Cognitive Computation manuscript No.
(will be inserted by the editor)

A Standardised Procedure for Evaluating Creative Systems

Computational creativity evaluation based on what it is to be creative

Anna Jordanous

Received: date / Accepted: date

Abstract Computational creativity is a flourishing research area, with a variety of creative systems being produced and developed. Creativity evaluation has not kept pace with system development with an evident lack of systematic evaluation of the creativity of these systems in the literature. This is partially due to difficulties in defining what it means for a computer to be creative; indeed, there is no consensus on this for human creativity, let alone its computational equivalent. This paper proposes a Standardised Procedure for Evaluating Creative Systems (SPECS). SPECS is a three-step process: stating what it means for a particular computational system to be creative, deriving and performing tests based on these statements. To assist this process, the paper offers a collection of key components of creativity, identified empirically from discussions of human and computational creativity. Using this approach, the SPECS methodology is demonstrated through a comparative case study evaluating computational creativity systems that improvise music.

Keywords Computational creativity · Creativity Evaluation · Cognitively-inspired evaluation · Methodology · Creativity

1 Introduction

How should we evaluate the creativity of computational creativity systems?

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. The question of computational creativity evaluation itself is non-trivial and can be considered one of the 'Grand Challenges' of computational creativity research [1].

The research in this paper is offered as a practical methodological contribution towards addressing this Grand Challenge. As will be described during the paper, the methodological approach taken is influenced by key findings in a wide range of creativity research, both computational and human.

Section 2 reviews previous work done in the area of evaluating computational creativity, considering how the development of computational creativity as a research field has developed the current culture of creativity evaluation and examining three existing frameworks for creativity evaluation [2–4]. Evaluative practice in computational creativity is shown to be in danger of falling into a 'methodological malaise' similar to that which had been identified as a potential hazard for artificial intelligence [5] and music informatics [6] research. This malaise occurs through lack of rigour, standardisation and systematic approach, and often through lack of any evaluation whatsoever to justify claims of systems being creative. In the context of creativity evaluation, this is probably at least in part due to problems in defining creativity. Clarification of the meaning of creativity is needed but several issues exist that hinder such clarification. In Section 3 existing definitions of creativity are examined, including dictionary, research and legal definitions, and different perspectives on creativity are explored. Section 4 derives an empirical definition of creativity from key contributions to the academic literature on creativity (both human and computational varieties), 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

Anna Jordanous
Centre for e-Research, King's College London, UK
Tel.: +44-20-7848-1988
E-mail: anna.jordanous@kcl.ac.uk

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.

Motivated and guided by the above considerations, and incorporating the components derived in Section 4, the three-step SPECS methodology for evaluation of computational creativity is presented in Section 5. This methodology is derived from observations on current evaluative practice and is based on directly tackling issues in creativity evaluation that currently hamper evaluation from being carried out. SPECS requires that the computational creativity researcher adopt and clearly state a definition of creativity to evaluate their system's creativity by,[1] then employs this definition as a set of standards for evaluation, each of which are to be tested by appropriate tests. Each step of the SPECS methodology is discussed and practical concerns relevant to each step are considered within those discussions.

A case study is presented demonstrating a practical application of the SPECS methodology. SPECS is used to evaluate in detail the creativity of four musical improvisation systems: GAmprovising [7], GenJam [8], Impro-Visor [9] and Voyager [10]. The results show that GenJam is perceived as most creative overall. Perhaps more importantly, though, formative feedback is gathered for each system to highlight their strengths and weaknesses in exhibiting creativity.

The findings of the SPECS evaluation in the case study are compared and contrasted with findings from various other methods of evaluation: surveys of human evaluation of the systems' creativity and the application of two existing evaluation methodologies [2,3]. Comparative results were fairly consistent across different methodologies in terms of summative evaluation, but the amount of formative feedback obtained from SPECS exceeded that of the other methods. Additionally, the task of using SPECS for creativity 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.

Some points with SPECS arose during its implementation and evaluation, where further development work could prove fruitful in a few areas outlined below. The SPECS methodology has however been shown to offer many benefits, as an evaluative tool which is offered to the computational creativity research community. Some reflections are made on the future use, longevity and contributions of the SPECS methodology.

[1] The components of creativity are strongly recommended as a basis for this definition.

2 Background

2.1 The role of evaluation and why it is needed

Evaluation highlights where progress is being made and how the evaluated item can be improved upon. Research progress can be demonstrated and tracked and we can learn from achievements and weak points of a system.

The evaluation process should be clearly stated, to be transparent and repeatable [11,5,6]. Evaluations can then be applied to related systems for more comparable and consistent evaluation results, and evaluation decisions can be critiqued and/or learned from. In artificial intelligence, to which computational creativity is closely related, Bundy describes how confused research aims, unclear evaluation criteria and poor continuity between projects reveal an underlying 'methodological malaise' [5, pp. 215-216]. Pearce et al. echo this,[2] noting that poorly-evaluated research 'has little practical or theoretical significance ... it becomes difficult for the reader to judge the value of the research, in terms of what it demonstrates and why this is important' [6, p. 3].[3] In both these cases, poorly-stated and poorly-evaluated research aims and objectives has been considered to hold back research progress, in two fields highly related to computational creativity.

Evaluation takes different forms. In a creativity evaluation context, *summative evaluation* provides a summary judgement of a system's creativity [2, for example]. *Formative evaluation* provides constructive feedback on the system's strengths and weaknesses, applying the results 'in the design of creative programs rather than in the assessment of established programs' [12, p.1]. The current work prioritises formative evaluation to learn from existing systems and inform future system development, whilst acknowledging the value in summative feedback for recognising creative achievements and contributions to creativity knowledge.

2.2 Computational Creativity research

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.

[2] In the context of music generation systems, which often contribute to computational creativity literature.

[3] Both Bundy [5] and Pearce et al. [6] stress that specific goals and value may vary across types of systems; this assertion is upheld in this paper.

Computational creativity is, according to the definition supplied by the steering committee for computational creativity research:[4]

> "Computational Creativity is the study and simulation, by computational means, of behaviour, natural and artificial, which would, if observed in humans, be deemed creative."

This definition has been developed over the past decade by general discussion, with the wording being refined over time by members of the steering committee and organisers of computational creativity research events. [13, 3, 1].

Computational creativity research follows both theoretical and practical directions and crosses several disciplinary boundaries across the arts, sciences and engineering. This has led to the emergence of a community with varying and occasionally disparate aims and motivations, from mainly artistic aims to scientific exploration of creativity, or the pursuit of software/hardware engineering achievements. Research within the field is influenced heavily by artificial intelligence, computer science, psychology and specific creative domains which have received attention from computational creativity researchers to date, such as art, music, reasoning and narrative/story telling [3, 14–16, provide examples].

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).[5] Computational creativity research events have been taking place regularly since 1999, developing from satellite workshops at artificial intelligence conferences[6] leading to autonomous workshops (2007-2008)[7] 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;[8] it can be said that computational creativity research is 'coming of age' [17].

In discussion at the most recent 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,[9] 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.[10] Some members of the steering committee have gone as far as to say that the tackling of creativity evaluation 'probably needs to be deferred until we are substantially more capable in general automated reasoning and knowledge representation' [1, p. 19].[11]

This view has not been echoed in the calls for papers for research 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" (2002-04), "the assessment of creativity in AI programs" (2003), "how we assess creativity in computers" (2007), "the assessment of creativity in AI programs" (2003), "Metrics, frameworks and formalizations for the evaluation of novelty and originality"[12] (2005) and the rephrasing "Metrics, frameworks and formalizations for the evaluation of creativity in computational systems" (2006, 2008). This last wording has appeared in the call for papers for all ICCC conferences to date (2010-2012).[13]

As will be discussed later in this paper, 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. Additionally, this culture may have developed as a side-effect

[4] Definition taken from the http://www.computationalcreativity.net website, which hosts relevant information about the field, including details of research events and of the steering committee who act to shape the general directions that computational creativity research takes.

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

[6] 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).

[7] 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).

[8] The pre-2004 workshops typically contained 10-15 papers, with program committee sizes around 5-15 depending on the event. This has grown to an average of 33 accepted papers and an average of 42 program committee members over the 2010-2012 conferences.

[9] Existing evaluation methodologies for computational creativity are examined later in this section of the paper.

[10] This paper will return later to these discussions at ICCC'11.

[11] This view was also expressed in [18].

[12] The combination of novelty and originality is often used as a reductionist definition of creativity [19–21, 2, 22–24]. Definitional issues shall be returned to later in this paper.

[13] 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 [3], which quickly became adopted more often than more formally stated predecessors such as [2, 19], as shall be shown later in this paper.

of the inclusive efforts to build up a community of computational creativity researchers; decisions on whether a paper should be accepted to a conference were often based around whether the paper would trigger interesting debate, rather than how academically rigorous its presentation was [25]. This was enhanced by the acceptance of 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 [25], position papers often report work in process rather than completed work, so have less requirements imposed for academic rigour in reporting. Position papers allow current work and new methods to be reported even if the work is not yet fully completed. Unfortunately, as that proceedings often do not clearly distinguish technical papers from position papers,[14] this distinction in quality can often be missed, making a lack of evaluative (and other academic) rigour seemingly more acceptable in this community.

In general, the issue of 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.' [11, 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 [25]. 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) reviewer's 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 development is a methodology for evaluation that is accepted as standard practice for tracking and evaluating progress in computational creativity. This issue of how best to evaluate computational creativity systems has generated many more questions than answers. The past ten years have seen discussion of how to evaluate the level of creativity demonstrated by computational creativity systems. In practice, the most significant results from these discussions have been Ritchie's empirical criteria approach [2,11], Colton's 'creative tripod' framework [3] and Colton and Pease's FACE/IDEA model [4], supplemented by a number of additional contributions to discussion [19, 12,26].

2.3 Ritchie's empirical criteria

Graeme Ritchie has proposed a set of formal empirical criteria for creativity [11,2]. 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.[15]

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?' [2, 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?' [2, p. 73]

Originally a set of 14 criteria in [11], four new criteria were added and two existing criteria revised in [2]. 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 criteria is parameterised, allowing further customisation of the criteria to individual definitions of creativity for different domains or systems.

The formal definitions of the 18 criteria can be found in [2]. Here, the criteria are deliberately presented informally, with descriptors such as 'suitable' and 'high' substituted for

[14] See for example the proceedings for ICCC'11 (`http:// icccc11.cua.uam.mx / proceedings/`), ICCC'10 (`http:// eden.dei.uc.pt/ ãmilcar / ftp / e-Proceedings_ICCC-10.pdf`) or IJWCC'07 (`http:// doc.gold.ac.uk / isms / CC07 / CC07Proceedings.pdf`) where a position paper is distinguishable from a full technical paper only by its number of pages.

[15] Whether creativity is contained in the creative process, or in the output generated by a system, or in both (and other aspects besides), is a debate which shall be returned to in greater depth later in this paper, in Section 5.1.2. At this stage of the paper, Ritchie's product-focussed perspective on this debate is highlighted; it will be argued in Section 5.1.2 that this can lead to disregarding of crucial evidence of creativity, and is a somewhat misunderstood interpretation of creativity.

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.
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.

In general, Ritchie's proposals acknowledge a number of theoretical issues, but are relatively impractical to apply for creativity evaluation. Several implementation decisions are left to the choice of the evaluator, which could possibly be taken advantage of for a more favourable evaluation of a particular system's creativity (for example one could weight highly the criteria which the system performs well against, or tweak parameters to best fit the system's interpretation of creativity). Ritchie's criteria have been reported as troublesome to implement [7,27], with some discussion in [28,29] being needed to state and justify implementation decisions.

2.4 Colton's creativity tripod framework

The creative tripod framework [3] emphasises the importance of considering the creative process when evaluating the creativity of a computer system. 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 does not demonstrate these three behaviours, then Colton argues that the system should not be perceived as creative.[16]

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, by evaluating the creativity of two of his own systems: the Painting Fool, an art-generation program [3] and HR, a mathematical reasoning system [30]. Colton justifies both these systems as creative by describing how each system is skilful, imaginative and appreciative. The issue of who decides if a system demonstrates the tripod qualities is not raised. In his examples, Colton chooses to evaluate his own systems, rather than using external judges.

Though not stated, it is assumed from the given examples that each tripod is considered equal, though it would be interesting to extend the tripod analogy to consider how 'balanced' the supporting tripod is and to consider whether or not each of the three qualities is equally important for creativity across all possible creative domains.

2.5 Computational Creativity Theory: the FACE/IDEA models

The FACE and IDEA models are offered as part of a wider research project to formally develop Computational Creativity Theory [4,31,32]. The FACE model is designed to represent creative acts and the IDEA model is designed to evaluate these acts. Collectively the models aim to distinguishes 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 [4] partly motivate their models in response to a consideration of how versions of the Turing test have been applied in discrimination tests [33] 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 the fact that the Turing test has not yet been passed by intelligent systems, setting evaluation benchmarks extremely high for the computational system if systems are to be judged at the same level of creativity as expected for humans.

Pease and Colton suggest as an alternative the FACE and IDEA model. The FACE model *(Frame, Aesthetic, Concept,*

[16] 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.

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 a 'beginning in our efforts to avoid some of the pitfalls of the TT' [4, 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.'

With these ambitious aims, it will be interesting to see how this work develops on its promising potential.

2.6 Survey of current evaluative practice for computational creativity

As described in the previous section, there are options available to the computational creativity researcher to use to perform creativity evaluation in their research. To examine how computational creativity evaluation is currently treated in practice, and identify trends, a survey of recent creative systems was conducted, examining how each system is evaluated. This survey was conducted to measure the frequency with which creativity evaluation is carried out in computational creativity research, to objectively quantify how commonplace creativity evaluation is as a research activity in computational creativity.

2.6.1 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.

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.[17] Discarding papers that did not report details of a creative system, a total of 75 papers were identified for review.

Details of any method of evaluation that was reported in the paper (either evaluation of creativity or of other factors) was recorded in the survey, formal or informal, objective or subjective. There was no discrimination between artistic methods or scientific methods; if any reflection was carried out on the system's achievements, strengths and/or weaknesses, through formal academic evaluation or through other means, then this was noted in the survey and recorded as an example of where evaluation took place.

2.6.2 Survey results: current evaluative practice

The key conclusions of the survey were that evaluation of computational creativity is *not* being performed in a systematic, rigorous manner, but instead current practice is variable and somewhat ad-hoc across the field:

- 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 (evaluation of creativity or of other factors of the system, such as quality of output).

[17] Proceedings from annual events in 2007-2010 were included in the survey, which was conducted in late 2010-early 2011. 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).

- 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 made any mention of existing creativity evaluation methodologies. The most represented methodologies were Colton's creative tripod [3] (used in 5 papers) and Ritchie's empirical criteria [2] (used in 4 papers). Other papers proposed new metrics which were not taken up by other papers, or mentioned creativity evaluation without actually assessing their system's creativity.
- Occurrences of creativity 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).

This survey shows that no evaluation methodology has been accepted as standard for evaluating and comparing the creativity of 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. Existing creativity evaluation methodologies have their critics and there is no consensus within the computational creativity community about which methodology to adopt, to allow the research community to measure progress using a common methodology.

Often the aim of evaluation has been to see if the systems contribute high quality results to a creative domain, for example if the results are aesthetically pleasing, highly valuable, accurate or if they compare favourably to a test set of typical results, or if the processes used by the system are of particular interest. This is related to but distinct from the aim of whether the systems can demonstrate behaviour that can be seen as creative, which is required by the above-stated definition of computational creativity as prescribed by the computational creativity research steering committee. These evaluative aims of quality and creativity can be confused, especially in the absence of a standard evaluation methodology for creativity, though these aims should not be treated as being mutually exclusive.[18] The survey results show, however, that currently the balance between evaluation of quality and evaluation of creativity is skewed towards evaluating quality. Whilst recognising the existence of and need for evaluation of other aspects of the system, this paper concentrates on the goal of creativity evaluation, the practice of which has been shown in this survey to be ad hoc and often ignored.

The results of this survey clearly demonstrate a critical point: computational systems are being presented as 'creative systems' without their creativity being justified; hence 'creative' becomes a descriptor of a system. This becomes a problem in that a fundamental aim of computational creativity research (according to the steering committee's own definition) is for systems to demonstrate behaviour *which would be seen as creative* if demonstrated by humans. Using this key objective as a descriptor of the outcome, without appropriate justification, is not a suitable way of demonstrating that the objective has been met. This is a relevant concern for all systems presented as computational creativity systems, regardless of whether the objectives are largely artistic, scientific, engineering-based or motivated by other concerns, e.g. accuracy of cognitive simulation.

[18] The case study reported below demonstrates the incorporation of value judgements and considerations of domain competence into creativity evaluation.

2.7 Reasons behind the lack of creativity evaluation

The survey findings reported above show 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 positive interest in such matters. However there are a number of points of contention and practical issues that arise when considering how to evaluate computational creativity.

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 (as discussed earlier in this paper). Questions arise about what exactly to test for, what interpretations of creativity should be used, who should perform the evaluation, when evaluation should be performed and what types of tests should be used. There is also a distinction to be drawn between the aim of evaluating creativity or evaluating quality; as the above-described survey of evaluative practice showed, these aims have become blurred to some extent.

In a discussion session at the ICCC'11 conference which partly focused on the current evaluative culture in computational creativity, several points were raised as to why researchers did not include (or did not report) creativity evaluation in their papers. For example, evaluation of creativity could include evaluations done in alternative environments to the traditional idea of formal evaluations. Following on from the above definition of computational creativity as computers demonstrating behaviour which would be deemed as creative in humans, a researcher might choose to define creativity in art (in whole or in part) as a positive audience reaction at an exhibition. *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. A test for this standard would therefore be to exhibit the artwork produced by a system and gauge audience response. Differing views arose during discussion as to whether this type of evaluation would be irrelevant content for a technical paper, or (as this paper advocates) if papers should always include evaluative feedback that is used to verify claims made that would otherwise be left unverified. It is hoped that the recent emphasis on including evaluation details in computational creativity publications will ensure that this content is seen as relevant for such papers.

There are scenarios where a formal evaluation procedure may not be appropriate, as for example in the case mentioned in the previous paragraph. One discussant at ICCC'11 mentioned different types of evaluation appropriate in computational creativity: experiments to evaluate the system, evaluation by peers in the domain in which the system is creative, expert evaluation and audience/target user evaluation.[19] If one of these types of evaluation is prioritised over others, this does not mean that the system is left unevaluated. The evaluation survey above highlights situations where systems were developed and presented without feedback *of any kind* having been elicited.

In terms of performing comparative comparisons with similar systems and evaluating systems with contextual reference to research progress in that area as a whole, one researcher noted the difficulty of finding 'even one other system thats doing exactly what you're doing',[20] hence causing practical difficulties in comparing systems like for like.[21] As a research field, though, activity in computational creativity research has increased greatly over the last decade or so (as described above) and many creative systems have now been developed.[22] *If* similar systems do 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 and how an individual system contributes to this research. This benefit has been demonstrated in research into narrative/story-generation systems [35], a long-standing and thriving research area within computational creativity research [36–39, 20, 40, 27, as example].

There are situations where a creative system operates in a niche where no other system exists; one comment was made about how a central aim of a creative system could be to generate products that no other systems generate, or to generate behaviour distinct from all other systems. For example, the ERI-Designer is believed to be the sole exemplar system of creativity in furniture arrangement [41, 42]. 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 [41], or with a considered comparison of the appropriate crossovers with systems operating in a reasonably similar domain.[23] 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. Some of the papers reviewed did extend and develop existing systems [43, 44] [45–47, for the "MEXICA"system and its variants] [42, 41, for the "ERI-designer" system]. Generally comparison between different versions of the system was limited but present in some cases.

A related issue here is whether it is 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. 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 recent debates in creativity research have concluded [48–51], 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, we do not need to be restricted to evaluating systems that are very similar, for meaningful comparisons to emerge.

Other points were raised around the difficulties of evaluating the concept of creativity itself: identifying reasonable evaluative criteria; the possibility of standardising such criteria over the diverse computational creativity community (and whether a community effort would be appropriate for determining such criteria); the scope of creativity itself and what a creative act entails; the multi-dimensionality of creativity and the degree to which human creativity should be used as a base model for our efforts. Issues were also raised as to the appropriateness of a single score as evaluation of a creative system, which was seen as a reason for not pursuing quantitative summative evaluation metrics: '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 viewpoint 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

[19] 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 [34, p. 1].

[20] All comments in this section are anonymised.

[21] Another issue mentioned during this part of the discussion was that it was often difficult to obtain up-to-date, maintained and fully-working materials to use for evaluation, such as the system's source code or products.

[22] The evaluation survey looks at 75 papers describing such systems.

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

on evaluation at the Dagstuhl seminar in computational creativity [34] that '[e]valuation can feed back into the system to affect (hopefully improve) future performance' [34, p. 1].

Clearly more investigation is needed into a suitable creativity evaluation approach, learning from what has been done so far. In order to evaluate computational creativity, a clearer understanding of creativity itself would be beneficial as a basis to inform such investigation. Section 3 examines how the word 'creativity' has been defined in human and computational creativity.

3 Definitions of creativity

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; there is a lack of consensus on the exact definition of the word *creativity* which hinders creativity research progress.

The difficulty of capturing in words an adequate definition of creativity should not discourage us from such an attempt [52,53,3], even though other researchers have been swayed away from this task, as has been reported elsewhere [54,55].

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 [56,57,36,37,58–60], combining novelty and value [19–21,2,22–24], combining exploration, transformation and association of concepts in a conceptual space [61,62], or defining a 'creativity-like' concept [19,63]. 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.

Increasingly, definitions of computational creativity are starting to refer to interpretations of human creativity in a computational manifestation [13,3,1] without further clarifying what human creativity is - showing the need for a working definition of creativity itself. Consequently it is important to examine how creativity is defined in human-focused research areas. There are complications in constraining this seemingly mysterious term [64,65,61,66] to definition. As Table 1 shows, dictionary-style definitions are inadequate and impractical for creativity evaluation, being restricted by their format to provide only cursory, shallow insights into what creativity truly is.

In creativity research, many have helped 'demystify' creativity. Models of the creative process [67–70] and psychometric tests of creativity based on the creative person generating products [71–73] are useful in a certain sphere, but the complexity of creativity requires a broader, more multidimensional treatment. The confluence approach to creativity [54,74,75] takes a reductionist approach, understanding creativity as a whole by breaking it down into smaller constituent parts.[24] In such research, the *Four Ps* construct [52, 76–78] 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.

In the search for a more precise and accurate definition of creativity, we can explore how creativity has been defined and used in the law, via some relevant examples (US and UK law). Here too, though, there is no standard definition of creativity to be found [79,80], despite the need to detect the presence of creativity for legal reasons [81–83].[25]

In general, several competing views of creativity exist. Sometimes differences of opinion do not need to be directly resolved but can co-exist, such as whether creativity is centred around cognitive function [61] or whether it is embodied and situated in an interactive environment [69,87], or whether creativity is domain-independent [48] or domain-specific [49,69]. 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 [67,88], one should focus on human everyday creativity, of which genius is a special case [52,89,61,90]. Similarly, P-creativity, or creativity that is novel at a personal level but not necessarily at a wider social level, encompasses H-creativity [61], creativity that is historically novel and has never been seen before on a global scale.

To satisfy the need for clear and defined benchmarks by which to evaluate progress in creativity research, particularly computational creativity research, there is indeed much useful contributory material towards a satisfactory definition. Despite the several offerings, however, no universally accepted definition is available. 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. What remains to be done is to draw this assortment of material together and unify the perspectives where possible to remove disciplinary separation and boundaries. Section 4 de-

[24] The work in Section 4 will adopt a confluence-style approach, seeking to capture a wider disciplinary spectrum of perspectives on creativity that has previously been attempted [54,74,75].

[25] What does appear when reviewing legal research are interesting discussions of whether or not computational creativity can be legally recognised [84,79,85,86].

Table 1 Dictionary definitions of 'creativity'. 'Creativity' often has no separate dictionary entry, but is included in definitions of 'creative' or 'create'. Definitions of 'creativity' that do exist may be defined using the words 'creative' or 'create'. Hence definitions are given for 'creative' (where available) and 'create' to supplement definitions of 'creativity'. For readability, some definitions are edited slightly to standardise formats and remove etymological/grammatical annotations.

Dictionary	creat-	Definition
Oxford English Dictionary 2nd ed. (1989) pp.1134-5	creativity	Creative power or faculty; ability to create.
	creative	Having the quality of creating, able to create; of or relating to creation; originative. b. Inventive, imaginative; of, relating to, displaying, using, or involving imagination or original ideas as well as routine skill or intellect, esp. in literature or art. c. Esp. of a financial or other strategy: ingenious, esp. in a misleading way. 2. Providing the cause or occasion of, productive of.
	create	1.a. Said of the divine agent: To bring into being, cause to exist; esp. to produce where nothing was before, 'to form out of nothing'. b. with complemental extension. 2. To make, form, constitute, or bring into legal existence (an institution, condition, action, mental product, or form, not existing before). Sometimes of material works. 3. To constitute (a personage of rank or dignity); to invest with rank, title, etc. 4. To cause, occasion, produce, give rise to (a condition or set of circumstances).
Collins English Dictionary (1998) p.371	creative	1. having or showing the ability to create. 2. inventive or imaginative. 3. characterized by imaginative and usually dubious bending of the rules.
	create	1. to cause to come into existence. 2. to invest with a new honour, office, or title; appoint. 3. to be the cause of. 4. to act (a role) in the first production of a play. 5. to be engaged in creative work. 6. to make a fuss or uproar
Chambers 21st Century Dictionary (1999) p.316	create	1. To form or produce from nothing. 2. to bring into existence; to introduce. 3. to cause. 4. to produce or contrive. 5. said of an artist etc: to use one's imagination to make something. 6. To make a fuss. 7. said of an actor: to be the first to play (a certain role). 8. to raise to an honourable rank.
The Penguin English Dictionary 2nd ed. (1969) p.174	creativity	creative power or faculty; ability to create
	creative	having power to create; related to process of creation; constructive, original, producing an essentially new product; produced by original intellectual or artistic effort
	create	make out of nothing, bestow existence on; cause, bring about; produce or make something new or original; confer new rank etc on; (theat.) be the first to act (a certain part); make a fuss
Webster's 3rd New International Dictionary (1961) p.532	creativity	the quality of being creative; ability to create
	creative	1. having the power or quality of creating; given to creation 2: PRODUCTIVE - used with 3: having the quality of something created rather than imitated or assembled; expressive of the maker; IMAGINATIVE
	create	1: to bring into existence; make out of nothing and for the first time 2: to cause to be or to produce by fiat or by mental, moral, or legal action 3: to cause or occasion - used of natural or physical causes and esp. of social and evolutionary or emergent forces 4a: to produce (as a work of art or of dramatic interpretation) along new or unconventional lines) b: to design (as a costume or dress)
The American College Dictionary (1963) p.284	creative	1. having the quality or power of creating. 2. originative; productive.
	create	1. to bring into being; cause to exist; produce. 2. to evolve from one's own thought or imagination. 3. to be the first to represent (a part or role). 4. to make by investing with new character or functions; constitute; appoint; 5. to be the cause or occasion of; give rise to.

scribes how this task has been tackled, reports the results of this work and reflects on the implications for the study of computational creativity.

4 Key components of creativity

This part of the paper aims to extract the common themes of creativity across disciplinary or domain specifics, rather than adding another to the mass of existing definitions. Combining several viewpoints across various perspectives in creativity research leads to a more inclusive summary of how we define creativity.

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

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 [50]. As discussed above, 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 efforts to capture and talk about creativity in words but the above discussion 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.

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 [71, 72,91–94] focus on *divergent thinking* and *problem solving* as key attributes of a creative person. In contrast, computational creativity research [58,20,19, for example] 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.

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 an approach of 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 [95,54,74,75] and investigates what these components are. Similar approaches have been applied to better understand other concepts that are difficult

to define in words, for example: consciousness [96], personality [97], flow [98] and musical perceptual preferences [99].

Some of the issues surrounding creativity definition have been debated for decades; clearly these cannot all be resolved satisfactorily in the scope of this paper. 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.

Cognitive linguistics advocates that the meaning of a word is dependent on the context it is used in [100]. 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 [101–103]. Using this premise, techniques from the field of computational linguistics have been used to empirically identify a collection of *components* of creativity.

4.2 Methodology for identifying components of creativity

A corpus of academic papers was collated, representing sixty years of research into the nature of creativity, from research perspectives such as psychology, education, computational creativity and others. This corpus (the *creativity corpus*) was analysed against a corpus of matched papers on subjects unrelated to creativity (the *non-creativity corpus*) using the log likelihood ratio (LLR) statistic on word frequencies in the two corpora.

694 words were identified which appeared significantly more often in the creativity papers corpus than expected. Lin's semantic similarity measure [104] and the Chinese Whispers clustering algorithm [105] were used to cluster these words into groups of words with similar meanings. Through these clusters and similarity data, 14 themes emerged, representing different components of creativity that collectively contribute to the overall meaning of the word 'creativity'. Individually these components make creativity more tractable and easier to understand, by breaking down this seemingly impenetrable concept into constituent parts. Together these components act as *building blocks* for creativity, each contributing to the overall presence of creativity.

4.3 Results: components of creativity

The 14 components of creativity identified in this linguistic analysis are presented in Figure 1. To summarise the meaning of each component:

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. 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.
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.
4. General Intellect
   - General intelligence and intellectual ability.
   - Flexible and adaptable mental capacity.
5. Generation of Results
   - Working towards some end target, or goal, or result.
   - Producing something (tangible or intangible) that previously did not exist.
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.

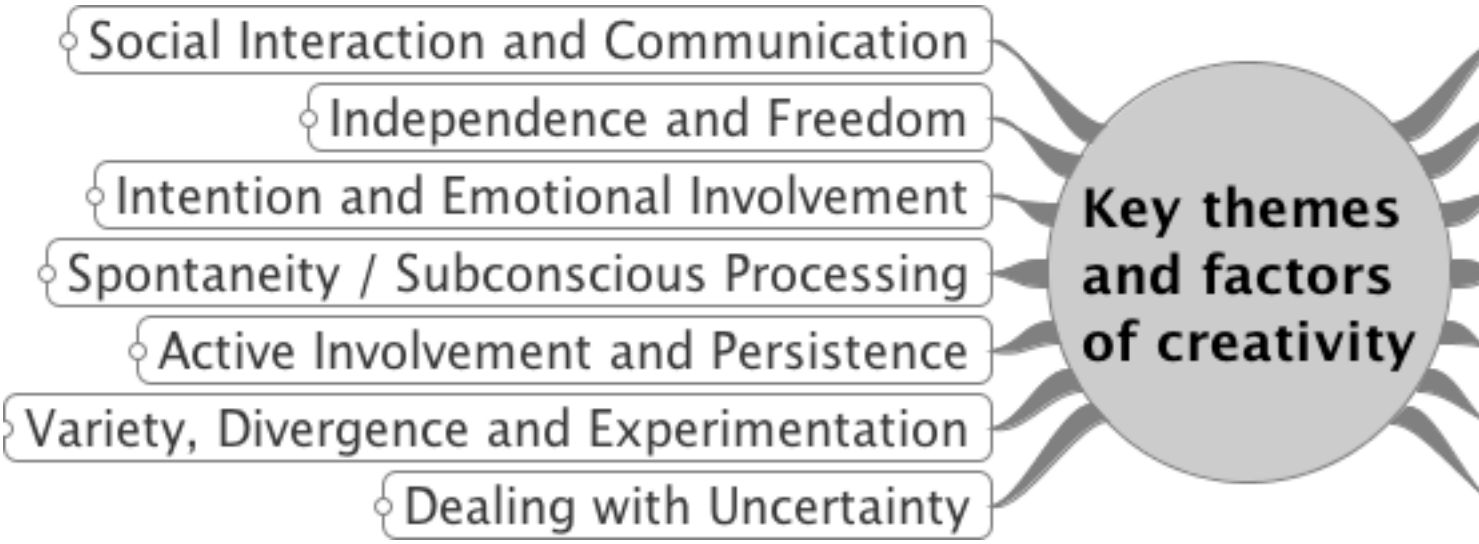


Fig. 1 Fourteen key components of creativity

  - 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.

The components collectively provide 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. Section 5 takes these components to be used as evaluation standards for computational creativity.

5 Towards a Standardised Procedure for Evaluating Creative Systems

To assist the researcher with this process, we now have several aspects with which to examine computational creativity systems, i.e. the components of creativity presented in Figure 1. This componential breakdown of creativity affords 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.

In response to the findings of the 2011 evaluation survey, *Evaluation Guidelines* have been proposed as heuristics to follow in computational creativity evaluation [106]. The Evaluation Guidelines require evaluators to clarify what is being evaluated under the term 'creativity' then perform tests strictly relating to that definition, rather than impose a single definition of creativity across all domains. The justification is that creativity is multi-faceted and complex, with

decades of creativity research so far failing to identify a universal definition. The Evaluation Guidelines are intended to be customisable to specific requirements for creativity in a wide range of domains, where the evaluator states and justifies the most appropriate existing evaluation suggestions. Existing evaluation methods can be incorporated if justified as relevant.

Practically, the Evaluation Guidelines make the evaluation process more relevant to creativity. By making the evaluation criteria transparent and available to other researchers, this helps avoid unnecessary duplication of effort. In their current state, the heuristical nature of the Evaluation Guidelines makes them less practical to apply than the other evaluation methodologies mentioned above.

This paper expands the Evaluation Guidelines into methodological steps to follow in evaluation, with guidance on related issues and the practicalities of each step in more detail and taking into account reflections from existing evaluation methodologies. These methodological steps are presented in Table 2 and are collectively referred to as the *SPECS* methodology: the *Standardised Procedure for Evaluating Creative Systems*. Here, each step is presented and discussed.

Table 2 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 is the most suitable.
      – *The 14 components of creativity in Figure 1 are strongly suggested as a base definition.*
   (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 the Figure 1 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 by 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.

5.1 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)?

As Ritchie [11] says, '[i]t is important to be explicit ... about the criteria that are being applied in making judgements of creativity' (p.3). Following from Jordanous [106], the customisation of a general definition of creativity to a specific type of creativity is preferred to defining creativity directly from the specific perspective of a given domain [19, 2]. Taking a domain-specific perspective on creativity risks over-specialising to that domain [51]. This has been evident in the tendency to move evaluative focus away from creativity to domain value [106]. Knowledge of a domain is important, however, when modelling 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'[107, p. 115], it is expected that those building a system appreciate requirements for creativity in that domain and should evaluate their systems accordingly.

Depending on how creativity is defined by the researcher(s), existing evaluation frameworks [2,3,19, for example] 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 the key components of creativity for a particular creative domain, the creative tripod [3] would be appropriate.

As the Figure 1 components were derived from general discussions about creativity, they should be used for Step 1a of SPECS (what does it mean to be creative in general) and can 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.

5.1.1 Identifying domain-specific requirements

While some aspects of creativity are common to all types of creativity, other aspects will vary in importance according to domain [107,19,2,51]. 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. The importance of each component can be investigated in many ways, such as consulting the opinion of experts and/or the general public, analysing prior research or consulting general knowledge about that field.

If the components of creativity in Figure 1 are used for Step 1a, the researcher should investigate the relative important of each component in their particular domain to weight the contribution of each component accordingly.

The two-step nature of Step 1 of SPECS, from a general definition to a specific characterisation of creativity in the given domain, is deliberate. This ensures that both domain-independent and domain-specific aspects of creativity [51] are taken into account when identifying what is necessary for creativity in a particular system. Taking this two-step approach prevents the adopted definition of creativity from being too tailored to the specific domain being tackled, to lessen the risk of creativity being defined for evaluation as something related to creativity in that particular domain but distinct to creativity itself.

5.1.2 The product/process debate

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.

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 [11]. 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. In response, Colton [3] cites examples from art to demonstrate the importance of the creative process when evaluating creativity, at least for art forms such as conceptual art.[26] Ritchie [62] concedes that when testing theoretical models, 'the mechanisms are the whole point' [62, p. 147].

While we can only use the material we have available to form an evaluation, previous evaluation experiments [33, 26] show that people often make assumptions about process in their judgements on product. The adage that a magician never reveals their secrets is analogous, as systems can appear less creative when you know how they work [11, 3, 26].[27] Our interpretation of how something was produced is also important - even if the processes are unknown, speculations may be made if people are repeatedly exposed to the systems (human or computational) they are evaluating [33].

[26] Conceptual art is where the concepts and motivations behind the artistic process form a significant contribution of the artwork.

[27] Colton's solution is to report systems in high-level terms only, rather than giving details of the program [3, p.8].

In human creativity research, creativity has been defined through the *Four Ps* [52]: *process*, *product*, *person* and *press* (or environment). Current 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 [3] who considers how an audience perceives the creativity of a system. Observations of a creative *person* in a *press*/environment may also be useful.

5.1.3 Distinguishing between P-creativity and H-creativity

Boden [61] distinguishes between *P-creativity* (psychological creativity), where a produced artefact is novel to the creator but has been discovered elsewhere, and *H-creativity* (historical creativity), where the produced artefact is unknown both to the creator and to society in general. Researchers should clarify which type of creativity is being addressed [33, 2, Edmonds (personal communications), 2009]. The computational creativity community has focused on P-creativity, mainly for practical reasons; it is far simpler to compare output against knowledge accessed by the system than to compare system output against the sum total of all knowledge in the world.

Pearce et al. [33] justify a focus on P-creativity through noting that H-creativity is a subset of P-creativity, where set membership is determined by historical and social factors. A focus on P-creative achievements would by definition include H-creative achievements as well. As Boden [107] points out, 'Our concern ... must be with P-creativity in general, of which H-creativity is a special case' [107, p. 112].

5.2 Step 2: Identifying standards to test for

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. For example, Ritchie [2] defines creativity as a combination of novelty and quality [2, pp. 72-73]. tested via his criteria. Colton [3] sees creativity as the combination of skill, imagination and appreciation, so these become standards. If using the components of creativity pictured in Figure 1, each component becomes a standard to be tested.

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*' [61, 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.

5.3 Step 3: Testing systems using standards

Test your creative system against the standards stated in step 2 and report the results.

The choice of evaluation tests depends on the standards chosen to be tested, the preferences, capabilities and equipment/facilities of the researcher(s) involved, and previous testing of those standards or related standards.

As the purpose of SPECS is to provide detailed feedback on the system rather than an overall 'creativity score', a single aggregated measure of the system is not necessary (and may be misleading). It is important, though, to give feedback on more important aspects for the domain in question, giving these more emphasis than less important aspects.

5.3.1 Preconceptions about computational creativity

Computational creativity researchers have stressed that creativity is 'in the eye of the beholder' [1,14]. As emphasised in the *Four Ps* approach to creativity mentioned above [52], the opinions of the audience are crucial in making, distributing and maintaining creativity judgements. People's evaluation of computational creativity can be influenced by preconceived notions and beliefs [33,61,3]. People may be reluctant to accept the concept of computers being creative, either through conscious reticence or subconscious bias [108]. On the other hand, researchers keen to embrace computational creativity may be biased towards give 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; this should be accounted for during evaluation.

5.3.2 Using quantitative and qualitative methods

Can creativity evaluation be conducted through purely quantitative measurements? Quantitative measurements can be 'an attractive approach to assessing creativity' [109, p.3] due to their objective nature, but this would require the subjective concept of creativity to be captured in quantitative measurements. 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 (capturing changes in opinion over time) [19,33,2,3,62].

Despite the quantitative form of Ritchie's criteria [2], the criteria depend heavily on the two rating schemes of typicality and value, usually obtained in practice through subjective assessment by human judges [28,29,109,7]. Ritchie [11] acknowledges that subjective ratings and appraisals depend on the particular judge's opinions, background knowledge and even perhaps on their current mood, though he does not discuss practical solutions such as using a number of different judges. Excess qualitative data given by judges is discarded. All but one of Pease et al.'s tests [19] are presented as formal quantitative tests similar to Ritchie's criteria, with one test using human judgements. The creative tripod framework [3] uses qualitative evaluation only: a system is creative if it is perceived to be skilful, imaginative and appreciative of its work.

It is recommended here to employ both quantitative and qualitative evaluation methods, to allow for quantitative comparison of systems whilst incorporating human judgement. Where possible, results for individual tests, representing the 'multidimensional structures' in creativity [107, p.113], should be viewed in combination.

5.3.3 Practical issues in using human judges

Soliciting human opinion in creativity evaluation is one 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 [107], but raises various practical issues to resolve. Human evaluators can say whether they think something is creative but may only be able to give limited explanation of their opinions. Human opinion is variable; what one person finds creative, another may not [1,14]. This may be influenced by the expertise of the judges, previous experience of the systems and/or similar systems, or preconceived notions about computational creativity. Large numbers of participants may be needed for a general consensus of opinion and there is no guarantee consensus can be reached. Time and resources are needed to devise and run suitable studies, as well as the need to obtain ethics approval, attract enough suitable participants and fund the paying of participants (or rely on goodwill).

5.3.4 The expertise and independence of the evaluators

Boden 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 [107]. Hence, she argues, their [expressed] opinions will generally be more informative and more grounded in fact.

In the survey on current evaluation practice [106], 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).

The issue of who should evaluate a system has been overlooked in previous discussions of creativity evaluation.

Two key contribution so far to computational creativity evaluation [2,3] do not make any recommendations on this matter, except for an underlying implication that the methodologies are tools for researchers to evaluate their own systems. Questions of impartiality arise when self-evaluating work; both Ritchie's criteria and Colton's creative tripod can be adapted (intentionally or unintentionally) to portray an evaluation as desired [26].

Transparent methods are vital for holding the researcher to account on their evaluative decisions. Using external evaluators helps avoid accidental or intentional bias [110,19], but availability, time constraints and willingness to participate can dictate who evaluates a system, often making internal evaluators (people involved in system development) a more attractive option, even with concerns of bias.

6 Application of the SPECS methodology to an evaluative case study

To exemplify the application of SPECS, the SPECS methodology has been used to evaluate the creativity of various computational creativity systems, generate feedback on the systems' performance and creativity and compare the systems against each other.

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' [111]. Lewis 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' [112]. Issues of choice and liberty were raised by Lewis, to do with having a choice of what expressive actions to perform and when to perform them. Neural evidence [113–115] shows that brain activity during improvisation relates to brain activity when making choices. Lewis 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 [116]. 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 [117].

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 expressed 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' [118, p. 140]

Pressing 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 [116] or components [117].

6.2 Musical improvisation systems being evaluated

The SPECS methodology was applied to compare and contrast the creativity of four musical improvisation systems:

- The improvisation system in [7], named *GAmprovising* for this study. GAmprovising [7] 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 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.
- *GenJam* [8]. GenJam [8] (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 [119], 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 [8] 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.
- *Impro-Visor* [9] looks for patterns and rules in existing improvisations, constructing grammar representations of these. To analyse existing improvisations, Impro-Visor breaks each solo up into short fragments, typically one bar long, and translates each fragment into more abstract

representations of individual notes, chordal structure and melodic contours (the musical shapes that the melodies make through raising or lowering pitch). Recurring or similar patterns are used to create a grammar (a set of rules that summarises all the abstract patterns). To generate melodies from the grammar rules it has learnt, Impro-Visor chooses notes to fit the patterns as closely as possible, using probability calculations. When generating an improvised solo, Impro-Visor randomly chooses one abstract pattern for each section of the tune, and chains them together to make a longer solo.

– *Voyager* [10] 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 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.

6.3 Applying SPECS

Resources and data on all four systems were collated to assist the application of the SPECS methodology to evaluate and compare the four systems' creativity.

6.3.1 Step 1a: Domain-independent aspects of creativity

Common components of creativity have been identified that we prioritise as important for creativity in general, across all domains.[28] The components will be used for Step 1a, as per the recommendations given above.

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. 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 summarise what these words meant to them in this context. These words were taken from the top results of a precursor study to the work deriving the components in Figure 1 [120], investigating how language use in writings about creativity differed from standard British English language usage as represented in the British National Corpus [121]. 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 [120]. In general, though, the answers given focused quite specifically on the relevant word or group of words, meaning that the data from the first ten questions mainly acted as practice trials for the eleventh question.

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 if there were any words which they felt were important for describing creativity in musical improvisation that have not been mentioned so far, and if so, what these words are and why they are 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.

Participants were asked about their musical experience and training and, if they were musicians,[29] what instruments

[28] This set of components is pictured in Figure 1.

[29] 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. The questionnaire distribution was however weighted towards professional musicians and improvisers.

Table 3 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

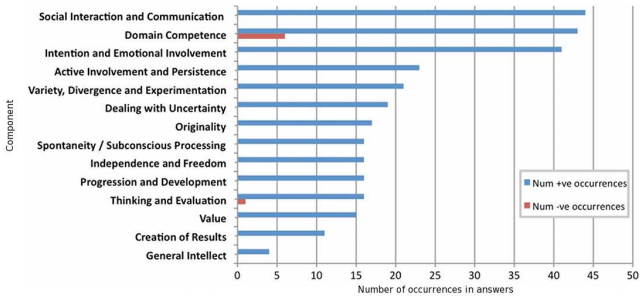


Fig. 2 Importance and relevance of creativity components to improvisation.

and genres they played. Participants came from different improvisatory backgrounds and with different levels of expertise and experience of a variety of musical styles.[30] 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 Their responses to the question about creativity, the debrief question and any follow-on correspondence post-questionnaire were analysed using the 14 components from Figure 1. 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. 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.

Figure 2 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 Figure 1, 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 4).

Table 4 Converting the I_c values into weights representing the importance of each component.

Component	I_c	weight%
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%

6.3.3 Step 2: Standards for evaluating the creativity of musical improvisation systems

Drawing upon the results from the above steps, the musical improvisation systems were evaluated along 14 standards, one for each of the 14 aspects in Figure 1. Again this follows the recommendations given above for SPECS.

6.3.4 Step 3: Evaluative tests for creativity in musical improvisation systems

There were six judges in total for Case Study 1.[31] The judges involved were experts in musical improvisation. 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.

[30] 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.

[31] 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 paper.

Each judge evaluated two systems each.[32] 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.[33] 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. They were also given example comments which collectively represented each component. These comments were taken from quotes in the annotated questionnaire data for Step 1b. These statements were used to help analyse the four 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).

Each system was evaluated in a dedicated 50-60 minute session. Judges would evaluate one system, take a break of at least 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 different orders, 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.

6.4 Results and discussion

Both quantitative and qualitative data was collected on each system during SPECS evaluation. The quantitative data was the judges' ratings for each component, and the qualitative data gathered was from comments made by judges during interview.

6.4.1 Judges' evaluation ratings

The evaluation data obtained from the judges (weighted and unweighted) is presented in Figures 3(a) to 3(d). Some observations can be drawn from the raw data from the judges' ratings. 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, as per Step 1b of SPECS.

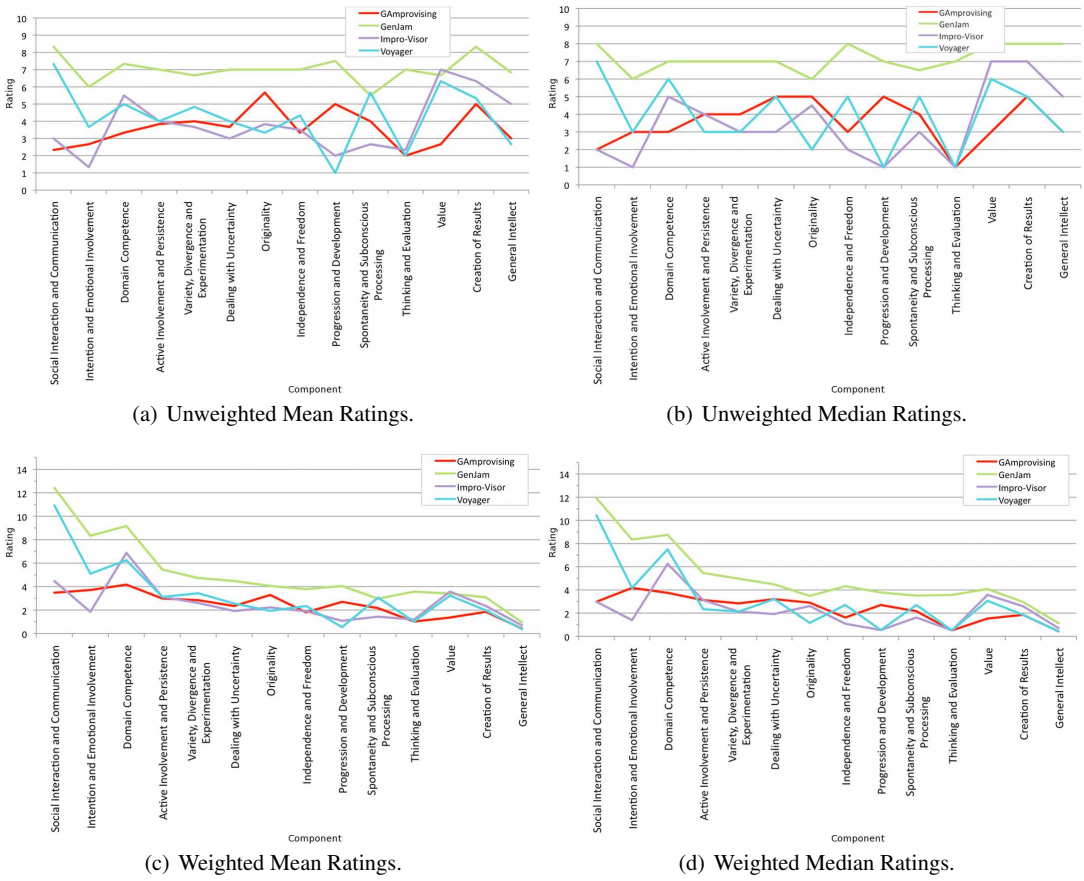
Unweighted ratings of components Overall in the unweighted ratings, GenJam performs best in 12 out 14 components when taking mean ratings per component across judges, and in all 14 when taking medians. For mean ratings of *Spontaneity and Subconscious Processing* and *Value*, GenJam is rated slightly lower on average than one other system; Voyager is rated 0.2 higher on average than GenJam for *Spontaneity and Subconscious Processing* and Impro-Visor is rated 0.3 higher on average than GenJam for *Value*. These differences are minimal though and GenJam still attracts ratings which average above the midpoint of 5 out of 10.

It is difficult to distinguish visually between the other three systems, with much crossover and no system emerging as consistently better or worse than the other two. 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. It should be remembered, though, that comments on individual components become more or less relevant to creativity in musical improvisation after the components have been weighted. 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.

Weighted ratings of components After weighting, 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 out of a maximum of only 1.4, rather than 10. Again GenJam is generally rated higher overall. A rough estimate score of the systems' relative creativity performance can be taken as the mean of the weighted ratings from all judges. GenJam scores a mean of 5.0. The next highest rated system is Voyager (3.3), which

[32] Judges were restricted to evaluating two of the four systems rather than all four due to practical restrictions on time.

[33] 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 as the mid-point between the two extremes of 0 and 10.



(a) Unweighted Mean Ratings.

(b) Unweighted Median Ratings.

(c) Weighted Mean Ratings.

(d) Weighted Median Ratings.

Fig. 3 Mean and median averages for all judges' evaluation ratings (before and after weighting by component importance) for the four systems.

is then followed by Impro-Visor (2.5) and GAmprovising (2.4).

As has been emphasised throughout this paper, though, summative feedback such as these scores are not so helpful for system development and learning from what other systems have done. For these purposes, formative feedback is more useful at giving detailed comparative critiques of how the systems are creative, and how they are not. 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 performance can be inspired in some cases by the judges' comments on the systems, which are reported per component in Section 6.4.2.

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).

Impro-Visor The components where Impro-Visor had been rated relatively highly, such as *Value* and *Generation of Results*, were quite unimportant components for musical improvisational creativity. Once the ratings had been weighted, the highest mean rating that Impro-Visor received was for

Domain Competence (6.9). As for the previous two systems, work on improving the three most important components would bring a large potential points gain (a total of 28). Additionally, given Impro-Visor's similar style of improvisation to GenJam, with both systems improvising solo jazz melodies over standard chord progressions, much could potentially be learned from GenJam in improving the autonomous improvisational capabilities of Impro-Visor.

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 four 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 [10], 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 [111].

Observations 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. Similarly, 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, as has been discussed above. Improvements in such areas are likely to be minimal compared to in the more important components.

GenJam generally stood out from the other systems. 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 three systems for 12 of the 14 components. Of the other three systems, GAmprovising seems to follow different patterns of ratings to the other two systems, which are quite similar. It also seems to have less peaks and troughs in the mean ratings trend shown in Figure 3(c), 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 less observations of note. One thing that can be noted, from comparing Figures 3(b) and 3(b) (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 three systems becomes less noticeable after weighting.

6.4.2 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

- *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 that the system demonstrated greater musical abilities than the other judges.
- *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 it's 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

- *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 [8] 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.

Impro-Visor

- *Social Interaction and Communication* was seen as limited.
- *Intention and Emotional Involvement* was also not demonstrated enough to satisfy the judges.
- On the other hand, *Domain Competence* was praised, except by Judge 3 who remarked that domain knowledge was dependent on what the system had been trained on.
- *Active Involvement and Persistence* divided opinions, with Judge 2 giving positive feedback based on the rigorous learning of grammars but Judges 3 and 6 finding the system often gets stuck in repetitions of similar style.
- Whilst Judge 6 remarked how output was not the same every time, all judges remarked on how *Variety, Divergence and Experimentation* was controlled by the user, not the system.
- Impro-Visor was perceived as being capable of *Dealing with Uncertainty* by Judge 6 but not necessarily able to deal with unknown situations by Judges 2 and 3.
- *Originality* was seen as occurring in predictable ways by chance.
- The *Independence and Freedom* in Impro-Visor was criticised, with the system seen to require a lot of human intervention (although Judge 2 remarked that it needed no more intervention than a human needs).
- *Progression and Development* was not seen by any of the judges.
- Judges were unsure about *Spontaneity/Subconscious Processing*, as there was little in the way of real time improvisation within the system.
- Impro-Visor was seen as lacking in *Thinking and Evaluation*, with most evaluation done by the user while the system generates solos.
- *Value* within Impro-Visor was perceived as fairly high. As Judge 6 remarked, 'It's not going to transform the domain but it ... has some value, even if playing mostly cliched licks'. The system's value as an education tool was recognised.
- For *Generation of Results*, the system did produce results though Judge 3 questioned whether the system knew it had produced results so could stop improvising.

– Impro-Visor was seen by Judge 6 as possessing some *General Intellect*, but the other judges felt it was more specialised to one task than generally intelligent.

Voyager

– *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 disagreed.
– *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 [10] was more sophisticated than the system itself.

To summarise, this case study has demonstrated the application of SPECS for creativity evaluation, generating much information from evaluating the four musical improvisation systems. Whilst GenJam seems to have emerged from the evaluation process as the most creative system, according to general consensus of opinion, the findings from SPECS that have been more useful have been the large amount of feedback for each system, to inform future development of the systems. For example GenJam's creator, Al Biles, could take inspiration from how Voyager exhibits *Spontaneity and Subconscious Processing* and the *Value* perceived in Impro-Visor and its products. For each system, strengths and weaknesses are identified for that system's creativity:

– GAmprovising's identified strengths were its ability to create 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.
– Impro-Visor was considered the system with highest value and again it had a good ability to create results. Much poorer scores were recorded for Impro-Visor's ability to develop its improvisations and to express emotions and intention; this last point was prioritised by survey participants alongside more expected abilities such as domain expertise and the ability to communicate and interact with other musicians and the audience.
– 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.

The case study has investigated which components make the greatest contribution to musical improvisation creativity. The SPECS results show that for all four 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 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 specific expertise and knowledge.

When building musical improvisation systems and intending to make them as creative as possible, this 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 6.3.2 and several being highlighted in judges' qualitative feedback), the top three components collectively account for 41.3% of musical improvisation creativity, according to the weights identified in Section 6.3.2, illustrated in Figure 2.

6.5 Evaluation of the SPECS methodology

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 [106].

As Section 2 described, other creativity evaluation methodologies and strategies have been proposed in the past. In the case study described above, alternative creativity evaluations to SPECS were also conducted:[34]

- A survey of human opinion, to try and capture a 'ground truth' for creativity evaluation [123]
- Ritchie's empirical criteria framework [2]
- Colton's creative tripod framework [3]

The evaluative process and results were compared and contrasted, to consider how usable and accurate the evaluation data was when obtained through each evaluative approach. Upon comparing results, the first thing to note was that there did not seem to be such a thing as a 'ground truth' for creativity, given varieties in people's opinions. This was found even when consulting large numbers of people (111 people took part in the survey for the case study).

Most creativity evaluation methods reported GenJam to be the most creative system overall, with GAmprovising as the least creative system, although methods disagreed on the relative placings of Voyager and Impro-Visor between these two extremes. Summative comparisons are of limited value to the researcher, though, particularly in terms of identifying strengths of the system to contribute to knowledge and weaknesses of the system to be improved. Ritchie's criteria in particular have proved this, as the level of abstraction away from the system itself in the feedback did somewhat obscure what could be learnt from that feedback. The loss of information was magnified as all qualitative feedback was disregarded and criteria were either reported as *TRUE* or *FALSE*, without any measurement of magnitude to distinguish how well a system had performed on a given criterion.

For the purposes of progressing in research, learning from advances and improving what has been done, formative evaluation feedback is more constructive. Colton's creative tripod framework, the human opinion surveys and the SPECS methodology all performed well at providing this feedback. In the implementation of SPECS using the components represented in Figure 1, SPECS gave feedback in the most detail but required the most information to be gathered for creativity evaluation.

Colton's creative tripod performed relatively well but did not uncover elements important for creativity outside of the tripod elements of *skill*, *imagination* and *appreciation*. The case study showed that other aspects were as important, or more so, for creativity in the domain of musical improvisation. For example, the ability to interact with and react to what is happening externally and communicate with others was considered most important for musical improvisational creativity. The ability to demonstrate intention to be creative and an emotional involvement with the creative process was also found to be very important for creativity. More generally, SPECS is more inclusive than Colton's tripod of the recent work in creativity theory that posits that the relevance and contribution of different aspects will vary according to the creative domain being investigated [50,51].

In the surveys of human opinion, participants reported difficulties in evaluating the systems' creativity. In particular, several people wanted a definition of creativity to refer to in creativity evaluation and did not want to rely on their own intuitive understanding. This may be due to biases against computational creativity or a lack of acquaintance with a computer being creative. Most participants did however appear to be positive or at least neutral towards computational creativity; whilst this might not stop subconscious biases affecting evaluations [108] it would reduce the likelihood of many overt negative biases being demonstrated. The difficulties may instead be due to people finding it difficult to objectively rate a subjective and ill-defined concept like creativity when asked to, although participants reported that they generally felt confident about their given responses.

SPECS compared well against the other creativity evaluation methods used, especially in terms of formative feedback generated, however a number of points for consideration did arise during the implementation of SPECS in the case study: the appropriateness of using the fourteen components of creativity reported in Figure 1, issues with identifying baseline standards for comparison of creativity evaluation results, various concerns surrounding the use of human judges and subjective evaluation data, practical issues in the creativity evaluation process and more general reflections on using SPECS for evaluation of creativity.

It should be noted that all of the above evaluation methods mentioned in this paper, and more, could be applied within the framework of SPECS. Although it is strongly rec-

[34] The FACE/IDEA models [4], published after the evaluations in this case study had been performed, will also be applied in the near future and results will be published in [122].

ommended that the Figure 1 are used as the base definition for creativity, in Step 1, a system evaluator could choose to use Colton's creative tripod framework [3] or Ritchie's empirical criteria framework [2] as the adopted definition of creativity, *if this decision is stated and justified by the evaluator as being the most relevant interpretation of creativity to adopt*.

The 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 allows computational creativity researchers to perform evaluation of creativity of their systems in a standardised and appropriate manner.

7 Summary and Conclusions

This paper focuses on evaluation of computational creativity, addressing the research question: how should we evaluate the creativity of a computational creativity system? There is a clear need for creativity evaluation in the field of computational creativity research, to track progress and identify strengths and weaknesses in our research.

The research question driving this research is:

> How should we evaluate the creativity of a computational creativity system?

This paper offers a practically applicable and useful answer in the form of the SPECS methodology.

- The SPECS methodology has arisen from the consideration of issues surrounding understanding and evaluating creativity.
- SPECS was demonstrated in use in a case study, utilising a working definition[35] of creativity derived in Section 4 and further investigations.
- This use of SPECS was evaluated in comparison to alternative methods of creativity evaluation using other methodologies and/or human judgement (Section 6.5).
- 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.

[35] Definitions of creativity are often supplied as a list of components or contributory aspects of creativity [73,54,61,50, as a selection of examples].The Figure 1 components are offered as a *working* definition because the empirical methods used to derive them are based around writings from the time period 1950-2009; as creativity changes over time, the components may need to be updated in the future, but for the present time, they are derived from writings from the last sixty years of research on creativity.

7.1 Future development of this work after feedback from evaluation

The FACE/IDEA models for computational creativity have recently been reported [4,32,31]. These models are a potentially important contribution to the creativity evaluation literature. Consequently FACE/IDEA models will be constructed for the four case study systems so that this theoretical framework can be considered and compared as an alternative evaluation methodology alongside the other methods highlighted in this paper. It will be interesting to see how these models apply to evaluation and what the relative strengths and weaknesses of such a creativity evaluation are in comparison to existing methods (including SPECS). Results of this comparison shall be reported in [122].

The SPECS methodology has been shown above to improve 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 (Section3) rather than just product, as well as matching creativity evaluation priorities to priorities of creativity in the relevant domain. There are areas in which further work on refining SPECS, and the Section 4 componential definition of creativity, would be very useful for increasing the usability and applicability of the methodology.

There would be value in future projects examining 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 factor analysis could potentially minimise the set of components,[36] by identifying any underlying patterns and common themes within the components (factor analysis) [124]. 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 Section 4 represents words (or clusters of words) used significantly more often than expected in connection with creativity.[37] As long as important evaluative data is not lost, though, some form of reduction such as the suggestions out-

[36] Principle Component Analysis (PCA) is another dimensionality reduction technique. PCA identifies the minimal representation of data by combining and merging components, using eigenvectors and eigenvalues for dimensionality reduction. As PCA therefore does not keep the components distinct and examine the importance of components individually, this has an adverse affect on how evaluative results can be used as *formative* feedback, as the correlations between component and results have been lost,

[37] Additionally, in the context of the specific domain investigated in the case study (musical improvisation), respondents to the questionnaire about creativity in musical improvisation collectively mentioned each component at some point (to varying degrees).

lined above would make the set of components from Figure 1 more manageable to apply within SPECS, speeding up the creativity evaluation process.

The devising of empirical and/or automated tests was not attempted in the case study in this paper. If adoption of such tests had led to flawed creativity evaluation results in the case study, 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.[38]

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 creativity 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.

7.2 Future use of the SPECS methodology

It will be interesting to see how other people use the SPECS methodology. Although SPECS has been demonstrated in one case study, the case study concentrated on one specific creative domain: musical improvisational creativity. 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 the creativity of more scientifically- or mathematically-orientated creative systems for which priorities will be different to the musical systems evaluated in this paper's case study. Another application it would be intriguing to see attempted is if SPECS is adopted for evaluating people's creativity, treating people as 'creative systems'.

[38] 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.

7.3 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 creativity 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 [125].

The components in Figure 1 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. There should be no reason why the SPECS methodology should not continue to be applicable in future decades, adapting to creativity as it exists in the future.

7.4 Summary of contributions

This paper 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 paper are:

- A survey of current evaluative practice trends in computational creativity (Section 2).
- The bringing together of several different research disciplines and academic backgrounds on creativity to inform the work in this paper (Sections 3 and 4).
- A working definition of creativity in the form of a collection of key components (Section 4).
- The SPECS methodology for a *Standardised Procedure for Evaluating Creative Systems* (Section 5), with example applications to evaluate a total of four systems.
- A review and comparison of existing creativity evaluation frameworks in computational research, in theory (Section 2) and in practice (Section 6.5).

7.5 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 (Section 2), 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' [5,6] that has arisen in computational creativity, as evidenced in the survey in Section 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 creativity evaluation that could significantly stilt research progress [5,6].

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 pictured in Figure 1. The case study offers an example of how SPECS can be practically applied.

As well as being a methodological contribution, the SPECS approach to creativity 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. 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.' [19, p. 8]

For the authors of the four systems evaluated in the case study, this paper provides evaluative analysis and comparison of the creativity of their systems. SPECS has provided detailed information on the strengths and weaknesses of these nine systems, both as individual systems and in comparison to the achievements and shortfalls in other systems. The way SPECS has been implemented in the case study makes it clear what work will be most productive in terms of creativity improvements. 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 usefulness of the collected qualitative and quantitative creativity evaluation data also extends further past those whose systems have been evaluated; those conducting research in modelling musical improvisational creativity can learn from the particularly detailed focus on musical improvisation systems and the feedback collected for the four systems. For those working in other domains, as well as seeing how creativity evaluation can be carried out using SPECS, it may be helpful to see how creativity is manifested in the case study domain, for potential cross-application to their own domain. Having said that, 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.

7.6 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 (Section 3). Although this paper 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. [67,71] to far more recent investigations, e.g. [53,87]; clearly this is an ongoing issue.

As Section 3 has reported, creativity research spans several different disciplines, each with their individual priorities and foci. The work in Section 4 brings together key contributions to research from a range of disciplines including psychology, education, management, artificial intelligence. Using empirical methods from computational linguistics, common themes across this range are identified to form the collection of components presented in Figure 1. Such an overview is of interest to researchers in human creativity as well as computational creativity.

This paper 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 creativity evaluation shall not be attempted within the scope of this paper; however this posits an intriguing use of the SPECS methodology.

7.6.1 Final comments

For those of us who tackle the ill-defined and highly subjective topic that is creativity research, 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 surely needed. The Standardised Procedure for Evaluating Creative Systems is offered to meet this need.

Acknowledgements Thanks to Nick Collins for his support and supervisory input during this work, and to all the participants who took part in the case study. Communications and discussions with Alison Pease and Steve Torrance have also been extremely beneficial, as have the comments by the three anonymous reviewers of this article. The quality of computational linguistics work to derive the components of creativity reported in this paper was greatly enhanced by the collaborative involvement and knowledge of Bill Keller.

The contents of this paper are the result of doctoral research conducted at the Department of Informatics, University of Sussex, UK, who provided a doctoral stipend to partially fund this work. Some financial assistance was also received from the Sir Richard Stapley Educational Trust and the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB).

References

1. Cardoso A, Veale T, Wiggins GA. Converging on the Divergent: The History (and Future) of the International Joint Workshops in Computational Creativity. AI Magazine. 2009;30(3):15–22.
2. Ritchie G. Some Empirical Criteria for Attributing Creativity to a Computer Program. Minds and Machines. 2007;17:67–99.
3. Colton S. Creativity versus the Perception of Creativity in Computational Systems. In: Proceedings of AAAI Symposium on Creative Systems; 2008. p. 14–20.
4. Pease A, Colton S. 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; 2011. .
5. Bundy A. What kind of field is AI? In: Partridge D, Wilks Y, editors. The Foundations of Artificial Intelligence. Cambridge, UK: Cambridge University Press; 1990. p. 215–222.
6. Pearce MT, Meredith D, Wiggins GA. Motivations and Methodologies for Automation of the Compositional Process. Musicae Scientae. 2002;6(2):119–147.
7. Jordanous A. A Fitness Function for Creativity in Jazz Improvisation and Beyond. In: Proceedings of the International Conference on Computational Creativity. Lisbon, Portugal; 2010. p. 223–227.
8. Biles JA. Improvising with Genetic Algorithms: *GenJam*. In: Miranda ER, Biles JA, editors. Evolutionary Computer Music. London, UK: Springer-Verlag; 2007. p. 137–169.
9. Gillick J, Tang K, Keller RM. Machine learning of jazz grammars. Computer Music Journal. 2010;34(3):56–66.
10. Lewis GE. Too many notes: Computers, complexity and culture in Voyager. Leonardo Music Journal. 2000;10:33–39.
11. Ritchie G. Assessing Creativity. In: Proceedings of the AISB Symposium on AI and Creativity in Arts and Science. York, UK; 2001. p. 3–11.
12. Colton S, Pease A, Ritchie G. 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. .
13. Wiggins GA. A Preliminary Framework for Description, Analysis and Comparison of Creative Systems. Knowledge-Based Systems. 2006;19(7):449–458.
14. Widmer G, Flossmann S, Grachten M. YQX Plays Chopin. AI Magazine. 2009;30(3):35–48.
15. León C, Gervás P. The Role of Evaluation-Driven Rejection in the Successful Exploration of a Conceptual Space of Stories. Minds and Machines. 2010;20(4):615–634.
16. Pérez y Pérez R. MEXICA: A Computer Model of Creativity in Writing [Ph.D. thesis]. University of Sussex. Brighton, UK; 1999.
17. Colton S, de Mataras RL, Stock O. Computational Creativity: Coming of Age. AI Magazine. 2009;30(3):11–14.
18. Wiggins GA. Closing the Loop: Computational Creativity from a Model of Music Cognition; 2008. COGS research seminar, School of Informatics, University of Sussex (October 2008).
19. Pease A, Winterstein D, Colton S. Evaluating Machine Creativity. In: Proceedings of ICCBR Workshop on Approaches to Creativity; 2001. p. 129–137.
20. Peinado F, Gervas P. Evaluation of automatic generation of basic stories. New Generation Computing. 2006;24(3):289–302.
21. Pereira FC, Cardoso A. Experiments with free concept generation in Divago. Knowledge-Based Systems. 2006;19(7):459 – 470.
22. Alvarado Lopez J, Pérez y Pérez R. A Computer Model for the Generation of Monophonic Musical Melodies. In: Proceedings of the 5th International Joint Workshop on Computational Creativity. Madrid, Spain; 2008. p. 117–126.
23. Brown D. Computational Artistic Creativity and its Evaluation. In: Computational Creativity: An Interdisciplinary Approach. No. 09291 in Dagstuhl Seminar Proceedings. Dagstuhl, Germany; 2009. .
24. Chordia P, Rae A. Tabla Gyan: An Artificial Tabla Improviser. In: Proceedings of the International Conference on Computational Creativity. Lisbon, Portugal; 2010. p. 155–164.
25. Pease A. Personal communications; 2012. In conversation.
26. Ventura D. A Reductio Ad Absurdum Experiment in Sufficiency for Evaluating (Computational) Creative Systems. In: Proceedings of the 5th International Joint Workshop on Computational Creativity. Madrid, Spain; 2008. p. 11–19.
27. Tearse B, Mawhorter P, Mateas M, Wardrip-Fonin N. Experimental Results from a Rational Reconstruction of MINSTREL. In: Proceedings of the 2nd International Conference on Computational Creativity. Mexico City, Mexico; 2011. p. 54–59.
28. Gervás P. 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); 2002. .
29. Pereira FC, Mendes M, Gervás P, Cardoso A. Experiments with Assessment of Creative Systems: An Application of Ritchie's Criteria. In: Proceedings of the Workshop on Computational Creativity (IJCAI 05); 2005. .
30. Colton S. Automated theory formation in pure mathematics. Distinguished dissertations. London, UK: Springer; 2002.
31. Colton S, Charnley J, Pease A. Computational Creativity Theory: The FACE and IDEA Descriptive Models. In: Proceedings of the 2nd International Conference on Computational Creativity. Mexico City, Mexico; 2011. p. 90–95.
32. Pease A, Colton S. Computational Creativity Theory: Inspirations behind the FACE and the IDEA models. In: Proceedings of the 2nd International Conference on Computational Creativity. Mexico City, Mexico; 2011. p. 72–77.
33. Pearce M, Wiggins G. 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; 2001. .

34. Brown P. Autonomy, Signature and Creativity. In: Computational Creativity: An Interdisciplinary Approach. No. 09291 in Dagstuhl Seminar Proceedings. Dagstuhl, Germany; 2009. .
35. Gervas P. Computational Approaches to Storytelling and Creativity. AI Magazine. 2009;30(3):49–62.
36. Meehan J. Tale-Spin. In: Schank RC, Riesbeck CK, editors. Inside computer understanding: five programs plus minatures. Hillside, NJ: Lawrence Erlbaum Associates; 1981. .
37. Turner SR. The creative process: a computer model of storytelling and creativity. Hillside, NJ: Erlbaum; 1994.
38. Bringsjord S. Artificial intelligence and Literary Creativity: Inside the Mind of BRUTUS. London, UK: Lawrence Erlbaum Associates; 2000.
39. Pérez y Pérez R, Sharples M. Three Computer-Based Models of Storytelling: BRUTUS, MINSTREL and MEXICA. Knowledge-Based Systems. 2004;17(1):15–29.
40. Peinado F, Francisco V, Hervás R, Gervás P. Assessing the Novelty of Computer-Generated Narratives Using Empirical Metrics. Minds and Machines. 2010;20(4):565–588.
41. Pérez y Pérez R, Aguilar A, Negrete S. The ERI-Designer: A Computer Model for the Arrangement of Furniture. Minds and Machines. 2010;20(4):533–564.
42. Aguilar A, Hernandez D, Pérez y Pérez R, Rojas M, Zambrano MdL. A Computer Model for Novel Arrangements of Furniture. In: Proceedings of the 5th International Joint Workshop on Computational Creativity. Madrid, Spain; 2008. p. 157–162.
43. Whorley RP, Wiggins GA, Pearce MT. Systematic Evaluation and Improvement of Statistical Models of Harmony. In: Proceedings of the 4th International Joint Workshop on Computational Creativity. London, UK; 2007. p. 81–88.
44. Whorley R, Wiggins G, Rhodes C, Pearce M. Development of Techniques for the Computational Modelling of Harmony. In: Proceedings of the International Conference on Computational Creativity. Lisbon, Portugal; 2010. p. 11–15.
45. Gervás P, Perez y Perez R. 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. London, UK; 2007. p. 13–20.
46. Montfort N, Pérez y Pérez R. 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; 2008. .
47. Pérez y Pérez R, Negrete S, Penãlosa E, Aṽila R, Castellanos V, Lemaitre C. MEXICA-Impro: A Computational Model for Narrative Improvisation. In: Proceedings of the International Conference on Computational Creativity. Lisbon, Portugal; 2010. p. 90–99.
48. Plucker JA. Beware of Simple Conclusions: The Case for Content Generality of Creativity. Creativity Research Journal. 1998;11(2):179–182.
49. Baer J. The Case for Domain Specificity of Creativity. Creativity Research Journal. 1998;11(2):173–177.
50. Plucker JA, Beghetto RA. Why Creativity is Domain General, Why it Looks Domain Specific, and why the Distinction Doesn't Matter. In: Sternberg RJ, Grigorenko EL, Singer JL, editors. Creativity: From Potential to Realization. Washington, DC: American Psychological Association; 2004. p. 153–167.
51. Baer J. Is Creativity Domain-Specific? In: Kaufman JC, Sternberg RJ, editors. The Cambridge Handbook of Creativity. New York, NY: Cambridge University Press; 2010. p. 321–341.
52. Rhodes M. An analysis of creativity. Phi Delta Kappan. 1961;42(7):305–310.
53. Plucker JA, Beghetto RA, Dow GT. Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research. Educational Psychologist. 2004;39(2):83–96.
54. Sternberg RJ, Lubart TI. The Concept of Creativity: Prospects and Paradigms. In: Sternberg RJ, editor. Handbook of Creativity. Cambridge, UK: Cambridge University Press; 1999. p. 3–15.
55. Veale T, Gervás P, Pease A. Understanding creativity: A computational perspective. New Generation Computing. 2006;24(3):203–207.
56. Newell A, Shaw JG, Simon HA. The Process of Creative Thinking. In: Gruber HE, Terrell G, Wertheimer E, editors. Contemporary Approaches to Creative Thinking. New York: Atherton; 1963. p. 63–119.
57. McCarthy J. Ascribing Mental Qualities to Machines. In: Ringle M, editor. Philosophical Perspectives in Artificial Intelligence. Atlantic Highlands, NJ: Humanities Press; 1979. .
58. Wiggins GA. Searching for computational creativity. New Generation Computing. 2006;24(3):209–222.
59. Jennings KE. Search Strategies and the Creative Process. In: Proceedings of the International Conference on Computational Creativity. Lisbon, Portugal; 2010. p. 130–139.
60. Ventura D. No Free Lunch in the Search for Creativity. In: Proceedings of the 2nd International Conference on Computational Creativity. Mexico City, Mexico; 2011. p. 108–110.
61. Boden MA. The creative mind: Myths and mechanisms. 2nd ed. London, UK: Routledge; 2004.
62. Ritchie G. Uninformed resource creation for humour simulation. In: Proceedings of the 5th International Joint Workshop on Computational Creativity. Madrid, Spain; 2008. p. 147–150.
63. Cohen LM. A Review of: "Expanding Visions of Creative Intelligence: An Interdisciplinary Exploration by Don Ambrose". Creativity Research Journal. 2009;21(2-3):307–308.
64. Williams F. The Mystique of Unconscious Creation. In: Kagan J, editor. Creativity and Learning. Boston: Beacon Press; 1967. p. 142–152.
65. Albert RS, Runco MA. A History of Research on Creativity. In: Sternberg RJ, editor. Handbook of Creativity. Cambridge, MA: Cambridge University Press; 1999. p. 16–31.
66. Kaufman JC. Creativity 101. The Psych 101 series. New York: Springer; 2009.
67. Poincaré H. 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]. New York: The Science Press; 1929. p. 383–394.
68. Wallas G. The Art of Thought. abridged ed. London, UK: C. A. Watts & Co; 1945.
69. Csikszentmihalyi M. Society, culture, and person: a systems view of creativity. In: Sternberg RJ, editor. The Nature of Creativity. Cambridge, UK: Cambridge University Press; 1988. p. 325–339.
70. Resnick M. Sowing the seeds for a more creative society. Learning and Leading with Technology. 2007;35(4).
71. Guilford JP. Creativity. American Psychologist. 1950;5:444–454.
72. Mednick SA. The Remote Associates Test. Boston: Houghton Mifflin Company; 1967.
73. Torrance EP. The Nature of Creativity as Manifest in its testing. In: Sternberg RJ, editor. The Nature of Creativity. Cambridge, UK: Cambridge University Press; 1988. p. 43–75.
74. Mayer RE. Fifty Years of Creativity Research. In: Sternberg RJ, editor. Handbook of Creativity. Cambridge, UK: Cambridge University Press; 1999. p. 449–460.
75. Ivcevic Z. Creativity Map: Toward the Next Generation of Theories of Creativity. Psychology of Aesthetics, Creativity, and the Arts. 2009;3(1):17–21.
76. Stein MI. A Transactional Approach to Creativity. In: Taylor CW, Barron F, editors. Scientific Creativity: Its Recognition and Development. New York: John Wiley & Sons; 1963. p. 217–227.

77. MacKinnon DW. Creativity: a Multi-Faceted Phenomenon. In: Roslansky JD, editor. Creativity: A Discussion at the Nobel Conference. Amsterdam, The Netherlands: North-Holland Publishing Company; 1970. p. 17–32.
78. Odena O, Welch G. A Generative Model of Teachers' Thinking on Musical Creativity. Psychology of Music. 2009;37(4):416–442.
79. Clifford RD. Random Numbers, Chaos Theory, and Cogitation: A Search for the Minimal Creativity Standard in Copyright Law. Denver University Law Review. 2004;82(2):259–299.
80. Mandel GN. To Promote the Creative Process: Intellectual Property Law and the Psychology of Creativity. Notre Dame Law Review (online). 2011;86(1).
81. Feist. Feist Publications, Inc. v. Rural Telephone Service Co. 499 US 340. 1991;111 S. Ct 1282, 113 L. Ed. 2d 358(Supreme Court).
82. Karjala DS. Copyright and Creativity. UCLA Entertainment Law Review. 2008;15:169–201.
83. Copyright, Designs and Patents Act. UK Government Legislation; 1988. Ch. 48/1988.
84. Noll AM. The Beginnings of Computer Art in the United States: A Memoir. Leonardo. 1994;27(1):39–44.
85. Holmes N. The Automation of Originality: When originality is automated, what becomes of personality? Computer. 2009;March 2009:98–100.
86. Warner J. The Absence of Creativity in *Feist* and the Computational Process. Journal of the American Society for Information Science and Technology. 2010;n/a.
87. Hennessey BA, Amabile TM. Creativity. Annual Review of Psychology. 2010;61:569–598.
88. Sternberg RJ. A three-facet model of creativity. In: Sternberg RJ, editor. The Nature of Creativity. Cambridge, UK: Cambridge University Press; 1988. p. 125–147.
89. Weisberg RW. Problem Solving and Creativity. In: Sternberg RJ, editor. The Nature of Creativity. Cambridge, UK: Cambridge University Press; 1988. .
90. Bryan-Kinns N. Everyday Creativity. In: Proceedings of the 7th ACM conference on Creativity and Cognition. Berkeley, California; 2009. .
91. Torrance EP. Torrance Tests of Creative Thinking. Bensenville, IL: Scholastic Testing Service; 1974.
92. Kraft U. Unleashing creativity. Scientific American Mind. 2005;April.
93. Runco MA, Dow G, Smith WR. Information, Experience, and Divergent Thinking: An Empirical Test. Creativity research journal. 2006;18(3):267–277.
94. Kaufman JC, Kaufman SB, Lichtenberger EO. 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. 2011;26(2):83–106.
95. Boden MA, editor. Dimensions of creativity. Cambridge, MA: MIT Press; 1994.
96. Seth AK. Explanatory Correlates of Consciousness: Theoretical and Computational Challenges. Cognitive Computation. 2009;1:50–63.
97. McCrae RR, Costa Jr PT. A Five-Factor Theory of Personality. In: Pervin LA, John OP, editors. Handbook of personality: theory and research. 2nd ed. New York: The Guilford Press; 1999. p. 139–153.
98. Romero P, Calvillo-Gamez E. 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). Girona, Spain; 2011. p. 100–105.
99. Huron D. Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. Music Perception. 2001;19(1):1–64.
100. Evans V, Green M. Cognitive linguistics: An introduction. Edinburgh, UK: Edinburgh University Press; 2006.
101. Oakes MP. Statistics for Corpus Linguistics. Edinburgh, UK: Edinburgh University Press; 1998.
102. Kilgarriff A. Comparing Corpora. International Journal of Corpus Linguistics. 2001;6(1):97–133.
103. Kilgarriff A. Where to go if you would like to find out more about a word than the dictionary tells you; 2006. Published at: `http://www.macmillandictionary.com/`.
104. Lin D. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Madison, WI; 1998. p. 296–304.
105. Biemann C. 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. Morristown, NJ: Association for Computational Linguistics; 2006. p. 73–80.
106. Jordanous A. Evaluating Evaluation: Assessing Progress in Computational Creativity Research. In: Proceedings of the Second International Conference on Computational Creativity (ICCC-11). Mexico City, Mexico; 2011. .
107. Boden MA. What is Creativity? In: Boden MA, editor. Dimensions of Creativity. Cambridge, MA: MIT Press; 1994. p. 75–117.
108. Moffat DC, Kelly M. 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; 2006. .
109. Haenen J, Rauchas S. 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; 2006. .
110. Pearce MT, Wiggins GA. Evaluating Cognitive Models of Musical Composition. In: Proceedings of the 4th International Joint Workshop on Computational Creativity. London, UK; 2007. p. 73–80.
111. Lewis GE. Improvising With Creative Machines: Reflections on Human-Machine Interaction (keynote talk). In: Proceedings of the 2nd International Conference on Computational Creativity. Mexico City, Mexico; 2011. p. xii–xiii.
112. Parker E. Drifting on a Reed (Keynote presentation). In: The Improvised Space: Techniques, Traditions and Technologies. London, UK; 2011. .
113. Csikszentmihalyi M. The Creative Person and the Creative System (keynote address). In: Proceeding of the seventh ACM conference on Creativity and cognition. Berkeley, California; 2009. p. 5–6.
114. Friis-Olivarius M, Wallentin M, Vuust P. Improvisation - the Neural Foundation for Creativity (poster). In: Proceedings of the 7th ACM Creativity and Cognition Conference. Berkeley, California; 2009. p. 411–412.
115. Berkowitz AL, Ansari D. Expertise-related deactivation of the right temporoparietal junction during musical improvisation. NeuroImage. 2010;49(1):712–719.
116. Berliner PF. Thinking in jazz: the infinite art of improvisation. Chicago Studies in Ethnomusicology. Chicago, IL: The University of Chicago Press; 1994.
117. Gibbs L. Evaluating Creative (jazz) Improvisation: Distinguishing Invention and Creativity. In: Proceedings of Leeds International Jazz Conference 2010: Improvisation - jazz in the creative moment. Leeds, UK; 2010. .

118. Bailey D. Improvisation: Its nature and practice in music. New York: Da Capo Press; 1993.
119. Biles JA. GenJam: A Genetic Algorithm for Generating Jazz Solos. In: Proceedings of the International Computer Music Conference. Denmark; 1994. .
120. Jordanous A. Defining Creativity: Finding Keywords for Creativity Using Corpus Linguistics Techniques. In: Proceedings of the International Conference on Computational Creativity. Lisbon, Portugal; 2010. p. 278–287.
121. BNC Consortium. The British National Corpus, version 3 (BNC XML Edition); 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: *http*: *//www.natcorp.ox.ac.uk/*.
122. Jordanous A. Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application [Ph.D. thesis]. University of Sussex. Brighton, UK; forthcoming.
123. Zhu X, Xu Z, Khot T. 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). Boulder, Colorado; 2009. p. 87–93.
124. Sauro J, Kindlund E. 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; 2005. .
125. Temperley N, Wollny P. Bach Revival; 2011. Available at Grove Music Online (Oxford Music Online): *http*: *//www.oxfordmusiconline.com/subscriber/article/grove/music/*01708 (retrieved 13th August 2011).