold value for typicality (α) is when one judge answers 'Agree' and the other judge answers 'Neutral'. Using the numeric representations of the Likert scale given below, $\alpha = \frac{(0.75+0.5)}{2} = 0.625$. This forms the lower parameter (inclusive) for T. The higher parameter 1, represents the highest possible rating (where both judges answer 'Strongly Agree': $\frac{(1.0+1.0)}{2} = 1.0$).
  - $T_{0,\beta}(\mathcal{R})$: In all cases, an acceptable level of atypicality is reached if at least one judge answers 'Strongly Disagree' or 'Disagree' to the question about typicality of a given artefact and the other judge answers 'Strongly Disagree', 'Disagree' or 'Neutral'. Therefore the threshold value for atypicality (β) is when one judge answers 'Disagree' and the other judge answers 'Neutral'. Using the numeric representations of the Likert scale given below, $\alpha = \frac{(0.25+0.5)}{2} = 0.375$. This forms the upper argument (inclusive) for T. The second argument to T, 1, represents the lowest possible rating (where both judges answer 'Strongly Disagree': $\frac{(0.0+0.0)}{2} = 0.0$).
  - 'Neutral' answers are discarded, for the purposes of this evaluation.
- $V_{\alpha,\beta}(\mathcal{R})$: the subset of items in $\mathcal{R}$ falling in an acceptable range of quality:
  - $V_{\gamma,1}(\mathcal{R})$: In all cases, an acceptable level of quality is reached if at least one judge answers 'Strongly Agree' or 'Agree' to the question about quality of a given artefact and the other judge answers 'Strongly Agree', 'Agree' or 'Neutral'. Therefore the threshold value for quality (α) is when one judge answers 'Agree' and the other judge answers 'Neutral'. Using the numeric representations of the Likert scale given below, $\alpha = \frac{(0.75+0.5)}{2} = 0.625$. This forms the lower argument (inclusive) for V. The second argument to V, 1, represents the highest possible rating (where both judges answer 'Strongly Agree': $\frac{(1.0+1.0)}{2} = 1.0$).
  - No criteria ask for the subset of low-valued items produced by the system.
  - 'Neutral' answers are discarded, for the purposes of this evaluation.
- $AV(F,X)$: taking the average value of function F over set X (here the mean average is used).
- $ratio(X,Y)$: the size of set X divided by the size of set Y (where $|Y| \neq 0$).
- θ, 'general comparison level in all criteria' (Ritchie, 2007, p. 79). A single acceptable threshold value for each criterion to pass to be satisfied. In keeping with previous applications of Ritchie's criteria (Gervás, 2002; Pereira et al., 2005, Case Study 1), $\theta = 0.5$.

In calculations, the Likert scales for typicality and value are converted to numeric form as follows:

1.0 = Strongly agree
0.75 = Agree
0.5 = Neutral
0.25 = Disagree
0.0 = Strongly disagree

As for Case Study 1, the revised criteria 8a and 10a (Ritchie, 2007) are used in place of criteria 8 and 10 and the unbalanced brackets in criterion 17 are corrected.

If $(\mathcal{R} - I) = \emptyset$, i.e. if no items in the inspiring set I are reproduced in the results set $\mathcal{R}$, '[a]ll of criteria 11-18 would be inapplicable'(Ritchie, 2007, p. 81). None of criteria 11-18 are therefore applicable as none of the systems generate artefacts from their inspiring set:

- Cook and Colton (2011) generates collages from the inspiring set images and the example collage is made from several inspiring set pictures and one inspiring set news story.
- The target limerick in Rahman and Manurung (2011) is not produced in the results set.
- Although *DARCI* (Norton et al., 2011) is able to re-render the three images in the inspiring set, it does not reproduce the original inspiring set images in its results.

Cook & Colton's collage generation module for *The Painting Fool*

1. $AV(typ, \mathcal{R}) > \theta \longrightarrow mean(\frac{(0.0+0.75)}{2}) = 0.375 \not> 0.5 \therefore FALSE$
2. $ratio(T_{\alpha,1}(\mathcal{R}), \mathcal{R}) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$
3. $AV(val, \mathcal{R}) > \theta \longrightarrow mean(\frac{(0.75+0.25)}{2}) = 0.5 \not> 0.5 \therefore FALSE$
4. $ratio(V_{\gamma,1}(\mathcal{R}), \mathcal{R}) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$
5. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{\alpha,1}(\mathcal{R}), T_{\alpha,1}(\mathcal{R})) > \theta \longrightarrow \frac{0}{0} = undefined \therefore$ not applicable
6. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), \mathcal{R}) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$
7. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), T_{0,\beta}(\mathcal{R})) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$

8a. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), V_{\gamma,1}(\mathcal{R})) > \theta \longrightarrow \frac{0}{0} = undefined \therefore$ not applicable

9. $ratio(I \cap \mathcal{R}, I) > \theta \longrightarrow \frac{0}{\infty} = 0 \not> 0.5 \therefore FALSE$

10a. $(1 - ratio(I \cap \mathcal{R}, \mathcal{R})) > \theta \longrightarrow 1 - \frac{0}{1} = 1 > 0.5 \therefore TRUE$

11. $AV(typ, (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable
12. $AV(val, (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable
13. $ratio(T_{\alpha,1}(\mathcal{R} - I), \mathcal{R}) > \theta \longrightarrow$ not applicable
14. $ratio(V_{\gamma,1}(\mathcal{R} - I), \mathcal{R}) > \theta \longrightarrow$ not applicable
15. $ratio(T_{\alpha,1}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable
16. $ratio(V_{\gamma,1}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable
17. $ratio(V_{\gamma,1}(\mathcal{R} - I) \cap T_{\alpha,1}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable
18. $ratio(V_{\gamma,1}(\mathcal{R} - I) \cap T_{0,\beta}(\mathcal{R} - I), (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable

Rahman & Manurung's poetry generator

1. $AV(typ, \mathcal{R}) > \theta \longrightarrow mean(\frac{(0.25+0.75)}{2}) = 0.5 \not> 0.5 \therefore FALSE$
2. $ratio(T_{\alpha,1}(\mathcal{R}), \mathcal{R}) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$
3. $AV(val, \mathcal{R}) > \theta \longrightarrow mean(\frac{(0.25+0.25)}{2}) = 0.25 \not> 0.5 \therefore FALSE$
4. $ratio(V_{\gamma,1}(\mathcal{R}), \mathcal{R}) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$
5. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{\alpha,1}(\mathcal{R}), T_{\alpha,1}(\mathcal{R})) > \theta \longrightarrow \frac{0}{0} = undefined \therefore$ not applicable
6. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), \mathcal{R}) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$
7. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), T_{0,\beta}(\mathcal{R})) > \theta \longrightarrow \frac{0}{0} = undefined \therefore$ not applicable

8a. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}), V_{\gamma,1}(\mathcal{R})) > \theta \longrightarrow \frac{0}{0} = undefined \therefore$ not applicable

9. $ratio(I \cap \mathcal{R}, I) > \theta \longrightarrow \frac{0}{1} = 0 \not> 0.5 \therefore FALSE$

10a. $(1 - ratio(I \cap \mathcal{R}, \mathcal{R})) > \theta \longrightarrow 1 - \frac{0}{1} = 1 > 0.5 \therefore TRUE$

11. $AV(typ, (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable
12. $AV(val, (\mathcal{R} - I)) > \theta \longrightarrow$ not applicable

13. $ratio(T_{\alpha,1}(\mathcal{R}-I),\mathcal{R}) > \theta \longrightarrow$ not applicable
14. $ratio(V_{\gamma,1}(\mathcal{R}-I),\mathcal{R}) > \theta \longrightarrow$ not applicable
15. $ratio(T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
16. $ratio(V_{\gamma,1}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
17. $ratio(V_{\gamma,1}(\mathcal{R}-I) \cap T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
18. $ratio(V_{\gamma,1}(\mathcal{R}-I) \cap T_{0,\beta}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable

Norton et al.'s image generator *DARCI*

1. $AV(typ,\mathcal{R}) > \theta \longrightarrow mean(\frac{(0.75+0.75)}{2}, \frac{(1.0+0.5)}{2}, \frac{(0.0+0.25)}{2}, \frac{0.5+0.0}{2}) = 0.46875 \not> 0.5 \therefore FALSE$
2. $ratio(T_{\alpha,1}(\mathcal{R}),\mathcal{R}) > \theta \longrightarrow \frac{2}{4} = 0.5 \not> 0.5 \therefore FALSE$
3. $AV(val,\mathcal{R}) > \theta \longrightarrow mean(\frac{(0.75+0.5)}{2}, \frac{(0.5+0.75)}{2}, \frac{(0.0+0.25)}{2}, \frac{0.5+0.25}{2}) = 0.4375 \not> 0.5 \therefore FALSE$
4. $ratio(V_{\gamma,1}(\mathcal{R}),\mathcal{R}) > \theta \longrightarrow \frac{2}{4} = 0.5 \not> 0.5 \therefore FALSE$
5. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{\alpha,1}(\mathcal{R}),T_{\alpha,1}(\mathcal{R})) > \theta \longrightarrow \frac{2}{2} = 1 > 0.5 \therefore TRUE$
6. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}),\mathcal{R}) > \theta \longrightarrow \frac{0}{4} = 0 \not> 0.5 \therefore FALSE$
7. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}),T_{0,\beta}(\mathcal{R})) > \theta \longrightarrow \frac{0}{2} = 0 \not> 0.5 \therefore FALSE$

8a. $ratio(V_{\gamma,1}(\mathcal{R}) \cap T_{0,\beta}(\mathcal{R}),V_{\gamma,1}(\mathcal{R})) > \theta \longrightarrow \frac{0}{2} = 0 \not> 0.5 \therefore FALSE$

9. $ratio(I \cap \mathcal{R}, I) > \theta \longrightarrow \frac{0}{3} = 0 \not> 0.5 \therefore FALSE$

10a. $(1 - ratio(I \cap \mathcal{R},\mathcal{R})) > \theta \longrightarrow 1 - \frac{0}{4} = 1 > 0.5 \therefore TRUE$

11. $AV(typ,(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
12. $AV(val,(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
13. $ratio(T_{\alpha,1}(\mathcal{R}-I),\mathcal{R}) > \theta \longrightarrow$ not applicable
14. $ratio(V_{\gamma,1}(\mathcal{R}-I),\mathcal{R}) > \theta \longrightarrow$ not applicable
15. $ratio(T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
16. $ratio(V_{\gamma,1}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
17. $ratio(V_{\gamma,1}(\mathcal{R}-I) \cap T_{\alpha,1}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable
18. $ratio(V_{\gamma,1}(\mathcal{R}-I) \cap T_{0,\beta}(\mathcal{R}-I),(\mathcal{R}-I)) > \theta \longrightarrow$ not applicable

Appendix G

Documents for external meta-evaluation of creativity evaluation methodologies

---

Example feedback sheet for external meta-evaluation of methodologies

Feedback sheets were provided to the external evaluators. These sheets reported the evaluation results obtained for their system using each evaluation methodology investigated (SPECS; Ritchie's criteria; Colton's creative tripod; survey of human opinion; and the FACE model). For each methodology, the sheets also included brief comparisons of evaluative performance across all systems.

The set of feedback below relates to the GenJam system and was sent to Al Biles. A similar set of feedback was prepared and sent to George Lewis as evaluative feedback relating to Voyager for all the methodologies listed above. Methodologies were presented in anonymised form (as shown below) so that the evaluators would be less likely to discover which was my methodology, avoiding the introduction of associated biases being introduced into their answers.

A third sheet was originally prepared for Bob Keller, relating to Impro-Visor, which was a fourth system originally included in this case study. The sheets for GenJam and Voyager also refer to Impro-Visor in their reported results and references are made to four evaluated systems, not three. We discovered during external evaluation that the musical examples used for some of the methodologies were not generated by the Impro-Visor system, but were composed by Bob Keller. This led to the removal of Impro-Visor from Case Study 1,[1] although feedback could still be obtained from Bob Keller on the methodologies (with the acknowledgement of the incorrect use of musical samples in implementing the methodologies).

Introduction

In this case study, the creativity of four musical improvisation systems has been evaluated:

- *GenJam* (Biles, 2007).
- *Impro-Visor* (Gillick et al., 2010).
- *Voyager* (Lewis, 2000).
- My own improvisation system, named *GAmprovising* for this study (Jordanous, 2010c).

[1] See Chapter 6 Section 6.2.

The research question and primary focus for the case study is:

How **creative** is this system as a musical improvisation system?

Five methodologies have been used to evaluate the creativity of your system and the other three systems. These methodologies have been anonymised for now and details of how the methodologies were implemented are deliberately brief for now, to focus on the results first. More details and references will be given separately. The findings of each creativity evaluation methodology are listed below.

[REFERENCES GIVEN HERE IN THE ORIGINAL DOCUMENTATION TO THE FOUR PAPERS CITED ABOVE]
[PAGEBREAK IN ORIGINAL DOCUMENTATION]

Creativity evaluation results: Set *CT*

Creativity is evaluated using three minimum criteria a system must meet in order to be potentially creative: *skill*, *imagination* and *appreciation*. The quantitative data are given in the form of ratings out of 10, supplied by judges with practical knowledge of both musical improvisation and computer music. The mean and standard deviation (s.d.) of these ratings are given. This methodology treats each criteria as equally important for potential creativity.

Some commentary is given on the main findings of this data. This is followed by the qualitative data collected from judges during evaluation: their comments on the system, from various perspectives. Finally, the four musical improvisation systems are compared to each other using this methodology, in terms of how creative they are and what the reasons are for this.

Quantitative feedback from the creativity evaluation

Skill: mean rating=7.3/10.0, s.d.=1.5
Imagination: mean rating=6.7/10.0, s.d.=2.5
Appreciation: mean rating=7.0/10.0, s.d.=1.0

Commentary on the quantitative data collected for GenJam GenJam demonstrated good performance on all three criteria. As the ratings showed, there could still be improvements along all three qualities, in particular its ability to demonstrate imagination. On the whole though, GenJam was found to perform consistently well in all three criteria.

Qualitative feedback from the creativity evaluation

Qualitative feedback highlighted the high skill levels of GenJam. GenJam was described as lacking inventiveness and imagination in its thought processes, though it could diverge to some limited extent. GenJam's ability to monitor itself, listen to, appreciate and think about what is happening and recognise what pitches are appropriate (or not) all contributed to boost the ratings data for this system for appreciation.

Comparison of the evaluation of creativity of the four evaluated systems

Skill: GenJam outperformed the other three systems in terms of skill, with a mean rating of 7.3. Impro-Visor and Voyager received fairly similar ratings, with means of 5.5 and 5.0, respectively. GAmprovising lagged behind the other three systems, with a mean rating of 3.3 (mean).

Imagination: As for *Skill*, GenJam scored higher ratings overall. The other three systems received similar mean ratings to each other. Voyager received a slightly higher mean rating, while GAmprovising and Impro-Visor both received the lowest mean ratings for this criterion.

Appreciation: GenJam again demonstrated considerably higher ratings than the other systems. While GenJam attracted ratings with a mean of 7.0, the other systems received much lower mean ratings, with a mean of 2.0 for GAmprovising and Voyager and 2.3 for Impro-Visor.

GenJam was considered more advanced in all three criteria, according to the judges' feedback. GenJam performed well in all three criteria, but the other systems were generally poorer at appreciation than at skill or imagination. It is hard to distinguish between Impro-Visor and Voyager, except that Voyager performed better for imaginative abilities. In general GAmprovising performed the least well of all four systems and noticeably under-performed in terms of skill in particular.

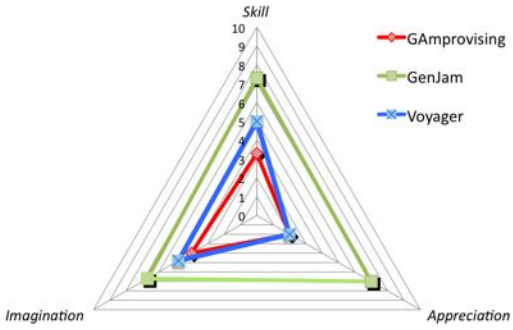


Figure G.1: Average performance (mean ratings out of 10) on the three creative criteria for *CT* (skill, imagination, appreciation)

[PAGEBREAK IN ORIGINAL DOCUMENTATION]

Creativity evaluation results: Set *SB*

Creativity is evaluated using a set of 14 components as criteria (see below), each weighted according to importance in musical improvisation creativity.

Quantitative results are given in the form of ratings out of 10, supplied by judges with practical knowledge of both musical improvisation and computer music. The mean and standard deviation (s.d.) of these ratings are given. The ratings are then weighted according to how important they have been found to be in musical improvisation creativity, in a separate study. There are 14 components, so if all components were equal then they would all be given a weighting of 7.1 (1/14). Therefore for weightings > 7.1, that component is more important than average, and vice versa for weightings < 7.1.

The weighted ratings, given last and in bold, are the most important data to consider. This data shows how highly the system was rated on that component, weighted by how important that component is. A comparison of the different systems evaluated is given in Figure G.

Some commentary is given on the main findings of this data. This is followed by the qualitative data collected from judges during evaluation: their comments on the system, from various perspectives. Finally, the four musical improvisation systems are compared to each other using this methodology, in terms of how creative they are and what the reasons are for this.

Quantitative feedback from the creativity evaluation

The 14 criteria are listed in descending order of importance for musical improvisation creativity (most important first). The ratings for GenJam and the other three systems are shown in Figure G.

1. Social Interaction and Communication
   - Communicating and promoting work to others in a persuasive, positive manner.
   - Mutual influence, feedback, sharing and collaboration between society and individual.

   Ratings: mean=8.3/10.0, s.d.=0.6, weighting=14.9, **weighted contribution to creativity=12.4**
2. Intention and Emotional Involvement
   - Personal and emotional investment, immersion, self-expression, involvement in a process.
   - Intention and desire to perform a task, a positive process giving fulfilment and enjoyment.

   Ratings: mean=6.0/10.0, s.d.=1.0, weighting=13.9, **weighted contribution to creativity=8.3**
3. Domain Competence
   - Domain-specific intelligence, knowledge, talent, skills, experience and expertise.

- Knowing a domain well enough to be equipped to recognise gaps, needs or problems that need solving and to generate, validate, develop and promote new ideas in that domain.

Ratings: mean=3.3/10.0, s.d.=1.5 , weighting=12.5, **weighted contribution to creativity=9.2**

4. Active Involvement and Persistence
   - Being actively involved; reacting to and having a deliberate effect on a process.
   - The tenacity to persist with a process throughout, even at problematic points.

   Ratings: mean=7.0/10.0, s.d.=1.0, weighting=7.8, **weighted contribution to creativity=5.5**
5. Variety, Divergence and Experimentation
   - Generating a variety of different ideas to compare and choose from, with the flexibility to be open to several perspectives and to experiment with different options without bias.
   - Multi-tasking during a process.

   Ratings: mean=6.7/10.0, s.d.=2.5, weighting=7.1, **weighted contribution to creativity=4.7**
6. Dealing with Uncertainty
   - Coping with incomplete, missing, inconsistent, uncertain and/or ambiguous information. Element of risk and chance, with no guarantee that problems can or will be resolved.
   - Not relying on every step of the process to be specified in detail; perhaps even avoiding routine or pre-existing methods and solutions.

   Ratings: mean=7.0/10.0, s.d.=3.0, weighting=6.4, **weighted contribution to creativity=4.5**
7. Originality
   - Novelty and originality - a new product, or doing something in a new way, or seeing new links and relations between previously unassociated concepts.
   - Results that are unpredictable, unexpected, surprising, unusual, out of the ordinary.

   Ratings: mean=7.0/10.0, s.d.=1.7, weighting=5.8, **weighted contribution to creativity=4.1**
8. Independence and Freedom
   - Working independently with autonomy over actions and decisions.
   - Freedom to work without being bound to existing solutions, processes or biases; perhaps challenging cultural/domain norms.

   Ratings: mean=7.0/10.0, s.d.=2.6, weighting=5.4, **weighted contribution to creativity=3.8**
9. Progression and Development
   - Movement, advancement, evolution and development during a process.
   - Whilst progress may or may not be linear, and an actual end goal may be only loosely specified (if at all), the entire process should represent some developmental progression in a particular domain or task.

   Ratings: mean=7.5/10.0, s.d.=1.3, weighting=5.4, **weighted contribution to creativity=4.1**
10. Spontaneity / Subconscious Processing
    - No need to be in control of the whole process; activities and thoughts may inform a process subconsciously without being fully accessible for conscious analysis.
    - Being able to react quickly and spontaneously during a process when appropriate, without needing to spend time thinking about options too much.

    Ratings: mean=5.5/10.0, s.d.=3.1, weighting=5.4, **weighted contribution to creativity=3.0**
11. Thinking and Evaluation
    - Consciously evaluating several options to recognise potential value in each and identify the best option, using reasoning and good judgment.
    - Proactively selecting a decided choice from possible options, without allowing the process to stagnate under indecision.

    Ratings: mean=7.0/10.0, s.d.=1.0, weighting=5.1, **weighted contribution to creativity=3.6**

12. Value
    - Making a useful contribution that is valued by others and recognised as an influential achievement; perceived as special; 'not just something anybody would have done'.
    - End product is relevant and appropriate to the domain being worked in.

    Ratings: mean=6.7/10.0, s.d.=2.3, weighting=5.1, **weighted contribution to creativity=3.4**
13. Generation of Results
    - Working towards some end target, or goal, or result.
    - Producing something (tangible or intangible) that previously did not exist.

    Ratings: mean=8.3/10.0, s.d.=1.5, weighting=3.7, **weighted contribution to creativity=3.1**
14. General Intellect
    - General intelligence and intellectual ability.
    - Flexible and adaptable mental capacity.

    Ratings: mean=6.8/10.0, s.d.=2.5, weighting=1.4, **weighted contribution to creativity=1.0**

Commentary on the quantitative data collected for GenJam Looking at GenJam's mean ratings for creativity, before they are weighted:

- Relative strengths: *Social Interaction and Communication* (8.3/10.0), *Creation of Results* (8.3/10.0).
- Relative weaknesses: *Spontaneity and Subconscious Processing* (5.5/10.0), *Intention and Emotional Involvement* (6.0/10.0).

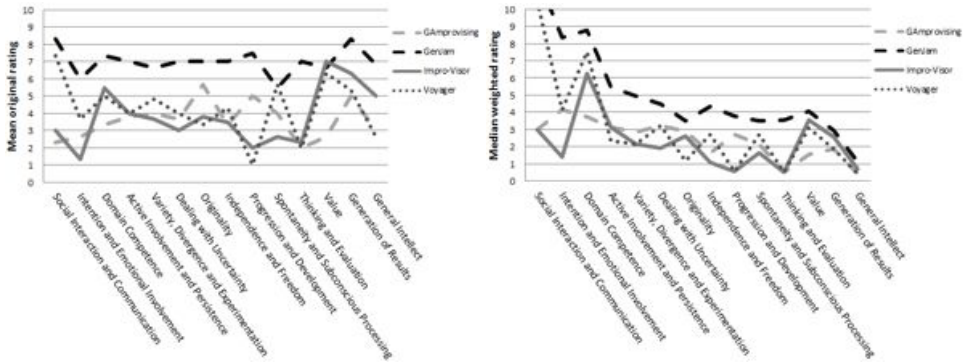
GenJam's weighted ratings were highest on average for all but two components: *Spontaneity and Subconscious Processing* and *Value*, two relatively unimportant components in this domain. It scored particularly well for *Social Interaction and Communication* (12.4). The areas where GenJam could make most improvement in terms of weighted creativity ratings are *Intention and Emotional Involvement* (potential gain of 5.6), *Domain Competence* (potential gain of 3.3) and to some extent, *Social Interaction and Communication* (potential gain of 2.5). Devoting attention to improving the top three components would be most effective for increasing the creativity of GenJam.

Qualitative feedback from the creativity evaluation

- *Social Interaction and Communication* in GenJam was praised highly, particularly in how it responds to what it hears.
- While attracting reasonable ratings for this component, GenJam was felt to reflect the *Intention and Emotional Involvement* of the human player rather than those inherent to the system.
- *Domain Competence* was also highly praised, with GenJam seeing as possessing a lot of relevant musical knowledge.
- *Active Involvement and Persistence* was seen as good though there was doubt as to whether it would become aware of problems occurring.
- GenJam was seen to be able to diverge quite a lot (*Variety, Divergence and Experimentation*) but some judges commented that its variation was limited by its programming.
- Judges reported difficulty with rating *Dealing with Uncertainty* due to lack of examples in the information they had been given. Ratings varied because of this.
- The level of *Originality* was considered to be fairly high by one judge, but other judges raised questions as to the extent to which GenJam could be original.
- *Independence and Freedom* in GenJam was seen as high in the autonomous version (Biles, 2007) although it needs training input beforehand.
- *Progression and Development* was noted by all judges in the context of the solo and overall, due to the use of genetic algorithm techniques.
- GenJam was seen to be fairly spontaneous within its programmed limits.
- *Thinking and Evaluation* was seen as being the user's responsibility, not the systems, and that the system could perform better for this, though it was able to constantly monitor what it was doing and behave rationally.

- *Value* was generally perceived as high, though one judge quickly found the solos to become boring. One judge was interested in playing with the system to practice improvising.
- The ability to generate end products was praised for *Creation of Results*.
- To some degree GenJam demonstrated *General Intellect*, through awareness of taste and alternative modes of thought. One judge commented: 'it could be nice if it decides what version to use, out of the different versions of the system by involving all different algorithms'. One judge was unconvinced by GenJam's intelligence, unlike the other judges.

Comparison of the evaluation of creativity of the four evaluated systems



(a) Unweighted Mean Ratings. (b) Weighted Mean Ratings.

Figure G.2: Mean averages for all judges' evaluation ratings for Results Set *SB* (before and after weighting by component importance) for the four systems.

Overall, GenJam was found to be the most creative of the four systems. Other systems perform better than GenJam in less important components (Voyager for *Spontaneity and Subconscious Processing*, Impro-Visor for *Value*); however GenJam's higher average ratings for nearly all components, and its particular strengths in components of greatest importance such as *Social Interaction and Communication*, demonstrate its greater musical improvisational creativity.

To improve on musical improvisational creativity, greatest gains can be made in all four systems by concentrating efforts first on more important components such as *Social Interaction and Communication*, *Intention and Emotional Involvement* and *Domain Competence* and by taking the judges' qualitative feedback into account. Specifically, GenJam's interactive abilities were praised alongside its ability to create results, whilst it could improve on its spontaneity and originality.

[PAGEBREAK IN ORIGINAL DOCUMENTATION]

Creativity evaluation results: Set *OS*

In this methodology, people were asked to rate how creative they thought each of the four musical improvisation systems were, on the following scale:

5 = Completely creative	3 = Quite creative	1 = Not at all creative
4 = Very creative	2 = A little creative but not very	

The participants also rated how confident they were about each answer they gave, on the following scale:

5 = Very Confident
4 = Confident
3 = Neutral
2 = Unconfident
1 = Very Unconfident

Quantitative feedback from the creativity evaluation
GenJam creativity ratings and participants' confidence in their answers

- Mean creativity rating = 3.0, corresponding to 'Quite creative' on the above creativity scale if rounded to 0 decimal places. Standard deviation = 0.9.
- Confidence in answers rated at a mean of 3.7, corresponding to 'Confident' on the above creativity scale if rounded to 0 decimal places. Standard deviation = 0.9.

The creativity ratings are illustrated in Figure G.3.

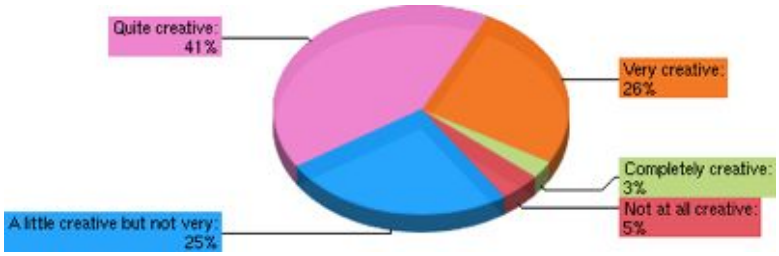


Figure G.3: Ratings of creativity for GenJam with method *OS*.

Qualitative feedback from the creativity evaluation
Qualitative feedback on systems' creativity This Section summarises the qualitative feedback given about each system. Positive comments are marked as '+', negative as '-' and neutral as 'n'. The number of comments making a particular point is given in square brackets after the description of that point. Some comments covered more than one point.[2]

Five participants had heard of GenJam before, in academic work (2 participants), a talk and demonstration (1 participant), seeing some videos in the past (1 participant) and through a musical colleague's use of a similar program (1 participant).[3]
64 people chose to comment further on the creativity of GenJam (receiving the most extra comments of all four systems), with comments averaging 176 characters each (s.d. 175).

- \- Some commented on the lack of direction and progression in GenJam's solos [2].
- \+ Others praised GenJam for its cohesion, melodic direction or natural flow [3].
- \+ Positive comments were received such as 'good idea' or 'like this, very good' [5].
- \- Others drew comparisons between GenJam's solos and 'muzak'/'lift music'[4] or computer generated music, with a lack of 'feeling' [10].
- \+ In contrast, GenJam's soloing was described as sounding more natural, authentic or human by some people [5].
- \- Some people found that the solos sounded good at first but then became too smooth-sounding, unsurprising and boring, like an improviser who had run out of ideas or an improviser taking

[2] A general point is that the comments seemed to centre around the systems' musicality and competence rather than the systems' creativity.

[3] This last participant's exact response was 'Computer programmer/musician where I work has done work with this, I think, and incorporated it into performances.', but as the GenJam system is not available for use by other musicians, it is likely that the participant was either a colleague of Al Biles or was not referring to GenJam but to a similar system.

[4] The terms 'lift music'/'muzak' are negative terms referring to generated music in an easy listening style, often artificially generated and/or played in a repetitive loop, which is often used for unobtrusive, unnoticeable background music.

too conservative an approach or lacking in flair [12]. Others found the system uninteresting or unmusical [2]. Questions were raised as to how much of the material in GenJam's solos was taken from pre-existing solos or repetitive [8] or whether GenJam thought about what it did and learned [1].

- Some commented that GenJam was poor at using silence in solos, a mistake often made by beginner soloists [2].

n GenJam's interplay with the trumpet player was praised by some [6] although it was sometimes criticised for not being sensitive enough to the human's playing [4].

- Comments were made about occasionally odd harmonic or rhythmic choices in GenJam's playing [5].

n Of the three pieces, some singled out the third extract (from 'The Rake') for particular positive mention [6] although although it also received some negative comments such as 'hilariously awful' [2] The second track divided opinion between positive [2] and negative [4] and the first track was given mostly negative specific comments [6] with only one positive comment [1].

Comparison of the evaluation of creativity of the four evaluated systems

Table G.1 shows how the creativity ratings compare across the four musical improvisation systems evaluated in this case study. The levels of confidence expressed by participants in their answers are also compared.

The final column in Table G.1 shows how participants ranked the systems overall in terms of creativity. Items ranked first by a participant are given 4 points, second 3 points, third 2 points, and fourth 1 point. All points are summed together for the overall ranking points score across all participants.

Table G.1: Comparison of the evaluation of creativity of the four evaluated systems with method *OS*.

System	Mean / s.d. creativity	Creativity (mean rounded to 0 dp)	Mean / s.d. Confidence	Confidence (mean rounded to 0 dp)	Ranking (points)
GenJam	3.0 / 0.9	Quite creative	3.7 / 0.9	Confident	1st (308)
Impro-Visor	2.9 / 0.9	Quite creative	3.8 / 0.9	Confident	2nd (301)
Voyager	2.7 / 0.9	Quite creative	3.5 / 1.0	Confident	3rd (250)
GAmprovising	2.5 / 0.9	Quite creative	3.5 / 0.9	Confident	4th (231)

As can be seen from Table G.1, the labels did not differentiate between systems' creativity, nor the participants' confidence in determining a creativity label. The numeric scores fared better, finding GenJam to be most creative (by a small margin). In the overall rankings, GenJam and Impro-Visor were rated very close to each other, though GenJam received 7 points more than Impro-Visor. Voyager came behind these two systems, followed by GAmprovising, with 19 points less than Voyager. The overall ranking scores, though, were all quite similar, rather than there being sizeable differences between the ranking scores, showing that there was no clear distinguishable agreement in participants' opinions. Confidence scores were slightly higher for Impro-Visor and GenJam than for Voyager and GAmprovising, but the difference is small.

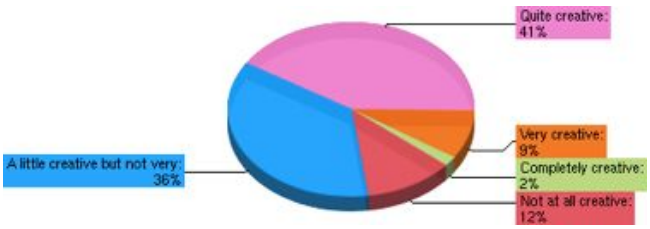
Figure G.4 show the variance in opinions across each system per rating. This figure shows that apart from it being slightly less likely for GenJam to receive ratings of 'A little creative but not very' and similarly for GAmprovising to receive ratings of 'Very creative', there is little noticeable difference between the distribution of ratings for each system.

[PAGEBREAK IN ORIGINAL DOCUMENTATION]

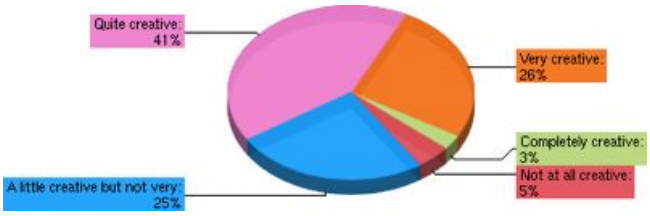
Creativity evaluation results: Set *FD*

Creativity is evaluated as the ability to produce creative acts, where creative acts are described by the attribution (or not) of four discrete criteria, representing measures/measurement methods:

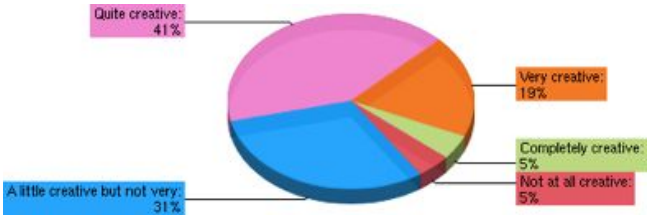
- E Example: example products of the system
- C Concept: a concept of a creative process taking place within the system, receiving input and generating output
- A Aesthetic: aesthetic measures employed by the system for self-evaluation
- F Frame: contextual commentary (expressed in natural language), framing what the system has done.



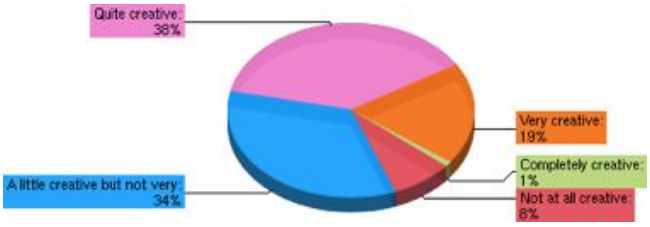
(a) GAmprovising.



(b) GenJam.



(c) Impro-Visor.



(d) Voyager.

Figure G.4: Ratings of creativity per system, showing the spread of opinion on each system's creativity with method *OS*.

For each of the four criteria, the *FD* method adds the superscript *g* if the system is able to *generate* items representing that criterion and the superscript *p* if the system has *processes* to generate representations or measures for that item, during system operation. (For example, E^g represents the ability to generate example artefacts, while E^p represents the ability to generate a method for generating example artefacts. A^g represents the use of aesthetic measures by the system, while A^p represents the ability to generate new aesthetic measures which can be used while the system is operating.)

Quantitative feedback from the creativity evaluation
No quantitative feedback is collected when this method is used to evaluate the creativity of a computational system.

Qualitative feedback from the creativity evaluation
GenJam's creativity can be represented in method *FD* by the ability to generate creative acts following the description: $< A^g, C^g, E^g >$.

It does not provide natural language descriptions of what it has done so the F(rame) criterion is not present. In its more recent incarnation, GenJam does use its own self-evaluation (based on pre-existing melodies that have been judged to be good), though it does not generate novel aesthetic measures, so can be attributed the A^g criterion. It employs a creative process and generates example outputs, so can also be attributed with C^g and E^g, though it performs no meta-generation of methods to create new products or new creative concepts for artefact generation.

Comparison of the evaluation of creativity of the four evaluated systems

Table G.2: Comparison of the evaluation of creativity of the four evaluated systems using method *FD*

/ System	Evaluation result
Impro-Visor	$< A^p, A^g, C^g, E^p, E^g >$
GAmprovising	$< A^g, C^g, E^p, E^g >$
GenJam	$< A^g, C^g, E^g >$
Voyager	$< C^g, E^g >$

Looking at Table G.2, no system produces creative activity that satisfies all four of the criteria, as none provide framing information, but GenJam, Impro-Visor and GAmprovising all satisfy three criteria. Voyager does not use or generate aesthetic measures, hence only satisfies two criteria. For further differentiation of GenJam, Impro-Visor and GAmprovising, GenJam only meets the ^{g} part of each of the three criteria as it generates items to satisfy the criteria but not the meta-generation of methods. GAmprovising does slightly better, with the generation of new methods for generating new artefacts (E^p), while Impro-Visor goes one better by also being able to generate new aesthetic measures (A^p). Thus from method *FD* we can conclude that Impro-Visor is evaluated as most creative, followed by GAmprovising, then GenJam, and lastly Voyager.

[PAGEBREAK IN ORIGINAL DOCUMENTATION]

Creativity evaluation results: Set *RC*

Creativity is evaluated using a set of 18 empirical criteria (see below). These criteria are based upon ratings of how typical and how valuable the improvisations were judged to be, as well as comparisons between the output of the system and any input the system uses as inspiration.

Quantitative feedback from the creativity evaluation

The formal set-theoretic definitions of the 18 criteria can be found in the original paper. Here, the criteria are deliberately presented informally, for a more immediate initial understanding of the criteria.

1. On average, the system should produce suitably typical output: **TRUE**.
2. A decent proportion of the output should be suitably typical: **TRUE**.
3. On average, the system should produce highly valued output: **TRUE**.
4. A decent proportion of the output should be highly valued: **TRUE**.
5. A decent proportion of the output should be both suitably typical and highly valued: **TRUE**.
6. A decent proportion of the output is suitably atypical and highly valued: **FALSE**.
7. A decent proportion of the atypical output is highly valued: ***undefined, N/A***.
8. A decent proportion of the valuable output is suitably atypical: **FALSE**.
9. The system can replicate many of the example artefacts that guided construction of the system (the *inspiring set*): **FALSE**.
10. Much of the output of the system is not in the inspiring set, so is novel to the system: **TRUE**.
11. Novel output of the system (i.e. not in the inspiring set) should be suitably typical: **TRUE**. *In calculations, this reduced to be the same criterion as Criterion* 1.
12. Novel output of the system (i.e. not in the inspiring set) should be highly valued: **TRUE**. *Same as Criterion* 3.
13. A decent proportion of the output should be suitably typical items that are novel: **TRUE**. *Same as Criterion* 2.
14. A decent proportion of the output should be highly valued items that are novel: **TRUE**. *Same as Criterion* 4.
15. A decent proportion of the novel output of the system should be suitably typical: **TRUE**. *Same as Criteria* 2 and 13.
16. A decent proportion of the novel output of the system should be highly valued: **TRUE**. *Same as Criteria* 4 and 14.
17. A decent proportion of the novel output of the system should be suitably typical and highly valued: **TRUE**.
18. A decent proportion of the novel output of the system should be suitably atypical and highly valued: **FALSE**. *Same as Criterion* 6.

Summary of the quantitative data collected for GenJam

- 13 / 18 criteria TRUE (Criteria 1-5, 10a-17)
- 7 / 11 distinct criteria TRUE (Criteria 1-5, 10a, 17)
- 4 / 18 criteria FALSE (Criteria 6, 8a, 9, 18)
- 3 / 11 distinct criteria FALSE (Criteria 6, 8a, 9)
- 1 / 18 criteria not applicable (Criterion 7)

Qualitative feedback from the creativity evaluation

No qualitative feedback is collected when this method is used to evaluate the creativity of a computational system.

Comparison of the evaluation of creativity of the four evaluated systems

From the results of applying this method, shown in Table G.3, GenJam emerged most creative, as it satisfied the most distinct criteria (3 more than the next best systems, Impro-Visor and Voyager) and falsified the least criteria (two or three less than the other systems). There was also only one undefined criterion, which was the same as for Impro-Visor and one less than the other two systems. Impro-Visor and Voyager were close in terms of relative creativity, both satisfying 4 out of 11 distinct criteria. For Impro-Visor 6 distinct criteria were evaluated as FALSE and 1 was evaluated as undefined, whereas

for Voyager 5 distinct criteria were FALSE and 2 were inapplicable. GAmprovising was found to be the least creative system, with 6 distinct criteria being FALSE, 2 inapplicable and only 3 distinct criteria being satisfied as TRUE.

[END OF FEEDBACK SHEET]

Provided explanations of each system

For each methodology evaluated by the external evaluators, details were given of the methodology and links were provided to papers and other relevant explanatory resources where available. Each methodology description was accompanied with a proviso:

> *(* I do not expect you to have a comprehensive and detailed understanding of how the methodology works and how it has been applied, and you are not required to look at the links mentioned in the optional Further Information. I hope, though, that the above explanation gives you enough information on which to base your answers to the following questions.).*

This proviso was intended to make it clearer to the evaluators to what extent I expected them to understand the methodology, and hopefully to reduce any perceived pressure on the evaluators in terms of how well they were expected to know each methodology. Links were provided for further information if this was desired, but further research beyond the description provided was optional.

Description of the FACE model (Presented as Method 'FD' in the feedback sheet)

> 'The results of this methodology are based on identifying creative acts performed by a computational system, along four criteria (Framing information, Aesthetic measures, use of Concept(s) and Examples). This is the FACE descriptive model of creative acts, as proposed by Colton, Charnley and Pease (2011).
>
> Processes used in the methodology: Each system was studied to see if it demonstrated each of the four criteria - either by generating items that satisfied the criterion (superscript g in the results) and/or by meta-generation - generating methods to generate items satisfying the criterion (superscript p).
>
> Model of creativity: The interpretation of what it meant for a system to be creative was that it produced creative acts, measured along the Frame, Aesthetics, Concepts and Examples model described above.
>
> Further information [optional]: If you are interested in finding out more about this evaluation methodology, the FACE model of creative acts is described in: Colton, S., Charnley, J., & Pease, A. (2011). Computational Creativity Theory: The FACE and IDEA descriptive models. In Proceedings of the 2nd International Conference on Computational Creativity, pp. 90-95 Mexico City, Mexico.' *[online link provided to the paper (Colton et al., 2011)]*

Table G.3: Comparison of the evaluation of creativity of the four evaluated systems with method *RC*.

System	# criteria (/18)	# distinct criteria (/11)
GenJam	13 T, 4 F, 1 n/a	7 T, 3 F, 1 n/a
Voyager	8 T, 8 F, 2 n/a	4 T, 5 F, 2 n/a
Impro-Visor	8 T, 9 F, 1 n/a	4 T, 6 F, 1 n/a
GAmprovising	5 T, 11 F, 2 n/a	3 T, 6 F, 2 n/a

Description of the SPECS methodology (Presented as Method 'SB' in the feedback sheet)

'The results of this methodology are based on determining and clearly stating what it means for musical improvisation systems to be creative (in general and specifically for musical improvisation), then identifying and carrying out tests corresponding to these determinations. This is the SPECS methodology, as proposed by Jordanous (2012).[5]

Processes used in the methodology: A general model of creativity was analysed in the context of what is important for creativity in musical improvisation (see the Model of creativity section, next). Then, expert judges (with knowledge of musical improvisation and computer music) were given information on the systems (papers, videos, recordings, website links) and given 60 minutes (with internet access) to research and learn about the systems. They were then asked to rate the system on each of the 14 criteria, out of 10. These ratings were then weighted according to the importance of that criterion. Each system was evaluated by three judges, who each evaluated two systems in a session that lasted two hours. In total, six judges were involved.

Model of creativity: The interpretation of what it meant for a system to be creative was based around taking a model of creativity as a general concept and fine tuning this general model around how creativity is displayed in musical improvisation. In a recent empirical study (reported in Jordanous 2012), 14 criteria were identified that collectively acted as "building blocks" for creativity. These criteria were used as the general model of creativity. To fine tune this model specifically towards musical improvisation creativity, 33 people were questioned on various aspects of musical improvisation and their replies were analysed to compare how often each of the 14 creativity criteria were mentioned in responses. (This study is also reported in Jordanous 2012). From this analysis, weights for each of the 14 criteria were calculated.

Further information [optional]: If you are interested in finding out more about this evaluation methodology, the SPECS creativity evaluation methodology and corresponding set of creativity components is described in: Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. Cognitive Computation, 4(3), 246279.'*[online link provided to the paper (Jordanous, 2012)]*

Description of the surveys of human opinion (Presented as Method 'OS' in the feedback sheet)

'The results of this methodology are based on asking people their opinions on how creative they thought the systems were.

Processes used in the methodology: An opinion survey was carried out across 111 participants, to capture their opinions on how creative they thought the various systems were. For each system, participants were given a short description of the system and how it works, and three short (30 seconds) recordings as examples of improvisations by the system. Based on this information, they were asked how creative they thought that system is.

Model of creativity: The interpretation of what it meant for a system to be creative was left open to the participants' interpretations (deliberately), to see how participants saw the creativity of the systems rather than to impose a particular definition of creativity.

Further information [optional]: If you are interested in finding out more about this evaluation methodology, the opinion survey can be viewed here. The questions specifically on Voyager are questions 13-17. For GenJam, look at questions 18-22, and for Impro-Visor, look at questions 28-32. (Please note that this part of the survey was presented in randomised order, so this ordering is not necessarily the order in which survey participants saw the pages.'[link provided to online version of the survey for Case Study 1's opinion survey (Chapter 8 Section 8.1.1)]

[5] At this point, anonymity was no longer maintained, hence there was a possibility that bias could have been introduced by the evaluator knowing that this was the methodology I was proposing. To have maintained anonymity at this point would however have reduced the information given on the methodology such as links to papers. In particular, if SPECS was the only methodology presented without links to papers, this may have aroused suspicions. A literature search for the SPECS methodology would soon have revealed Jordanous (2012), showing the methodology to be mine.

Description of Colton's creative tripod methodology (Presented as Method 'CT' in the feedback sheet)

'The results of this methodology are based on the "creative tripod" model, as proposed by Colton (2008). Using the creative tripod, potential candidates for creative systems can be identified by looking at three aspects of the system: its ability to demonstrate skill, imagination and appreciation.

Processes used in the methodology: Expert judges (with knowledge of musical improvisation and computer music) were given information on the systems (papers, videos, recordings, website links) and given 60 minutes (with internet access) to research and learn about the systems. They were then asked to rate the system on various criteria, out of 10, including the system's skill, appreciation and imagination. Each system was evaluated by three judges, who each evaluated two systems in a session that lasted two hours. In total, six judges were involved.

Model of creativity: The interpretation of what it meant for a system to be considered potentially creative was based on checking if the system demonstrated skill, imagination and appreciation. If a system does demonstrate all three of these "creative tripod" qualities, it could potentially be a creative system.

Further information [optional]: If you are interested in finding out more about this evaluation methodology, the creative tripod methodology for evaluation of computational creativity is described in: Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In Proceedings of AAAI Symposium on Creative Systems, pp. 1420.'*[online link provided to the paper (Colton, 2008b)]*

Description of Ritchie's criteria methodology (Presented as Method 'RC' in the feedback sheet)

'The results of this methodology are based on assessing 18 empirical criteria for creativity, as proposed by Ritchie (2007).

Processes used in the methodology: The 18 criteria for creativity concentrate on the typicality and value of the improvisations produced by the system. Each criterion represents a particular permutation of one or both of these aspects. Some of the criteria also measured novelty of the system's products against an inspiring set (existing examples as inspirational material in constructing improvisations). To obtain ratings of typicality and value for products of each system, a survey was carried out amongst 89 participants. Three 30-second excerpts were chosen for each system (a total of 12 excerpts in total for participants to listen to). Participants rated each of the 12 excerpts for typicality and value as a musical improvisation. This survey can be viewed here. For the Voyager excerpts, look at the three tracks for questions 12-23. For GenJam, look at the three tracks for questions 24-35, and for Impro-Visor, look at the three tracks for questions 48-59. (Please note that this part of the survey was presented in randomised order, so this ordering is not necessarily the order in which survey participants saw the tracks.

Model of creativity: The interpretation of what it meant for a system to be creative was based on how many of the 18 criteria proposed by Ritchie (2007) were satisfied. Following the lead of previous evaluations performed using this methodology, the set of 18 criteria were treated as equal to each other, rather than being weighted individually. Some of the criteria turned out to be equivalent to each other in calculations, due to the system making no use of an inspiring set (existing examples as inspirational material in constructing improvisations). In these cases, the two equivalent criteria were considered as one distinct criterion, when comparing different systems.

Further information [optional]: If you are interested in finding out more about this evaluation methodology, Ritchie's empirical criteria for creativity attribution are described in: Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. Minds and Machines, 17, 67 99.' *[online link provided to the paper (Ritchie, 2007)]*

'Details of the specific calculations for this implementation of Ritchie's criteria are temporarily available here (and will be available in my thesis at a later date).' *[online link provided to a document presenting these calculations]*

Appendix H

Meta-evaluation of SPECS using SPECS

Meta-evaluation is seldom addressed in the literature, with some exceptions (Pease et al., 2001; Machado et al., 2003).[1] To some extent, SPECS already contains some meta-evaluation when implemented using the 14 components derived in Chapter 4. One of the components, *Thinking and Evaluation*, evaluates how the creative system uses evaluation. Advocating similar points to those in Pease et al. (2001), Chapters 8 and 9 examine how accurate, useful and implementable the SPECS methodology is, in comparison to other methodological tools and in comparison to human assessments of creativity. One question has however not yet been addressed: *How creative is SPECS?*

Creativity of the SPECS methodology itself is not a specific requirement of this thesis work, but as this thesis focuses on the study of creativity, this question becomes an intriguing one to consider as an aside.[2] Section 10.2.4 outlines how a creative system need not necessarily be computationally based. On this premise, it is feasible to treat a methodology as a potentially creative system. Given that the creativity of the SPECS methodology is not a priority, a descriptive evaluation of SPECS guided by the Chapter 4 components is given, rather than a full application of SPECS being performed. This gives some intriguing feedback about SPECS from an alternative perspective:

1. *Active Involvement and Persistence* SPECS methodology requires active involvement in implementation on the part of the user, although it does not particularly demonstrate any persistence if problems are encountered.
2. *Generation of Results* SPECS does indeed produce evaluative results through Step 3, specifically targeted towards the standards identified in Step 2.
3. *Dealing with Uncertainty* Case Study 2 has shown that if information is missing, SPECS can still be implemented, although its accuracy is limited to the information available and is weakened if important information is unavailable (as most methods of evaluation would be). If information is unknown on how to define creativity in a particular domain then SPECS requires

[1] Machado et al. (2003) incorporated a self-evaluation facility in their Artificial Arts Critics (AACs). Feedback from the environment where AACs were situated was used to refine the 'Dynamic Evaluation' stage of evaluation in real time, though Machado et al. do not give clear details of this. Pease et al. (2001) included discussions of how to measure the success of their assortment of tests, setting out two criteria: how the tests reflect human evaluations of creativity, to be tested empirically, and how applicable the findings are in improving future work on modelling creativity. Sadly such testing does not appear to have been performed in subsequent work by Pease et al.

[2] Hence its inclusion as an appendix rather than in the main text.

the researcher to carry out further work to use the methodology, although this thesis does offer the researcher information towards a general definition of creativity.

4. *Domain Competence* Step 1b requires an understanding of specific domain priorities for creativity and knowledge of what contributes to creativity in that domain. In considering competence in the domain of how to evaluate creativity, the methodology provides guidance and heuristics but not detailed instructions.
5. *General Intellect* Step 1a requires an understanding of what contributes to general creativity; this thesis provides assistance to that in one way. The methodology also requires user contributions in extracting standards to be tested from definitions, identifying how to test these standards, performing the tests and analysing the results. The methodology itself does not have general intellect of its own in this respect.
6. *Independence and Freedom* Without a user to perform the evaluation, the SPECS methodology is of limited use. In the context of the independence of the methodology, it requires additional information in order to be usable (e.g. supplying a creativity definition for Step 1) and cannot be used independently of all other information. SPECS does however work independently of any other creativity method, though there is freedom to adapt SPECS, so other methods can be incorporated if the user chooses.
7. *Intention and Emotional Involvement* It cannot be said that SPECS demonstrates any intention or emotional involvement with the process of evaluating creativity.
8. *Originality* Being a novel contribution, no similar methodology to SPECS exists. The closest in application is probably the creative tripod framework (Colton, 2008b), however SPECS approaches the task of evaluation very differently to Colton.
9. *Progression and Development* As discussed throughout this thesis, SPECS easily allows for adaptation to progress and development of creativity. Within the SPECS steps, a definition of creativity is developed into standards for creativity which then facilitate progress towards evaluative tests. The basic framework behind SPECS is fixed, though, although it is reasonable to expect that future research into SPECS could develop the framework if needed.
10. *Social Interaction and Communication* Implementing SPECS involves interaction with a user who performs each step of SPECS. Within the methodology, each step interacts with the previous step by taking information derived in that previous step. SPECS also requires that the user interacts with that domain in deriving a specific definition of creativity.
11. *Spontaneity and Subconscious Processing* Being formed of higher level instructions, it could be argued that SPECS demonstrates a subconscious level of processing, though it could also be argued that this level of abstraction is taken on by the user rather than the methodology itself. In other regards it is difficult to attribute spontaneity or subconscious processing to SPECS.
12. *Thinking and Evaluation* Evaluation is what SPECS is entirely focused towards. The methodology itself does not perform self-evaluation and it does not reason over the results.
13. *Value* Throughout this thesis SPECS is shown to have value in several different ways, from providing formative feedback towards improving systems, to gaining a greater understanding of creativity, via tracking progress in computational creativity research achievements.
14. *Variety, Divergence and Experimentation* Within SPECS there is much scope for experimentation and trying a variety of tests out, although the user is restricted to working with the methodological steps given and the information supplied from different tests. The key to this component is that the user is free to experiment, as long as the processes they are using are clearly stated, for transparency of research methods.