PhD IN COMPUTER SCIENCE

ON THE POSSIBILITY OF RECALLING WITHOUT SEEING:

EVIDENCE FROM STATE-TRACE ANALYSIS OF THE EXPERIENTIAL BLINK.

BY WILLIAM JONES

SUPERVISED BY PROFESSOR HOWARD BOWMAN

49805 words over 199 pages (Date: 19 Sept 2019)

Abstract

In recent years, there has been much debate about how our conscious perception of the world relates to our ability to store and process information about it. Given the tight coupling between these two processes, it is not surprising that the nature of this relationship has proved difficult to establish with any degree of certainty. Recent findings however, have provided an opportunity to quantify evidence about how such a relationship might manifest. In this thesis, we follow up on one particular such set of findings: examining the relationship between participants' subjective experience of stimuli and their ability to encode them into working memory during the attentional blink.

This problem is tackled in three progressive steps. Firstly, we attempt to establish with certainty that a difference does exists between these two cognitive processes. In order to quantify the distinctness of the two cognitive processes, we make use of state-trace analysis. Having established that the two cognitive processes are in some way distinct, we examine more closely what form their relationship takes; what kind of relationship of dependency exists between the two measures, is it possible to have one without the other? Finally, we attempt to provide a theory and computational model of the above results.

Our findings provide evidence that working memory encoding and subjective experience are dissociated in some manner. Further examination yields evidence that it is possible that working memory encoding may exist as a necessary but insufficient condition for subjective experience. We develop a theory of this behaviour based on targets being encoded simultaneously, but only experienced in serial, and build a computational model of these results by integrating with an existing model – the Simultaneous Type/Serial Token model of attention. The predictions this model makes strongly match those observed in human participants.

Table of Contents

1. Introduction

One of the motivations of modern neuroscience is to understand how the physical matter of the brain gives rise to the complex spectrum of behaviours we observe from it. Creating suitable tools by which we can quantify all of these behaviours as we understand them however, is a challenge. In this thesis, we are concerned with one particularly challenging area: conscious perception or subjective experience. To be specific, what we are precisely interested in in this thesis is what a participant internally experiences of an external stimulus. For the avoidance of any doubt and for the sake of a consistent terminology, we are from now on going to refer to this phenomenon as *subjective experience*. Subjective experience is a challenging area because it is not a directly observable phenomenon. What a participant experiences of a stimulus is only available to them directly and unlike, for example, whether that stimulus has been encoded into working memory, is not directly amenable to an external test.

In light of these difficulties, it is tempting to either write off subjective experience as impossible to study (Chalmers 1995) or as a phantom that arises out of our own ignorance of the subject matter (Hacker 2010). However, neither of these options will prove fruitful to neuroscience in the long term. Simply giving up is not an option, and the only way neuroscience will progress beyond these ideas if they are incorrect is to rise to the challenge of them and to understand why. In this thesis, we explore this subjective experience through comparison to a more directly measurable phenomena – working memory encoding. In particular, it is widely agreed that working memory and subjective experience are closely related, but there are many instances in which they appear to be in some way separable (Velichkovsky 2017). In light of this, by establishing how working memory encoding and subjective experience are related or co-dependent and most importantly, how they are not, we can inform not just the existing debate on their seperability, but contribute to the understanding of what underlies each process.

We propose that this examination should take place in three stages. Firstly, any constructive comparison of the two cognitive processes presupposes that the two are separable in some way. As we will discuss, there is some evidence that this is the case, but our first step should be to establish this with high confidence. Secondly, if we can establish the distinctness of the two processes, we should

then examine in more detail the relationship between the two. It is clear that the measures are closely related in some way, but to what extent are they coupled together? For example, is it possible to elicit cases in which working memory encoding occurs in the absence of subjective experience, or subjective experience in the absence of working memory, or both, or neither? Finally, if we can establish evidence for some relationship between the two measures, we must then attempt to provide a theory and model of our results. While our findings may be interesting in and of themselves, these results are only as useful as they are interpretable in the context of other work. Developing a theory and modelling it provides an excellent opportunity to be critically assessed in a broader context.

Research

We now discuss how the research chapters of this thesis approach this problem. Replication of much of the work done, using an entirely separate dataset, is available in Appendix E.

State-trace Analysis of the Attentional Blink

In this chapter, we probed the first question that any constructive examination of the relationship between working memory encoding and subjective experience presupposes – that the two are separable. What we were attempting to quantify was the functional distinctness of the two processes; i.e. that the two are not mutually dependent. Based on previous literature, we identified an appropriate paradigm over which to assess this distinction was the Attentional Blink.

A standard approach for demonstrating this distinctness is to look for *functional dissociations*. These arise when it is possible to independently modify behaviour on different tasks that embody each of our different processes. Unfortunately, as we will discuss in the literature review, recent research (Bogartz 1976, Dunn, Kirsner 1988, Henson 2006, Davis-Stober, Morey et al. 2016) has provided evidence that such methods may not be as robust as originally thought. We therefore adopted a more recently developed methodology known as state-trace analysis that does not have the same vulnerabilities.

Unfortunately, one of the weaknesses of this state-trace method is that, while it is excellent for evaluating the presence of dissociations, it is not possible on the basis of this state-trace analysis on its own to come to any conclusions about

the properties of these dissociations. In this chapter we therefore further developed a “post hoc” method that allowed us to assess which data points contributed to the dissociation found most strongly. This, combined with existing results (Pincham, Bowman et al. 2016), allowed us to come to some conclusions about the nature of the relationship between working memory encoding and subjective experience.

Methods in State-Trace Analysis

While we took care to validate the methods used in the previous chapter, the state-trace analysis we performed highlighted several areas in which the technique could be improved. Primary among these was the creation of an appropriate prior for the Bayesian methodology upon which the quantification of our state-trace analysis is based. In this chapter, we therefore developed a new method for taking an existing prior, and modifying it to more accurately reflect our current dataset. To prevent this influencing our results, the method made use of a contrast that is independent from our hypothesis of interest. After validating this method with test data, we then applied it to the state-trace analysis performed in the previous chapter.

Modelling Subjective Experience

Having established that it is possible to dissociate working memory encoding and subjective experience, and having gained some evidence about the nature of the relationship of these two cognitive processes, we attempted to establish a working theory and model of the results. Broadly, our theory was that the working memory encoding of multiple targets was able to progress simultaneously, but that the experience of the same targets could only occur in serial.

There were several directions we could have taken the modelling of these results, but we opted to build our theory into an existing model. Our model of choice was the Simultaneous Type/Serial Token model (Bowman, H., Wyble 2007). It already naturally dealt with the simultaneity/seriality dichotomy and was able to provide both behavioural and electrophysiological predictions. Having created this model, we validated our findings by comparing the results from human data to predictions made from the model.

Meta-experience in the Attentional Blink

In this chapter, we re-examine a weakness of our original analysis – that many of our conclusions are based on analysis performed on averages, but such averages do not fully characterise what occurs. While this is unlikely to compromise our existing conclusions, it opens up an interesting question –how does subjective experience match up to working memory encoding on a trial to trial basis?

The question we are fundamentally addressing here is *metacognition*: how well our subjective reports reflect success at encoding targets into working memory. This is a topic that has been explored extensively in the metacognition literature, with several tools developed for quantifying this relationship. Unfortunately, none of these existing tools are suitable for assessing our dataset, and as part of this chapter we develop a new measure for metacognition.

A complication of nomenclature arises when applying this method to our own data however, as we explicitly capture measures of *subjective experience*, instead of measures of *confidence* that a true metacognitive measure requires. To avoid confusion we introduce a new term for the "metacognition" calculated across our own data that uses subjective visibility ratings instead of confidence reports: *meta-experience*.

We apply this method to determine how meta-experience changes during the attentional blink. We use these new findings to validate the model developed in the previous chapter. This further informs our understanding of the working memory encoding/subjective experience relationship.

Core hypothesis

To summarise, we wish to assess the relationship between working memory encoding and subjective experience. This assessment takes the form of three research hypothesis that progress, one from another:

1) *Working memory encoding and subjective experience can be dissociated*
2) *Assuming working memory encoding and subjective experience are dissociated, the relationship of the dependency between the two (if any) between these two processes can be established*

3) *Assuming a relationship of dependency (or lack thereof) between working memory encoding and subjective experience can be established, this can be modelled (in the Simultaneous Type/Serial token model of attention)*

In this chapter, we have covered the broad motivation for the thesis, the research questions it will attempt to solve, and a summary of how each of our research chapters is going to address them. We now move on to the first major section of the thesis, the literature review, in which we will cover the major theoretical background to the work that will be undertaken.

2. Literature review

Consciousness

In this thesis, a term that has been and will be used extensively is *consciousness*. Unfortunately, the word *consciousness* is a somewhat ambiguous term whose definition has been co-opted into defining a very large number of different concepts (Zeman 2005), many of which are just as ill-defined as the word consciousness itself (Antony 2001). As a starting point, we will borrow the definition from the Stanford Encyclopaedia of Philosophy (Zalta, Nodelman et al. 2005) "The state of quality of being aware of an external object or something within oneself", though it will quickly become apparent that such a definition is inadequate. Even from this starting point however, it is easy to see why consciousness has been a topic of fervent study and debate; even this apparently uncertain and incomplete definition strikes at the heart of our understanding of ourselves. However, while it is clearly desirable to measure something so important, we must proceed carefully to avoid confusion arising over this multifaceted concept.

In this subchapter, we review some of the existing literature around consciousness and how it relates to our working memory encoding/subjective experience dichotomy. Before we begin however, we wish to make clear one point of nomenclature. One concept referred to extensively in the discussions about consciousness in the literature is the internal experience of an external stimulus, what a participant internally experiences of an external stimulus presented to them. Some authors, notably (Pincham, Bowman et al. 2016) whose data we extensively make use of, discuss the internal experience of an external stimulus as *conscious experience* or *conscious perception*. For consistency, and for the avoidance of doubt, in this work we will refer to this as *subjective experience* unless directly quoting another work. Further, we refer to any measure of this subjective experience as *subjective report*.

The Easy and Hard Problems of Consciousness

The problem of defining and measuring consciousness has been widely considered. One framing of the problem that has gained a lot of traction is the partitioning of the problems of consciousness into two distinct groups – the *easy*

problems of consciousness and the *hard* problems of consciousness (Chalmers 1995). The easy problems are those that are vulnerable to typical quantitative analysis, specifically:

"[…]are those [problems] *that seem directly susceptible to the standard methods of cognitive science, whereby a phenomenon is explained in terms of computational or neural mechanisms"*

Whereas the hard problems are those problems that are not. The classic example of a hard problem is that of *qualia*, the phenomenal character of an item - what it is like to have an experience of a stimulus; for example, ones internal experience of the "redness" of red.

While this distinction is interesting, it has caused a lot of debate. Criticisms of Chalmers work have generally come from the viewpoint that the hard problem is for some reason or another either a false dilemma (Kalat 2014, Hacker 2010) or is not truly hard (Dennett 2000), or they have specifically attacked the dualist view that is somewhat central to his argument (Carruthers, G., Schier 2017). The theory has also been criticised based on a similar intuition to the one given in the introduction to this section – Dehaene (Kalat 2014) has argued that the hard problem only arises because of our insufficient understanding of what consciousness is, and that with evolved understanding, the problem will disappear. Hacker (Hacker 2010) has made a somewhat similar, though perhaps more vitriolic argument:

"*The whole endeavour of the consciousness studies community is absurd – they are in pursuit of a chimera. They misunderstand the nature of consciousness. The conception of consciousness which they have is incoherent. The questions they are asking don't make sense. They have to go back to the drawing board and start all over again. It's literally a total waste of time."*

Regardless on which side of the debate one stands, this framing of the problem of consciousness brings forth the distinction between working memory encoding and subjective experience we made in our introduction. Whether or not a stimulus is, for example, encoded into working memory is a quantity that is potentially different from how a person internally experiences a stimulus.

Subjective vs Objective Measures of Consciousness

This working memory encoding/subjective experience debate has been previously discussed extensively in the literature, though in terms of the differentiation between the “phenomenological awareness" of the phenomenal character (subjective experience) of a stimulus versus the “access consciousness" (working memory encoding) of a stimulus that is available for report (Block 1995).

One well debated viewpoint on the relationship of these two quantities is that phenomenological awareness *overflows* conscious access in the sense that we internally experience more of stimuli than is available for objective report later. Put another way, phenomenological awareness is a necessary but not a sufficient condition for conscious access. Block (Block 1995) is a notable proponent of this viewpoint, initially on the basis of the Sperling paradigm. The Sperling paradigm (Sperling 1960) (Figure 1) measures the recall of participants in two conditions. The first condition is one in which a grid of letters are presented for 500ms, followed by a blank screen. Participants are then asked to report all the letters they can. Typically, participants can report 4. The second condition is identical to the first, but after the grid has been removed, participants are cued on a row to report. Participants still managed to recall 4 items, but those items will be on the cued row. Importantly, though participants are only ever able to report 4 items, they subjectively feel that they can see the whole grid. Block cites this (Block 2007) as evidence that the participants are on some level richly experiencing the whole grid, but are unable to report all of it.

Dehaene (Dehaene, Changeux et al. 2006) has contested the validity of this experiment on the basis of the change blindness paradigm. The change blindness paradigm uses an alternating series of images that are subtly different from one another to demonstrate that participants will not notice small changes between the pictures (Simons, Rensink 2005). It is argued that change blindness shows that participants are overconfident on their ability to report, and instead of seeing a scene, they often merely suffer from the illusion of seeing because they know they can reorient their attention to any part of the scene at a given moment to obtain information about it (Dehaene, Changeux et al. 2006). This result has led some to conclude that participants are not actually experiencing the whole grid

as they subjectively feel in the Sperling paradigm, and that, at any given moment, very little of the scene is actually being processed – a *sparse* experience. (De Gardelle, Sackur et al. 2009) have also demonstrated a modified Sperling test that introduces the presence of unexpected "letter-like" pseudo letters into the grid. Despite the introduction of these pseudo-letters, participants still felt that they were experiencing a grid consisting entirely of letters. In light of these results, the authors propose that the "illusion of seeing" is the result of our expectations of what we will see and partial information about the scene. Block has criticised the validity of the results of this modified experiment (Block 2011) on the basis of low contrast stimuli and the introduction of a backward mask. He cites Fragile Visual Short Term Memory (a four seconds lasting fragile VSTM store with a capacity that is at least a factor of two higher than robust VSTM) (Sligte, Scholte et al. 2008) as one which demonstrates something similar without the same problems – and incidentally as one that supports a rich experience.

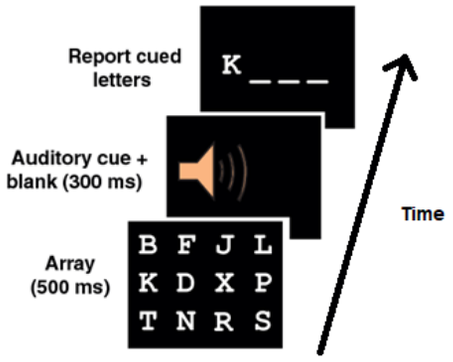


Figure 1) An example of the Sperling paradigm (Sperling 1960). Adapted from (Kouider, De Gardelle et al. 2010) who perform a modified version of the experiment. An array of letters is presented for 500ms. Participants are then either presented an aural cue which indicates to the subject to focus report on a certain row, or allows to report freely. In both instances, participants will be able to report approximately 4 stimuli, but when cued on a row to report, those 4 stimuli will be in that row.

In further support of a rich experience, (Vandenbroucke, Fahrenfort et al. 2014) record fMRI during an inattentional blindness task based on Kanizsa figures (an optical illusion). Notably, the Kaniza illusion requires conscious processing of its inducers to be recognised (Harrison, Tong 2009) and the neural signature unique to the processing of the Kanizsa was present in both those who were inattentionally blind and those who were not. This is argued to indicate what has been claimed by some for a while, that during inattentional blindness the unreported stimuli are perceived, but are simply not accessed (Vandenbroucke, Fahrenfort et al. 2014). Bronfman et al. (Bronfman, Brezis et al. 2014)

demonstrate that participants in another modified Sperling experiment still retain some information about unattended rows. Specifically, the researchers found that participants could still accurately report on a high/low colour variance of separately cued rows.

This large body of literature debates whether phenomenological consciousness is a necessary condition for access consciousness. The dual question, whether access consciousness is a necessary condition for phenomenological experience has been less discussed. The existence of such a phenomenon is debatably less intuitive – it implies that we can somehow access and report stimuli that we at no point subjectively experienced. Despite this, there is a small body of literature that has probed the subject.

An example (Block 1995) from the original paper that spawned the Phenomenal-Access consciousness distinction that attempts to show that such a phenomenon is plausible is a thought experiment around an extreme form of blindsight. Patients with blindsight have a damaged visual cortex and are consciously blind in part of their visual fields, but under some circumstances will respond to visual stimuli that they cannot "see" (Humphrey 1974). We are asked to consider the possibility of a "Superblindsight" patient that, unlike a normal Blindsight patient, can prompt themselves to what is in their blind field and evaluate it. Such a Blindsight patient would have two different experiences of a view item: the experience of knowing through experience and "Just knowing" what is in their blind field – thus access consciousness without phenomenological consciousness. (Lamme 2001) also discusses blindsight as an indication that conscious awareness and conscious access are separable. Block himself admits that this idea is only a thought experiment, but others (e.g. (Bogen 1997)) have proposed a case that is potentially more realisable, the case of split brained patients. Split brain patients are those literally with a split brain – their right hemisphere is separated from the left. (Bogen 1997) enumerates several facts about such patients:

"(1) in most of these patients speech is produced only by the left hemisphere, (2) the speech is evidence that P [Phenomenal] and A [Access] coexist in that hemisphere, and (3) verbal denial of information that has been delivered only to the right hemisphere (and rationally acted upon) reflects the existence of an

independent capacity in the right hemisphere, that is, an A-consciousness different from the A-consciousness of the left hemisphere."

Since the right hemisphere has its own independent access consciousness, this begs the question: does it also have its own independent phenomenal consciousness? If it does, it would be a very persuasive and replicable argument for access consciousness without phenomenal consciousness.

Further to this, there are also several other paradigms that would seem to show some kind of access to stimuli for which there has been no phenomenal experience. For example, flash suppression (Hsieh, Colas et al. 2011), visual masking (Van den Bussche, Hughes et al. 2010), or episodic face recognition (Heathcote, Freeman et al. 2009). However, we would argue that these paradigms provide evidence for a weaker claim: that of *influence* without experience. In every case, the identity of the un-experienced stimulus is not directly reportable, it merely influences the report of, or response to, something else. One result that does seem to demonstrate a direct report is a study by (Pincham, Bowman et al. 2016), which demonstrates (among other results) an apparent *free recall* of stimuli in the absence of conscious awareness. The authors find that participants report stimuli at high level even when reporting zero subjective experience (Note, this result is found in the appendix of the paper, not the main body). A somewhat similar result also arises from studies by (Soto, Silvanto 2014) and (Trübutschek, Marti et al. 2017), which explores the possibility of stimuli being maintained in working memory in the absence of any subjective experience. However, we note that while these results are compelling, they are about a slightly different question – working memory maintenance, rather than encoding.

Measuring subjective and objective report

So our discussion has been in abstract terms such as the easy/hard problems, phenomenal/access consciousness and working memory encoding/subjective experience. While these are useful tools for thinking about the problems of consciousness, in order to apply the scientific method we need these to be relatable to some measurable phenomena.

Describing a measure for working memory encoding is simple. To draw upon the easy/hard dichotomy, working memory encoding is an easy problem. It is directly amenable to the traditional methods of cognitive neuroscience. Assessing

whether a stimulus has been encoded into working memory can be done by simply asking a participant to report on the result, and comparing this report to the stimulus that was presented. We term these types of measures *objective,* as in contrast to measures of the harder problems such as internal experience that are by definition *subjective* measures.

In contrast, measuring subjective experience is significantly more difficult. To borrow the easy/hard dichotomy again, Overgaard (Dienes, Z. 2015) writes about the search for such measures as the equivalently hard problems of empirical consciousness research. It is difficult to attempt to evaluate the “best" third party measure of an internal state that is, by definition, only available to the participant themselves. One approach is simply to ask participants to report on their own state. This report might be on a dichotomous scale (Lamy, Salti et al. 2009) to maximise statistical power, a larger scale with more than 2 bins (Overgaard, Rote et al. 2006) or a more continuous scale (Sergent, Dehaene 2004). Regardless of the method, this must be done carefully as participants often have difficulty using poorly designed scales, or those with many options (Sandberg, Timmermans et al. 2010).

Such measures often run into difficulties, but are still widely used. On one hand, given that the measure is fundamentally introspective, asking participants about their percept is potentially the most effective approach. On the other hand, it has been argued as far back as (James 1898) that any such subjective measure must be retrospective - and therefore subject to memory effects. Furthermore, subjective report scales can also be subject to response bias (Timmermans, Cleeremans 2015). In particular, subjects may withhold reports simply because they are not confident about them instead of because they have zero subjective experience (Garner, Hake et al. 1956), or make up responses when questioned (Nisbett, Wilson 1977). Subjective report is also known to be subject to changes in instruction (Overgaard, Sorensen 2004). An alternative that has been proposed is to probe subjective experience by asking participants to wager on the correctness of their response (Persaud, McLeod et al. 2007). This is proposed to be a more direct measure of subjective report. However, its effectiveness has been debated. It has been argued that it is no more effective than questioning

participants directly, and is known to be subject to participants risk aversion (Dienes, Zoltán, Seth 2010).

We have now covered some of the literature behind the working memory encoding/subjective experience dichotomy presented in the introduction, and some methods by which these might be evaluated. In terms of the research question of this thesis however, we now need to go one step further: despite introducing and discussing many phenomena that seem to show a separation of subjective experience and working memory encoding, we have not yet discussed a formal mechanism by which this can be demonstrated. These separations have canonically been explored by trying to find functional dissociations (though recent literature has proposed improvements), a concept we now cover in more detail.

Dissociating of Cognitive Processes

The question of "modularity of mind" - that is, to what degree the mind is made up of functionally independent components - has been of great importance in the fields of philosophy, psychology and neuroscience. While there are many diverse views, e.g. (Fodor 1983), (Carruthers, P. 2006) and (Prinz 2006), a large number of contemporary models and theories incorporate modularity in some way. This makes the ability to separate functionally independent mental processes and thus to be able to demonstrate this modularity - or lack thereof – critical to modern cognitive neuroscience. As we saw in the previous section, one of the areas over which the independence of processes is particularly contested is the one posed by this, particularly whether subjective experience of the character of a stimulus (the "*phenomenological awareness*" of it) and the ability to encode it into working memory for retrieval (the "*access consciousness*" of it) are distinct.

Tackling such problems is usually performed by looking for *functional dissociations*. When we say two mental processes are functionally dissociated, we specifically mean that they are in some way functionally independent from one another. To put it another way, there is no complete relationship of dependency between the two. i.e. neither is both a necessary and sufficient condition for the other to occur. Evidence for such a dissociation is seen to arise when we find variables that allow us to independently modify performance on two separate tasks, providing putative evidence that the cognitive processes

embodied by the tasks are in some way separate (Dunn, Kirsner 1988). Such dissociation logic has been widely applied, and made an important contribution to the investigation of functional independence in the mind in such diverse sub-fields as short term memory (encompassing working memory as one part of this) and long term memory (Warrington 2014), word comprehension (Cousins, York et al. 2016) and consciousness (Cohen, Cavanagh et al. 2012). Despite their merits, some authors have argued that functional dissociations can be improved upon (Bogartz 1976, Dunn, Kirsner 1988, Henson 2006, Davis-Stober, Morey et al. 2016), and we will discuss these results and proposed alternatives extensively further below.

Functional Dissociations

The functional dissociation is a technique that has been widely implemented across the fields of psychology and neuroscience as a marker of the functional distinctness of mental processes. There are several types of functional dissociations, but all arise when one is able to independently modify performance on a set of one or more tasks in a set without affecting performance on other tasks in the set. The ability to differentially affect behaviours on different tasks is seen as evidence that the mental processes underlying them are in some way functionally separate. There exist three types of dissociations widely used in the literature (Shallice 1988): single dissociations, uncrossed double dissociations and crossed double dissociations.

A single dissociation is the simplest example. In this instance, we have a manipulation in which one independent variable is modified, and there are two different tasks over which performance is assessed. In a single dissociation, we find that it is possible to increase performance on one task without affecting performance on the other. This is seen to provide evidence that the two cognitive processes underlying the tasks in question are functionally dissociated. However, it has long been known (Teuber 1955) that single dissociations only constitute weak evidence of a functional dissociation. Indeed, the strongest conclusion that it is possible to come to from a single dissociation is that the cognitive processes underlying the task are *not the same*. The possibility still exists that one is either a necessary or sufficient condition for the other to occur.

The problem is that there is no reason to believe that increased (or decreased) cognitive function translates well, or at all, to increased or decreased task performance (Loftus 1978); classic examples of how this might occur are floor or ceiling effects. Floor or ceiling effects occur when experimental limitations prevent performance from getting better or worse respectively (Krantz, Tversky 1971). For example, we may ask participants to identify a stimulus in a stream of distractors. Depending how rapidly the stream is presented, we would expect performance to increase or decrease, with faster streams resulting in worse performance. However, beyond a certain point, performance will either be at nil or perfect, and changes in the presentation rate will no longer affect performance. This is a classic floor (or ceiling) effect. Because of these issues, it is generally proposed that authors should not attempt to use single dissociations as a marker of functional dissociation, but should instead use a double dissociation (Shallice 1988). There are two kinds of these dissociations we discuss, crossed and uncrossed double dissociations.

An uncrossed double dissociation is simply the occurrence of two, opposite single dissociations (Dunn, Kirsner 1988). Instead of finding a single manipulation over which we modify one independent variable, we find two different manipulations. Once again, we have two tasks over which performance is assessed, say task A and task B. An uncrossed double dissociation arises when modifying the independent variable of one of the manipulations changes performance on task A while performance on task B stays the same, while doing the same to the other independent variable does the opposite – performance on task B changes while task A performance is static. Strictly, a double dissociation does not remove the issues that a single dissociation demonstrates, but in practice, with good experimental design, it makes them much less likely (Dunn, Kirsner 1988). However, there is no way to preclude effects of this kind entirely, therefore despite the merits of an uncrossed double dissociation, where possible a crossed double dissociation is often preferred (Shallice 1988).

Crossed double dissociations is a special case of the uncrossed double dissociation in which it is possible to increase performance on one task while simultaneously decreasing performance on the other through modifying the levels of only a single manipulation. This results in a crossover interaction of the

levels of this variable and the two task performances. Since there is no static task performance, a dissociation of this kind is not vulnerable to claims of floor or ceiling effects, though it is still subject to the generalised problem that the existence of these effects demonstrates (Newell, Dunn 2008), which we will now discuss in more detail.

Dissociations – Failure Conditions

In order to properly discuss the problems with dissociation logic, we will formalise slightly (for an extensive theoretical treatment, see (Bamber 1979)). Let our experimental manipulations be the set of independent input variables $\mathrm{i_p}$. These are then related to cognitive function, our latent variables $\mathrm{l_q}$, through functions g_n. This cognitive function is then related to our task performance, our dependent output variables $\mathrm{o_r}$ through functions f_m. See Figure 2 for an example. Inside this framework, the problem of demonstrating functional dissociations comes down to determining whether a single latent variable is sufficient to explain how the pattern of independent input variables relates to the levels of the dependent output variables. If a single variable is sufficient then we may conclude that there is no dissociation (Figure 2A), but if it is not, a dissociation is required (Figure 2B)).

Framed in this way, our floor and ceiling effects from the last section are simply regions of the functions f_m where the output function does not change. Figure 3 shows why this is problematic, with a slightly more complicated example of the general result that floor effects demonstrate. The models are polar opposite in terms of the dependency of the output variables, in A) they are mutually dependent and in B) they are independent. However, because of the nonlinearity of B that somewhat resembles a floor effect, A) and B) are almost indistinguishable for almost any range of inputs. The levels of the second dependent variable may equally be changing because they are functions of two independent variables, or because they are functions of the same independent variable that is not responsive over this range for the second output. This is problematic because dissociation logic would always come to the conclusion that a dissociation existed.

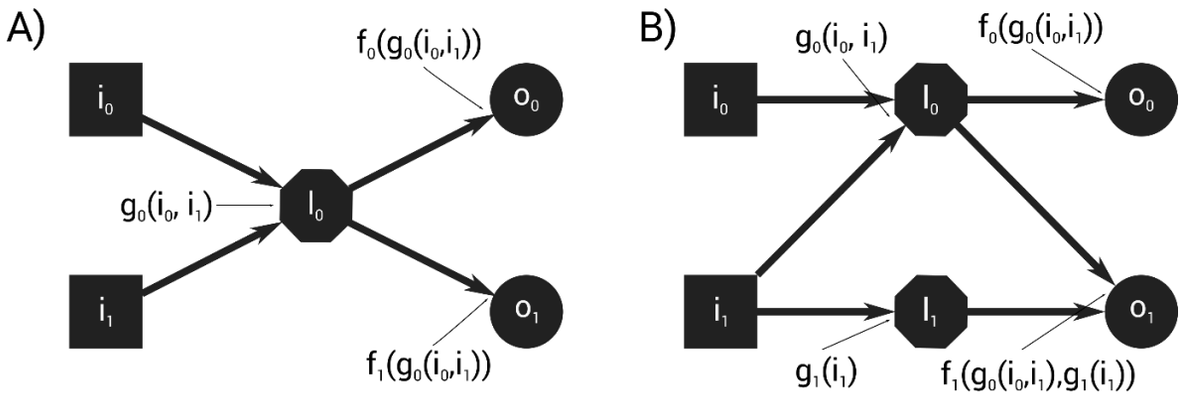


Figure 2) Examples of two different systems under the framework described above. i_p are our independent input variables, l_q our set of latent cognitive functions and o_r our set of dependent output variables. g_n are the functions relating our input variables to our latent variables and f_m our functions relating our latent variables to our output variables. A) A potential system in which no dissociation exists. Both of our dependent variables are reliant on the same underlying latent cognitive process and so no functional dissociation between the two exists. B) One potential system in which a dissociation does exist. In this instance the level of our output variables is determined by two different cognitive processes.

This brings us to an interesting problem of the dissociation logic we discussed in the previous section. These methods are actually making quite a strong implicit assumption within the framework we have created here – that the functions f_m are approximately uniform in their behaviour. Floor or ceiling effects are one example of behaviour that contradicts this but many have asserted that there is no reason to believe that more subtle behaviours cannot occur (Dunn, Kirsner 1988, Bamber 1979, Prince, Brown et al. 2012). (Prince, Brown et al. 2012) describe these effects as *scale dependent interactions*, in which nonlinearity in the translation between latent and dependant variable can occur and (Loftus 1978) details several accounts of the effects of different kinds of response functions that account for bounds on the dependant variables. For these reasons, despite their wide use in the literature, it has been argued that while functional dissociations are certainly indicative, they do not strictly provide either a necessary or sufficient basis for determining the separation of mental processes (Bogartz 1976, Dunn, Kirsner 1988, Henson 2006, Davis-Stober, Morey et al. 2016). As we have seen in Figure 3, it is possible to construct cases in which we can create the types of dissociations we have discussed without separate mental processes, and it is easy to see that similarly it is possible to do the opposite. The question that of course arises in light of this is, if not through dissociation logic, how can we separate the cases in which a single latent variable is required from

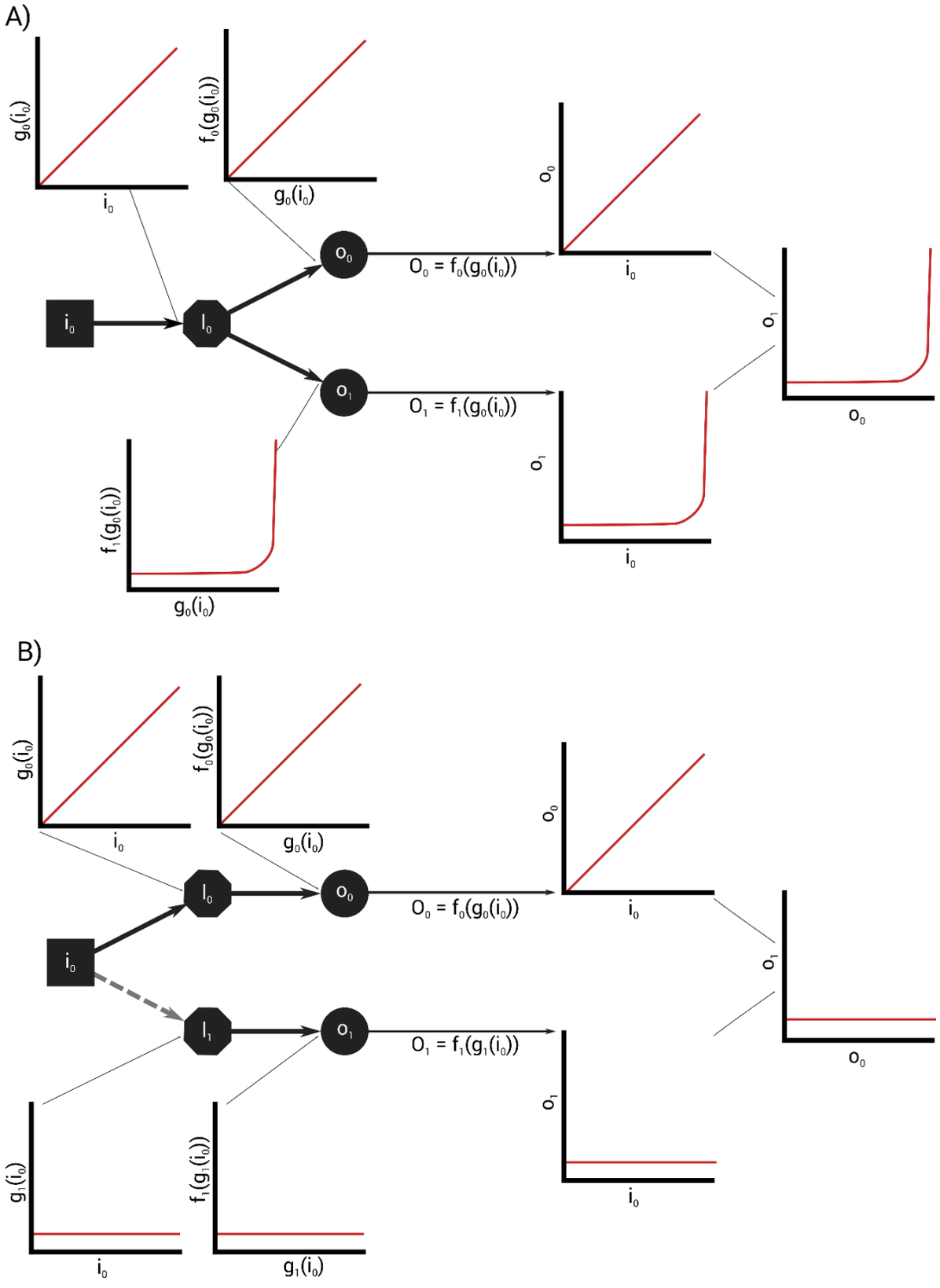


Figure 3) Example of how the scale dependent interactions described by (Prince, Brown et al. 2012) can cause traditional dissociation logic to reach erroneous conclusions. In this instance, we have two systems – A) in which a dissociation does not exist and B) in which one does. In A) the output variables are reliant on the same underlying cognitive process, but in different ways. In B), the output o_1 is completely unrelated to the input i_0 even indirectly, no change in i_0 can affect it. Conversely, o_0 is directly related to i_0 through the latent variable l_0– this gives us a classical dissociation. The problem arises because o_1 in A) is highly insensitive over most of the range of i_0

that we are sampling. For almost the entire range of sampled i_0 values, the behaviour of A), a system without a dissociation and B) a system with a strong dissociation will be almost entirely indistinguishable based on the behaviour of o_0 and o_1. Despite this, dissociation logic would conclude B) – that a dissociation exists.

the ones in which multiple are required. Fortunately, this question has been tackled, through a tool known as state-trace analysis.

State-Trace analysis

State-trace analysis is a tool for testing the validity of a number of different theoretical models for how a system of latent variables modulates the observed relationship between the independent variables and dependent variables. This is quite a powerful tool with broader applications than we specifically use here (Loftus 2002). We are interested in one facet of state-trace analysis in particular – distinguishing whether a single latent variable is sufficient to explain the difference between two different dependent variables. This case is known as *dependent variable* state-trace analysis (Prince, Brown et al. 2012). This explores the specific hypothesis that a single latent variable is sufficient to explain our pattern of results. We have described previously the difficulties associated with dissociation logic, scale dependent interactions. Our problem arises because these "scale dependent interaction" functions f_n are effectively unconstrained. Though there may be choices of f_n that are theoretically unpalatable, we have not set any explicit bounds on then. Because of this, it is impossible for us to know whether any behaviour that we see is arising, for example, as the result of one or many latent variables or simply as a result of a particular f_n.

State-trace analysis solves this by constraining the relationship between latent cognitive functions of dependant variables f_n to be at least monotonic. Given this assumption, then for a model containing only a single latent variable, the relationship between our two task performances cannot fail but to be monotonic. Any violation of this in the observed data makes the data logically inconsistent with a single latent process model, and thus demonstrates a functional dissociation. See Figure 4 for an example of this monotonicity, and how non-monotonicity might arise with more than one process. Note that the opposite is not true; a model that contains two latent variables does not necessarily cause a non-monotonic relationship between task performances. A monotonic model is

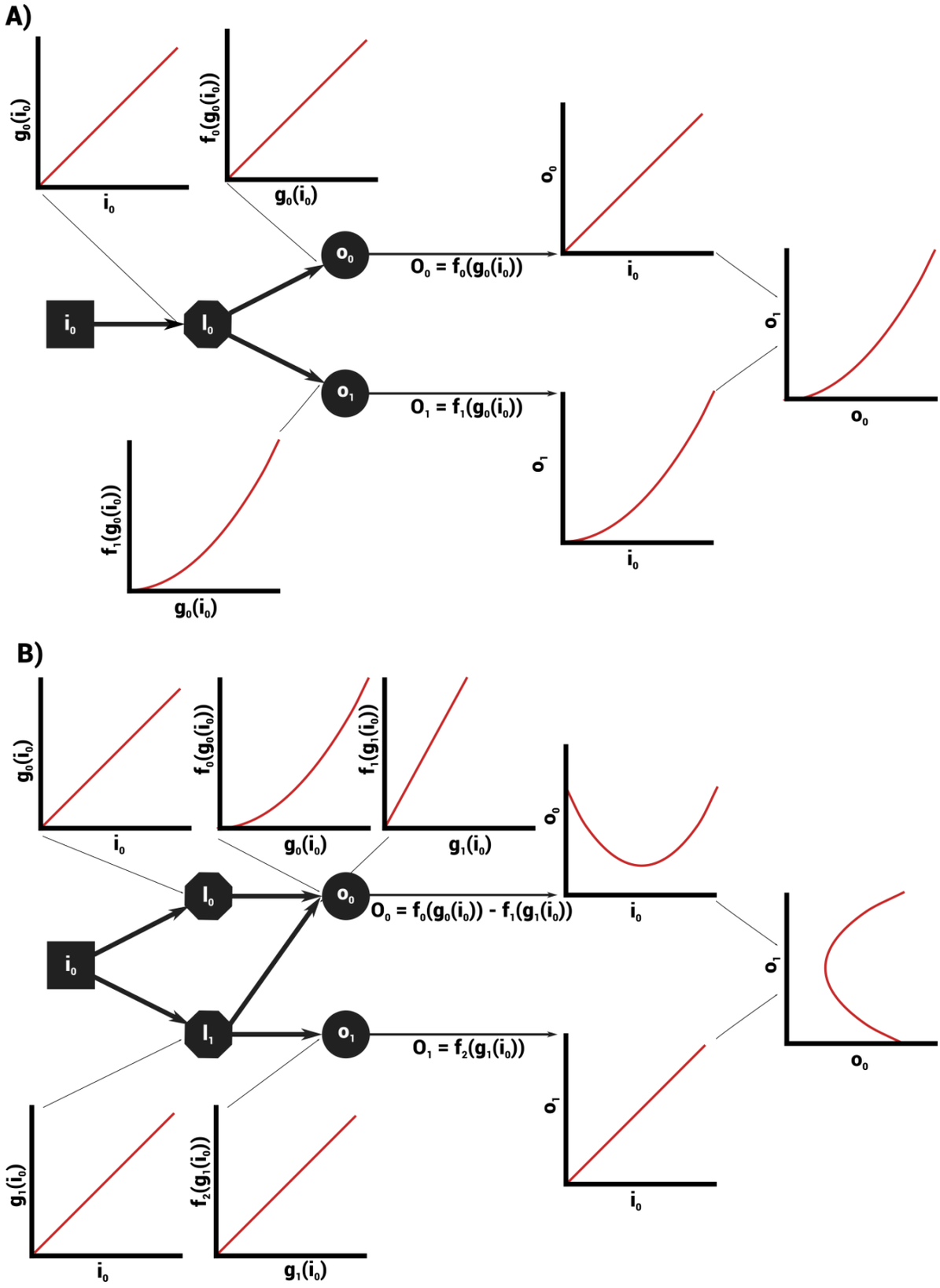


Figure 4) Example of how both monotonic and non-monotonic results can occur in state-trace analysis. A) An example of how a monotonic state-trace may arise with one single latent process. Notably in this instance, because all functions f_n are functions of the same latent variable and monotonic, the relationship between our O_0 and O_1 must also be monotonic. B) An example of how a non-monotonic state-trace might arise, even though all functions f_n are monotonic. In this instance we have also notably kept our functions g_m monotonic (and simple). Even despite, this we are able to demonstrate a non-monotonic relationship between O_0 and O_1.

therefore potentially consistent with either outcome. This criterion of monotonicity is not selected randomly. Monotonicity of this translation is at least an assumption that, implicitly or explicitly, is very widely relied upon (Krantz, Tversky 1971) and indeed, violations of this assumed monotonicity are likely to make constructive inference almost as hard as no assumption at all (Loftus 1978). Ultimately, without the assumption of monotonicity, we are left with two choices – translations of cognitive function to task performance f_n are either *non-unique* in the sense that multiple levels of cognitive function may equally translate to the same task performance, or *non-continuous* in the sense that translation of cognitive function to task performance occurs in sudden jumps instead of a smooth continuous fashion. Overall, this provides a material improvement over dissociation logic. In Figure 3, for example, dissociation logic would, potentially incorrectly, assert a dissociation exists over most of the range of inputs. State-trace analysis on the other hand, would (correctly) identify that there is not enough information to distinguish whether a dissociation exists over the same inputs. The more general case of uncrossed double dissociations incurs a similar fallacy to the one proposed in Figure 3, as it is simply the conjunction of two single dissociations (Dunn, Kirsner 1988). Crossed double dissociations are also similarly confounded, but would require any single process model to have at least one "negative" relationship in the sense of increasing cognitive function leading to decreased task performance (Dunn, Kirsner 1988). For an extensive discussion of this, see (Dunn, Kirsner 1988).

We describe state-trace analysis informally in terms of a state-trace plot, see Figure 5. We have a state factor consisting of our two tasks, with the performance on each task forming an axis on our graph. We then plot on this graph each level of our dimension factor, the variable that we are varying across our tasks. If we can draw a monotonically increasing (or decreasing) curve joining all the levels of our dimension factor, the relationship between our task performances across our variable is monotonic. In all other cases, it is non-monotonic. In the context of our attentional blink experiment, identity report and judging visibility are our two tasks so they give us our state factor, and the lags are the measure that we are varying across both tasks, so they give us our dimension factor. Plotting report accuracy on one axis and visibility on the other, we are trying to determine

whether it is possible to draw a monotonic curve joining the data across each of our lags.

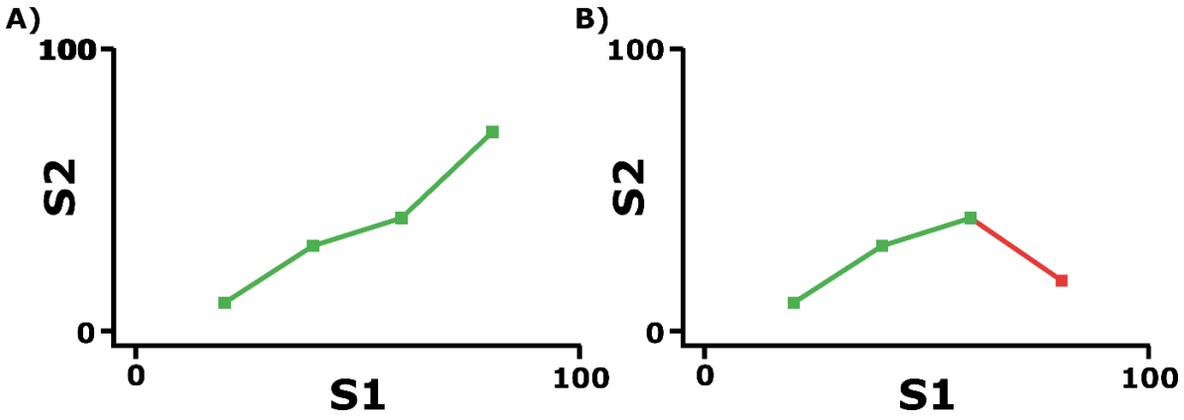


Figure 5) A) Example of a monotonic state-trace plot across 4 levels of a dimension factor D. It is possible to draw a monotonic (increasing) curve joining all points, therefore the relationship between the levels of our state factor is monotonic. B) Example of a non-monotonic state-trace plot across 4 levels of a dimension factor D. The point furthest to the right makes drawing either a monotonically increasing or monotonically decreasing curve impossible, therefore the relationship between the levels of our state factor is non-monotonic.

Often, we also include a "trace" factor in analysis of this type. While our current description of "traceless" state-trace analysis sets out a clear criterion for evaluation of a dissociation, actually eliciting such a dissociation is another matter. Often, it is desirable to introduce another manipulation into ones analysis to sweep out this behaviour, a *trace* factor (Prince, Brown et al. 2012). Notably, further than merely being desirable, a trace factor is required when one would otherwise only examine two points on a state-trace plot, since it is impossible to demonstrate a non-monotonic relationship from only two data points (Bamber 1979).

While adding a trace factor can be very helpful, it does come with complications. Since the trace factor is a convenience designed to sweep out the behaviour in the underlying system instead of the measure of interest itself, one must be careful that it does not compromise their analysis. It is therefore important to set the trace factor such that it is monotonic across the levels of the measure of interest, such that the non-monotonicity from the trace factor is not confused with non-monotonicity of the question of interest (Prince, Brown et al. 2012) (See (Davis-Stober, Morey et al. 2016) for an example of how this information can be taken advantage of). Similarly, potential interactions between the trace factor and

other factors must also be considered. One must also be careful that the trace factor we introduce is constructive. When the levels of the trace factor do not overlap on either of the dimensions, the state-trace plot is ineffective at diagnosing dimensionality (Prince, Brown et al. 2012). Instead, one effectively has two separate two dimensional state-trace plots, which could be consistent with either a multi-dimensional or unidimensional model.

Statistical methods

A long term challenge that state-trace analysis has faced is statistical quantification. Historically, there have been several attempts to make use of various methods of null hypothesis testing (Loftus, Oberg et al. 2004, Bamber 1979) in order to quantify the evidence for non-monotonicity. However, all of these methods suffer from the same problem – they can only quantify evidence against the null hypothesis (usually the monotonic ordering). Even if the underlying data is entirely consistent with the null hypothesis, noise in the data will *almost surely* (in the mathematical sense) lead to deviations from this in a data sample. Since evidence cannot be gathered for the null, only against it, this will (potentially) lead to bias. The *fallacy of classical inference* is an example of why this might be problematic. The fallacy of classical inference states in broad terms, that as the number of samples increases, so too will the effect size required to reject the null at a given alpha decrease, and that ultimately with enough samples we will reject the null even if the difference between conditions is trivially small (Friston 2012). Furthermore, one of the strong points of state-trace analysis on its own is that it can help assess when complex patterns of behaviour can be modulated by a simple underlying structure. When one makes use of null hypothesis testing, there is no benefit to choosing simple models of the data, and indeed, a more complex model will fit often the data better; as such, null hypothesis methods are effectively discarding this advantage. Additionally, these methods all rely to a greater or lesser degree on parametric assumptions of the data. One of the beauties of the original state-trace analysis was that it was essentially assumption free, and an ideal statistical test will allow this to continue.

As a solution to these problems, several more recent pieces of work advocate the usage of a Bayesian approach (Sense, Morey et al. 2016, Davis-Stober, Morey et

al. 2016, Prince, Brown et al. 2012). The Bayesian approach allows evidence to be compared for both possible hypotheses, monotonic or non-monotonic. Furthermore, Bayesian statistics does not necessarily rely on any parametric assumptions, allowing the non-parametric nature of the analysis to be preserved. That is not to say that Bayesian analysis is perfect. Firstly, it requires the specification of a prior, and it is not always clear what a suitable prior may be. That said, the prior can also often be a chance to take advantage of prior knowledge (Davis-Stober, Morey et al. 2016), so if this uncertainty is overcome this can become an advantage rather than a disadvantage. Secondly, computation of the posterior often either requires a method such as Gibbs sampling that is highly computationally expensive (Prince, Brown et al. 2012), or special parametric assumptions that allow it to be computed analytically (Davis-Stober, Morey et al. 2016).

All of the statistical methods we have discussed so far are about analysis at the single subject level. However, we will almost always have more than one subject in a study. Unfortunately, state-trace analysis at the group level presents some difficulty. Primarily, it is not possible to apply summary statistical methods like averaging to state-trace data (Newell, Dunn 2008, Prince, Brown et al. 2012). It is possible both to average multiple non-monotonic datasets into a monotonic dataset, and vice versa. On the basis of the Bayesian method, there are several approaches that allow evidence to be assessed across a group of participants. The simplest method is the Grouped Bayes factor (Prince, Brown et al. 2012). This method assumes that the Bayes factors from each participant are independent from one another. The Grouped Bayes factor is then the product of all individual Bayes factors, effectively quantifying the evidence that the whole group is monotonic versus the whole group being non-monotonic. This method has the advantage of being very simple but aside from the (perhaps strong) assumption of independence, also relies on the implicit assumption that participants' results are approximately homogenous.

If all participants are either monotonic or non-monotonic, it is a very sensible measure of the group level effect. However, if there are participants in each direction, particularly if the results of some participants are very strong compared to others, it rapidly becomes less useful. Imagine a case in which we have nine

participants, 8 of which evaluate a Bayes factor of 10, and 2 of which evaluate a Bayes factor of 0.002. The grouped Bayes factor would be 0.32, indicating marginal evidence in favour of the effect that the two are showing at the group level, but this would not be a sufficient summary of the behaviour. A recent proposal by (Davis-Stober, Morey et al. 2016) attempts to solve this by supplementing the grouped Bayes factor with another Bayesian analysis that assesses the homogeneity of the Bayes factors which the authors label the aggregated Bayes factor (ABF). While this method provides a very valuable tool, it suffers from several limitations. It can only unambiguously demonstrate the dangerous case of heterogeneity, but cannot entirely support homogeneity. When homogeneity is supported, individual level tests need to be examined. For this reason, the authors advocate its usage alongside the grouped Bayes factor already introduced.

RSVP and the Attentional Blink

We have so far discussed some of the literature background to working memory encoding and subjective experience and given some techniques by which we might separate the two. However, in order to provide any such separation with the best possible opportunity to manifest we must choose an appropriate experimental paradigm. A paradigm that is well placed to shed light on this topic, and has been used previously (Sergent, Dehaene 2004) to explore the all-or-none nature of subjective experience, is the attentional blink (AB). The attentional blink is a phenomenon seen during RSVP (Rapid Serial Visual Presentation) in which participants frequently fail to detect a second target for a short time after the presentation of an encoded first target; see T2|T1 accuracy in Figure 6 (Raymond, Shapiro et al. 1992, Bowman, H., Wyble 2007). We propose that the attentional limitations causing this impairment are potentially informative: such impairment may help distinguish the results of failed accessibility from other characteristics of perception such as subjective experience. Others have argued similarly (Cohen, Cavanagh et al. 2012), but few attempts have been made to empirically explore this question, with (Pincham, Bowman et al. 2016) being an exception. Another advantage of the attentional blink is that the paradigm has been so extensively modelled. While experiments and dissociations can tell us about specific effects, placing findings in larger theoretical context is pivotal to the forward progress of

science, and one of the most powerful tools for achieving this is modelling. The large number of models, including many with computational implementations, makes this a potentially fruitful direction to take our research question.

In this section, we introduce the RSVP and attentional blink paradigms as well as looking at how subjective experience has been explored in the context of the attentional blink up until now. We also take the opportunity to introduce several models of the attentional blink.

The Attentional Blink

Rapid serial visual presentation (RSVP) is a presentation technique in which multiple stimuli are presented rapidly (usually 6-20 items per second (Raymond, Shapiro et al. 1992)) one after the other in a fixed location. The stream of stimuli is typically composed of targets to be identified and distractors to be ignored that are distinguished by some feature - for example the colour or type. As items are presented in the same spatial location, each item acts as a mask for the previous item. Between this masking and the rapid presentation rate, it becomes difficult to identify any single item in the stream, and this makes it ideal for testing the limits of the attentional and perceptual systems.

When more than one target is presented in an RSVP stream, both targets cannot always be processed to the same level. The attentional blink (AB) is a “blink of the mind's eye" that presents as a deficit in performance on a second target when more than one target is to be identified in an RSVP stream. It arises approximately 100-500ms after the presentation of the first target. It is a particularly robust finding that has been demonstrated over a large number of studies over a wide range of task conditions (Martens, Wyble 2010). Typically, the AB is elicited using alphanumeric stimuli, but images, letters, digits or words will all elicit the blink. For an example of a typical attentional blink RSVP stream and the associated results, see Figure 6.

In the context of the attentional blink, there are several terms of particular importance: targets, lag, and Stimulus Onset Asynchrony, as well as several findings that bear discussion – particularly lag 1 sparing. The attentional blink paradigm is typically performed over two *targets* which are labelled T1 and T2 for the first and second target respectively, though sometimes more targets are

used. These targets are labelled as T3, T4, etc. for the third and fourth and so on. The main parameter of the attentional blink is the relative serial positions at which the two targets are presented, known as *Lag*. For example, at Lag 1 there are no intervening distractors between the targets, while at Lag 2, the two targets are separated by one intervening stimulus. Also relevant is the presentation rate of stimuli is known as *Stimulus Onset Asynchrony* (SOA). SOA specifically refers to the number of milliseconds between the onset of one stimulus and the onset of the next. Finally, the classic attentional blink finding (Figure 6), arises when lag is plotted against T2|T1 accuracy (second target accuracy, given the first target was correct). Excluding Lag 1, typically, as the two targets approach one another, accuracy is significantly reduced compared to recovery baseline (lags 7 and 8). A typical blink is shown in Figure 6. Performance at Lag 1 however is often at and sometimes even above recovery level. This is known as *Lag-1 sparing*, and is itself a common finding of the attentional blink (Wyble, Brad, Bowman et al. 2009). Interestingly, when multiple targets are presented with no intervening distractors, this sparing can be "spread" to up to 4 further targets (Olivers, Van Der Stigchel et al. 2007) in a finding that is often called "spreading the sparing".

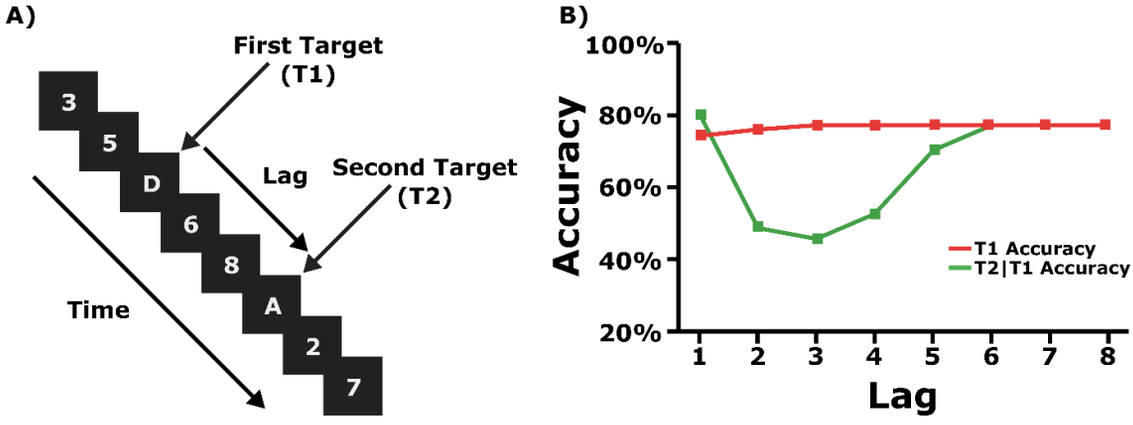


Figure 6) A) A typical attentional blink RSVP stream. Participants are instructed to report the two letters at the end of the stream. B) Illustration of expected accuracy for T1 and T2|T1 at each lag during a typical attentional blink study with an SOA of approximately 80-120ms.

Incorporating Subjective Experience - The Experiential Blink

The attentional blink has been extensively studied with respect to report accuracy and working memory encoding; however, for the purposes of our research question we are interested in it as a tool for distinguishing these measures from subjective experience. Up until recently however, subjective experience during the attentional blink has largely been explored obliquely through studies attempting

to determine the all-or-none nature of subjective experience. The question of the all-or-none nature of subjective experience considers whether a subjective experience of a stimulus can be graded, or whether subjective experience is a binary phenomenon – either on or off.

One of the most widely cited papers providing evidence for bifurcation of subjective experience is presented by (Sergent, Dehaene 2004). Importantly for the research question of this thesis, it achieves this by analysing subjective report during the attentional blink. In this instance, subjective report is seen to behave very similarly to the classical pattern of behaviour report accuracy shows during the attentional blink (See Figure 7). There has though, been debate about the interpretation of these results. Notably, the 21 point response scale used for subjective report has been discussed (Overgaard, Rote et al. 2006, Nieuwenhuis, de Kleijn 2011), and it has been argued that participants have difficulty using scales with a large number of response points (Sandberg, Timmermans et al. 2010).

A more recent study (Nieuwenhuis, de Kleijn 2011) changed to a 7 point scale, and made use of a more traditional character identification task. The authors also perform experiments using both direct subjective report, and a post decision wagering method that requires participants to bet on the certainty of their outcome. With these changes, the authors find graded responses for both direct subjective report and post decision wagering. In terms of how subjective experience behaves during the attentional blink, the authors find a remarkable contrast to report accuracy. While report accuracy demonstrates an attentional blink with lag 1 sparing, the subjective report in both instances had considerable less lag 1 sparing. This can be seen in experiments 3 and 4 in Figure 7(B), which uses both the revised experimental paradigm and improved scale.

Both of these articles have only studied subjective experience during the attentional blink as a method of assessing the all-or-none nature of subjective report. One article that has directly studied how subjective experience and working memory encoding differ during the attentional blink is a study by (Pincham, Bowman et al. 2016). In this article, the authors perform two experiments, one assessing the behaviour of subjective report during the attentional blink by sampling a large number of lags, and another additionally

assessing electrophysiological behaviour with the same experimental paradigm, but over a smaller number of lags in order to concentrate data in a few conditions in order to obtain robust ERPs.

The authors of this work find the behaviour of subjective report during the attentional blink to be very similar to the results of (Nieuwenhuis, de Kleijn 2011) (See Figure 7(C)). One notable difference is that where the subjective visibility results from (Nieuwenhuis, de Kleijn 2011) did show a minimal level of lag 1 sparing, the data from (Pincham, Bowman et al. 2016) shows none. Nonetheless, the authors did observe a second blink of the mind's eye for subjective report. They call this the *experiential blink*, a term that we will use. This experiential blink is distinct from the attentional blink, as it occurs over a different measure (subjective visibility versus report accuracy), and significant because it exhibits different behaviour (absence or reduction of sparing at earlier lags). Interestingly, exploring this finding in more detail, the authors also raise the possibility that they have found evidence for a case in which participants are encoding targets into working memory without experiencing them (Pincham, Bowman et al. 2016) in a phenomenon they call *sight-blind recall*, a term we will also adopt. In terms of electrophysiological results, the authors find the P3 component differentially affecting report accuracy and subjective report, specifically that the P3 indexes subjective report more closely than report accuracy. Specifically, the authors suggest that that when reporting poor subjective visibility on the second target, participants show a weakened and shortened P3 compared to when reporting high subjective report.

Taking these three studies together, there is significant evidence that the attentional blink is indeed a good paradigm with which to distinguish working memory encoding and subjective experience. In particular, there is a significant difference in the pattern at Lag 1. Additionally, a goal we discussed in the introduction was not just to empirically explore the relationship between working memory encoding and subjective experience, but additionally to provide a simulation model of any results. In light of this, we now turn to providing a brief overview of some contemporary models of the attentional blink.

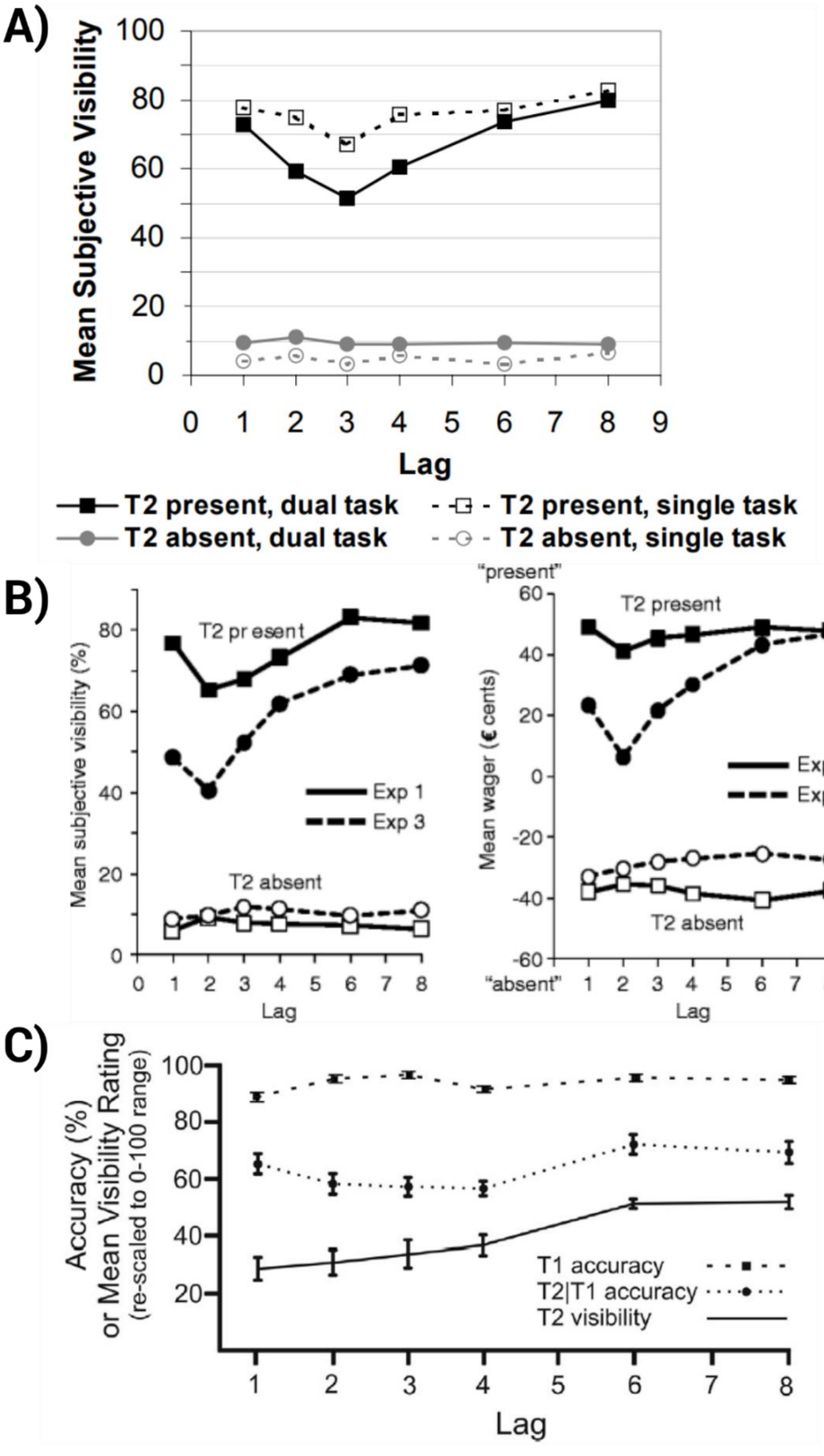


Figure 7) Subjective Report during the Attentional Blink across 3 experiments. A) (Sergent, Dehaene 2004) In the dual task participants were required to identify the first target, while in the single task they were not. In all conditions participants were required to make a subjective visibility judgement on the T2. B) (Nieuwenhuis, de Kleijn 2011). Experiment 1 replicates (Sergent, Dehaene 2004) without a single task condition and with a 7 point response scale. Experiment 2 is

the same as Experiment 1, but substitutes subjective report with post decision wagering. Experiment 3 changes the word presentation task used in (Sergent, Dehaene 2004) to a character identification task that is more representative of typical attentional blink research. Experiment 4 is as experiment 3, but substitutes subjective report with post decision wagering. C) (Pincham, Bowman et al. 2016) behavioural results from Experiment 1. In this experiment, participants were required to report the identity of both targets, but only the subjective experience of the second. Subjective report was on a 6 point scale.

Theories and Models of the Attentional Blink

When the attentional blink was first discovered (Raymond, Shapiro et al. 1992), there was much focus on explaining why it was occurring. In particular, there is evidence that missed targets during the attentional blink are processed quite extensively, both from behavioural and neurophysiological studies (Martens, Wyble 2010). On this basis, many of the dominant theories of the attentional blink were initially based on *central capacity limitations*. That is, that the attentional blink was arising as a limitation in capacity of some central attentional mechanism. The specifics of these models varied, but were generally united in assuming that all stimuli in a stream were fully processed up to the point of conceptual representation (Martens, Wyble 2010). Given that the attentional blink cannot be "trained out" (Braun 1998, Taatgen, Juvina et al. 2009), these *central processing* accounts that assume that the attentional blink is the result of a fundamental cognitive limitation seem reasonable. However, some more recent findings called these assumptions into question.

Firstly, it is possible to reduce the attentional blink by redirecting the participant's focus away from target identification. This has been done by adding task irrelevant visual motion or flicker (Arend, Johnston et al. 2006), changes in task instruction (Ferlazzo, Lucido et al. 2007), and most compellingly, the introduction of a second irrelevant task (Taatgen, Juvina et al. 2009). It is also possible to identify multiple targets in a row, as long as there are no intervening distractors, an effect called "spreading the sparing" (Olivers, Van Der Stigchel et al. 2007). All of these results present some difficulty for accounts that propose that the attentional blink emerges due to a fundamental information processing limitation of the brain. It is hard for any model that relies on an information processing limitation to explain why performance *increases* with the addition of a second

task. As a response to these findings, further accounts have been proposed, of which we review a selection here.

Locus Coeruleus Model

One model is the Locus Coeruleus Model (Nieuwenhuis, Gilzenrat et al. 2005). The locus coeruleus is a small nucleus in the brainstem that is widely connected (Figure 8), known to be critical for the regulation of cognitive performance through norandrenergic projection to the cortical mantle , and that has been argued to be central to the deployment of attention. In monkeys, the neurons of the LC fire a powerful burst in response to visually presented targets (Nieuwenhuis, Gilzenrat et al. 2005). Following this burst of activity is the widespread release of norepinephrine (NE) to cortical areas, providing an attentional enhancement that allows the target to be encoded into working memory (Nieuwenhuis, Gilzenrat et al. 2005). Although NE release potentiates processing in cortical areas, it is also thought to be locally autoinhibitive. For this reason, following the burst of activity in the LC is a refractory period in which further LC discharge is rarely observed (Martens, Wyble 2010, Bowman, Howard, Wyble et al. 2008).

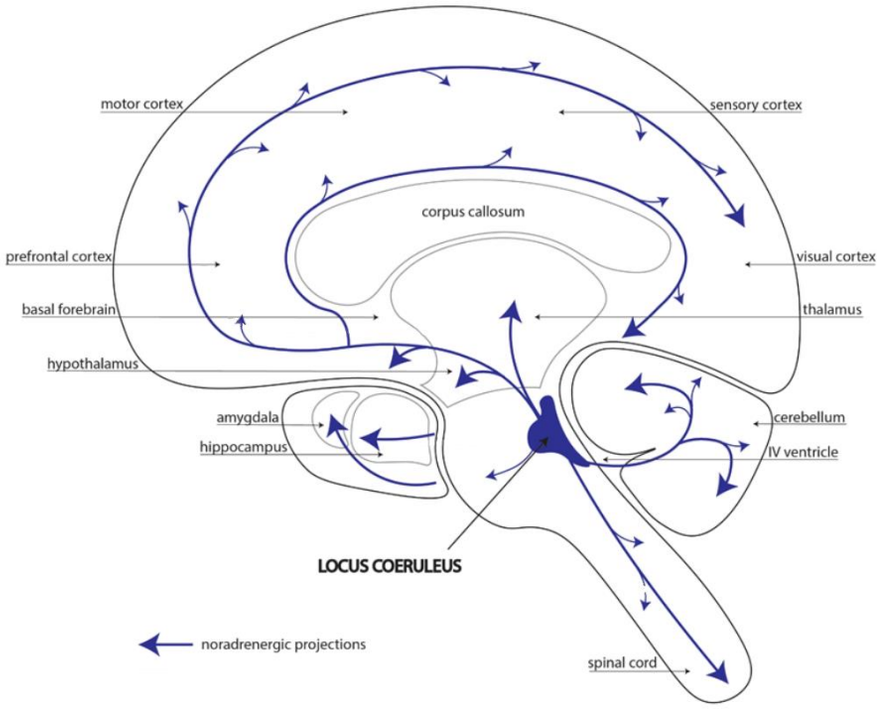


Figure 8) Noradrenergic projections from the locus coeruleus. Adapted from (Feinstein, Kalinin et al. 2016).

The LC-NE model arises from noting that this behaviour is remarkably similar to the timescale of the attentional blink. Since the locus coeruleus is so involved in

attention and cognition, it is no stretch to consider that the refractory period of the LC system is what mediates the attentional blink. The explanation for lag 1 sparing in this instance would be that the presence of the second target is close enough to the first target to benefit from the initial release, but targets at later lags are not. Though it does provide a potential explanation of the attentional blink (though see (Bowman, Howard, Wyble et al. 2008)), it is noted that if the AB is mediated by noradrenergic mechanisms, the application of an adrenergic agent should directly affect the refractory period of the blink. (Nieuwenhuis, Van Nieuwpoort et al. 2007) failed to find this effect, suggesting that if the adrenergic system does modulate the attentional blink, it does so indirectly.

Boost and bounce model

Another model is the boost and bounce model (Olivers, Meeter 2008), see Figure 9. It proposes that the inhibition seen during the attentional blink has a functional role in working memory encoding, as a way to keep non-targets from intruding upon working memory encoding. Broadly, a seen target triggers an attentional "boost" that helps the target reach working memory. In order to prevent items other than the intended target from being encoded into working memory, this boost is followed by a "bounce". This bounce inhibits attention and blocks further processing. It is proposed that this dynamic is what causes the attentional blink. Second targets that fall into this bounce window are inhibited and are less likely to be encoded. Lag 1 sparing arises because two targets next to one another end up both benefiting from the boost – the inhibitory process is not triggered by the presence of a target. While the simplicity of this model is a major virtue, it is also something of a challenge; it has been speculated whether it is capable of fully accounting for the attentional blink (Martens, Wyble 2010). Furthermore, it is noted that the attentional blink can be found in the absence of a distractor following the second target (Nieuwenstein, Potter et al. 2009). This is a problem for the boost and bounce model, as a distractor would be necessary to generate the "bounce" required for the blink. This said, it does explain many key attentional blink findings. Notably, it is consistent both with the presence of a secondary task improving attentional blink performance (Olivers, Meeter 2008, Taatgen, Juvina et al. 2009), as well as the spreading the sparing findings in which multiple targets

are presented (Olivers, Meeter 2008).

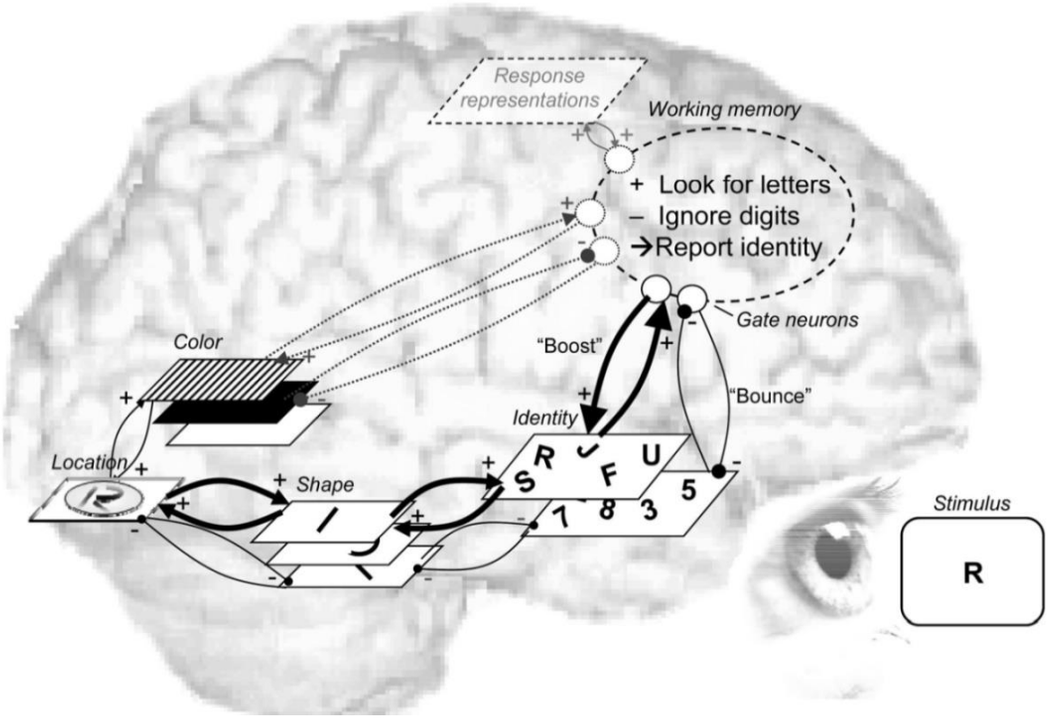


Figure 9) The boost and bounce model of attention, adapted from (Olivers, Meeter 2008). A demonstration of how the boost and bounce model would work in an RSVP paradigm. The boost and bounce are delivered by excitatory and inhibitory gate neurons. When a target arrives, a strong attention enhancement is triggered by an excitatory gate that remains open until a distractor hits. This distractor triggers a strong response from an inhibitory gate, which causes the blink.

Threaded cognition model

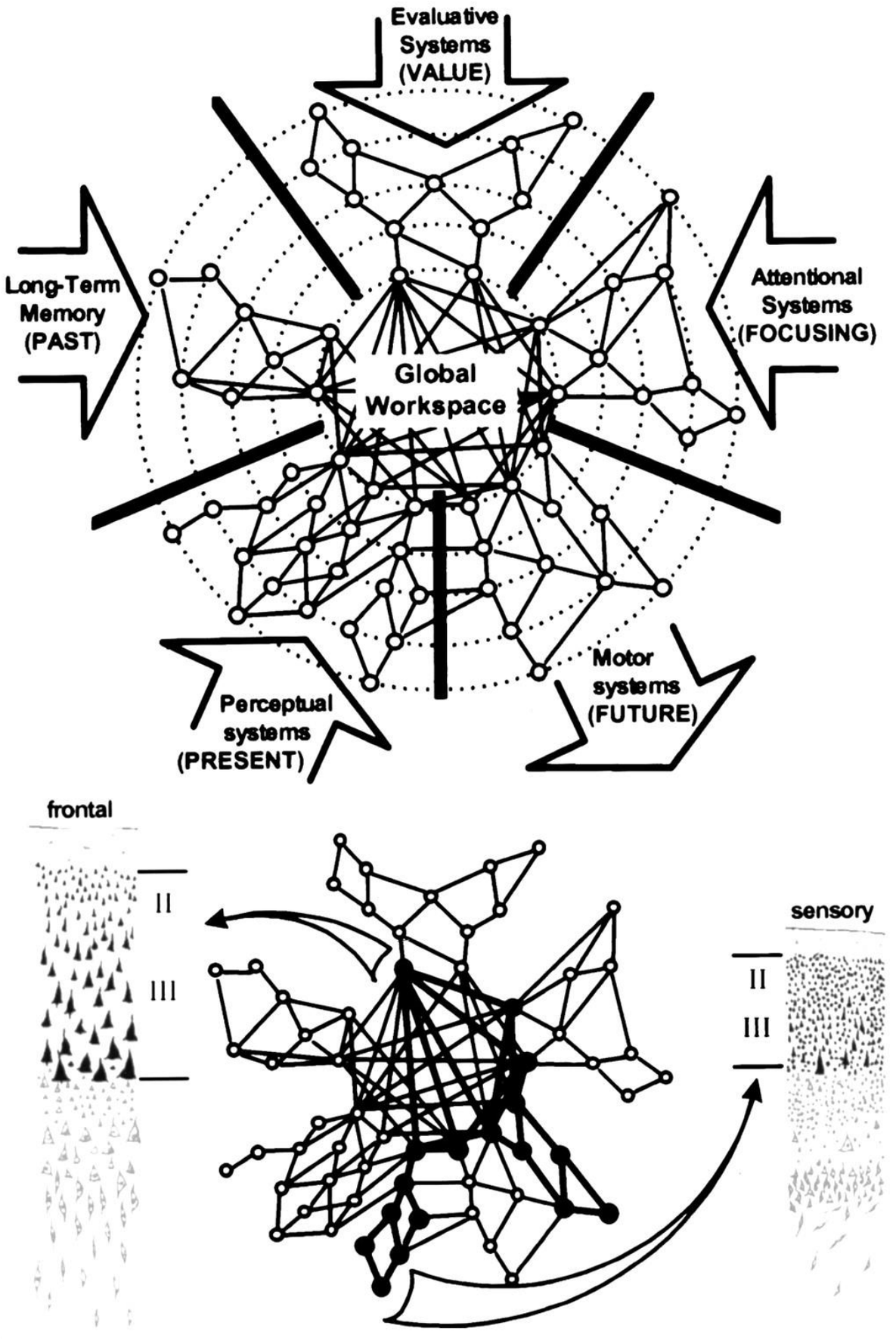
The threaded cognition account (Taatgen, Juvina et al. 2009) is similar to the boost and bounce model discussed above in that it predicts the attentional blink to be the result of an overzealous attentional control mechanism, but proposes a significantly different mechanism. Broadly, it is proposed that many individual cognitive resources can be used in parallel but any one resource can only be used for one given single task (thread). Different threads compete for resources, and in the absence of cognitive control, once a resource is freed up, the first thread that requires it will take it up. The attentional blink arises because inside this system, a rule has been set up to protect target consolidation. This rule suspends further target detection in order to prevent targets and distractors from being consolidated into one percept. Importantly, such a rule is clearly overzealous – it is an overexertion of cognitive control because T1 consolidation occurs even without such a mechanism. The model is able to replicate distraction reducing the AB, and findings that some people do not show the blink (Martens, Munneke et al. 2006, Taatgen, Juvina et al. 2009). One difficulty of the model from the point of view of the research question of this thesis is that it is not a neural network.

This makes it unclear how to relate it to electrophysiological findings, something we will do extensively later in the thesis.

Global Workspace Model

The Global workspace model was first suggested by Baars (Baars 1997). The core idea of global workspace theory is that conscious content is globally available to diverse conscious mechanisms without any need for further processing. This has been expanded on by several sets of researchers (Franklin, Graesser 1999, Shanahan 2006), but by far the most dominant model is by Dehane et al. (Dehaene, Kerszberg et al. 1998), who provide a neural implementation – the "Neuronal global workspace" (Dehaene, Changeux 2011). This model assumes that many different, highly specialised information processors are joined via long range connections. This network of connections forms a higher level unified space in which information is broadly shared and broadcasted back to lower level processes. In this model, a piece of information becomes conscious by being shared across this global network. At any given moment, many pieces of information are being processed unconsciously in parallel by these modular processors. The piece of information that is conscious is decided by a winner-takes-all competition (aided by top down attention) between the neural populations representing each piece of information.

The Neuronal Global Workspace model has been developed in great detail. Much of the model has been shown to have strong physiological plausibility (Dehaene 2014, Dehaene, Changeux 2011, Dehaene, Kerszberg et al. 1998), and it correctly predicts a range of behaviours in, for example, Conscious perception (Dehaene, Changeux 2011) Inattentional Blindness (Dehaene, Changeux 2005) and (most importantly for us) the Attentional Blink (Dehaene, Sergent et al. 2003). In this model, the attentional blink arises as the result of this winner-takes-all competition for global workspace activation. The act of a piece of information becoming conscious is what makes it available for report. The widespread pattern of activation that arises as the result of the T1 blocks the entry of the T2 into the global workspace for a window of about 200ms. At early lags, this inhibits T2 performance, but at later lags, this becomes less relevant. While a computational version of the model exists and the model does explain why stimuli presented closely together are both processed to a high level, it struggles



.

Figure 10) The Global Neuronal Workspace model (adapted from (Dehaene, Kerszberg et al. 1998). (Upper), A schematic of the 5 main types of processor proposed to be linked by the global workspace. (Lower) A link between two processors is established through the activation of distributed workspace neurons.

to explain the lag 1 sparing effect, and the “spreading the sparing" findings (Martens, Wyble 2010). Currently, there is no immediate explanation for the behaviour at the lag 1 data point, though the presentation of two concurrent targets shows similar results (Taatgen, Juvina et al. 2009).

The Simultaneous Type/Serial Token model

The Simultaneous Type/Serial Token (STST) (Bowman, H., Wyble 2007) and its extension, the episodic Simultaneous Type/Serial Token models (eSTST) (Wyble, Brad, Bowman et al. 2009) are a pair of neural network models of the attentional blink. Both models propose a two stage account of the blink that bears some resemblance to the two stage model by (Chun, Potter 1995). The (e)STST models build on a *type*/*token* distinction to simulate how items are bound into temporal contexts. In this definition, the *type* of a stimulus encompasses all of its instance invariant properties, the features that do not change between occurrences. Take the letter K for example; parts of its type are its semantic features (e.g. it is a letter, it is after J in the alphabet) and its visual features (e.g. its shape and colour). Conversely, a token represents a specific episodic occurrence of a type and particularly, where it occurred in time relative to other items. In the STST model, in the first stage, types are processed in parallel with many types simultaneously but fleetingly represented, and it is the act of sequentially binding a type to a token (tokenisation) in the second stage that creates a solidified representation in working memory. This binding is achieved using a resource called the binding pool, which provides a neutrally plausible mechanism by which types can be bound to tokens.

In the (e)STST models, a component of the model called the *blaster* provides a boost to highly salient items from the first stage that allows them to reach the threshold for binding in the second stage. Similarly, in order to provide protection to the tokenisation process and to prevent it from being corrupted, after having fired, the blaster receives a powerful inhibitory signal which prevents it from firing again for a short period. In this context, lag 1 sparing arises when the two targets end up both benefitting from the same blaster firing and potentially end up being bound into overlapping temporal contexts (i.e. to the same token). Since the two targets are bound to these overlapping temporal contexts, this account predicts an increase in order errors at early lags, something which is seen in human data

(Chun, Potter 1995). The blaster is also where the main difference between the STST model and the eSTST model arises. Since the firing of the blaster is time limited, the original STST model struggles to provide an account for the "spreading the sparing" findings that occur over a comparatively long period. In the eSTST, model the model dynamics have been changed such that multiple target items in sequence allow the blaster to remain active. This allows an account of the blink that is not strictly time limited, and therefore accounts for the spreading the sparing finding.

Through these mechanisms, the Simultaneous Type/Serial Token model creates an account of working memory encoding that is consistent with many attentional blink findings. The original model struggled to account for the spreading the sparing effect in which multiple targets not separated by distractors would be encoded accurately. With the additional enhancements provided by the eSTST model, though, this is no longer a problem. Of note for our research question, there also exists a computational model of STST from which it is possible to generate both behavioural data, and also "virtual" ERP's (Wyble, Bradley, Bowman 2005, Craston, Wyble et al. 2009) that closely mimic the results from human participants.

Metacognition

The core research question of this thesis concerned with the relationship of working memory encoding and subjective experience. So far, in this literature review, we have discussed these processes in the context of the literature on consciousness, reviewed techniques by which they (working memory encoding and subjective experience) might be separated and discussed an appropriate experimental paradigm over which to do so. One interesting facet of the analysis we have discussed so far, however, is that they are based on averaged behaviour. For example, in the last chapter, we discussed the experiential blink paradigm, and in particular the lag 1 data point at which report accuracy is high, and subjective experience is low. While this is an interesting finding in its own right, based on just this averaged behaviour, it is difficult to interpret the results. We are unable to distinguish whether this result is because participants are losing the ability to accurately reflect on their experience, or whether their criteria for reporting subjective experience simply is changing at lag 1.

If participants are losing their ability to accurately reflect on their experience, we may see more low subjective reports attached to correct trials (and vice versa for high subjective report and incorrect trials), and given the fairly high accuracy rate of correct report, it would not be surprising if this resulted in worse subjective report despite high report accuracy. This would indicate that participants are suffering from a failure of self-reflection. Conversely, though, it could simply be that at Lag 1, the view of two targets with no distractor is an *odd* percept instead of a *poor* percept, and this uncertainty causes participants to shift their criteria for reporting subjective experience of targets to be more conservative. This would also result in a worsening of subjective report despite high report accuracy, but lends itself to a completely different theoretical interpretation than the first example.

What we are searching for to distinguish these behaviours is a measure of *discriminability*. In this case, how well our high and low subjective reports distinguish between correct and incorrect trials. Such a measure has been quantified in the field of signal detection theory (SDT), and has been widely applied in examining how objective performance matches up to confidence in the study of *metacognition*. In this section, we provide a (very) brief overview of Type 1 and Type 2 SDT for unfamiliar readers, and discuss how these techniques have been applied in the field of metacognition. Note that, as we discussed in the introduction, when applying such methods to our own data in which we collected *subjective report* instead of *confidence*, the same theory is applicable, but we will refer to such measures as meta-experience.

Signal Detection Theory

To begin our brief introduction to SDT in the context of metacognition, we first set out some nomenclature. Metacognition is about thoughts that are about other thoughts, a multileveled concept. We therefore distinguish the “Type 1" measures as the first level measures, those that are attempting to objectively evaluate the world. In contrast to this, we have our “Type 2" measures that are those about these first level thoughts. To tie this back to a concrete example, in our experimental data, our Type 1 measure is objective task performance, while our Type 2 measure is the confidence in the results of the Type 1 Task. To evaluate

metacognition from this, we are asking the question “How well do our Type 2 results distinguish the levels of our Type 1 performance?"

There are several approaches one might take to solve this problem. The simplest approach would be to somehow evaluate the accuracy of our Type 2 reports, either directly (Persaud, McLeod et al. 2007) or through a correlation (Kornell, Son et al. 2007, Nelson 1984): how often does a correct response correspond to high confidence, and an incorrect response correspond to low confidence? While this does give a simple and intuitive measure of parity between Type 1 responses and confidence, it also demonstrates the problem that many potential measures suffer from – that these measures are contaminated by *bias*. In terms of our later SDT framework (e.g. Figure 12), changing c may lead to a change in correlation, even for a constant d' (Nelson 1984). Say, for example, a change in our reporting criterion shifts out true positive and false positives rates from 70% and 30% to 60% and 40%. This is very likely to lead to a change in correlation, but does not necessitate a change in discriminability. Our problem specification is to determine how well the levels of our Type 2 measure *discriminate* between the levels of our Type 1 measure, a measure we will from now on refer to as *metacognitive sensitivity*. While these proposed measures do evaluate this, they are also affected by the overall propensity for correctness or incorrectness, what we will refer to from now on as *metacognitive bias*. Take for example, two participants who are asked to perform a detection task, and rate their confidence in their response. What we are interested in is how well their confidence ratings *discriminate* correctness or incorrectness – metacognitive sensitivity. If one participant has a propensity for overconfidence that is reflected in consistently high confidence ratings, while another has a similar propensity for under confidence, this (metacognitive) bias should not affect our measure of discriminability as long as their confidence ratings equally distinguish type 1 correctness. An alternative class of approaches that does distinguish between the metacognitive bias and metacognitive sensitivity, and can provide a measure that is metacognitive bias free, are those based on Signal Detection Theory.

SDT attempts to evaluate the discriminability of our measure independent of any individual level of bias. The basic SDT framework is as follows. Take a task for which there are two different possible trial outcomes Signal (S) in which a

stimulus is present and Noise (N) in which it is not. Participants are attempting to evaluate these outcomes, and produce one of two different responses, Respond Signal (RS) if they believe the signal to be present, and Respond Noise (RN) if they believe the signal was not present. When either a signal or a noise trial is presented, it produces evidence according to normal distributions with equal variance on a one dimensional internal decision axis. A criterion c is set on this axis and when evidence is above this criterion, the participant reports Signal, and when it is below they report Noise. See Figure 12 for an illustration.

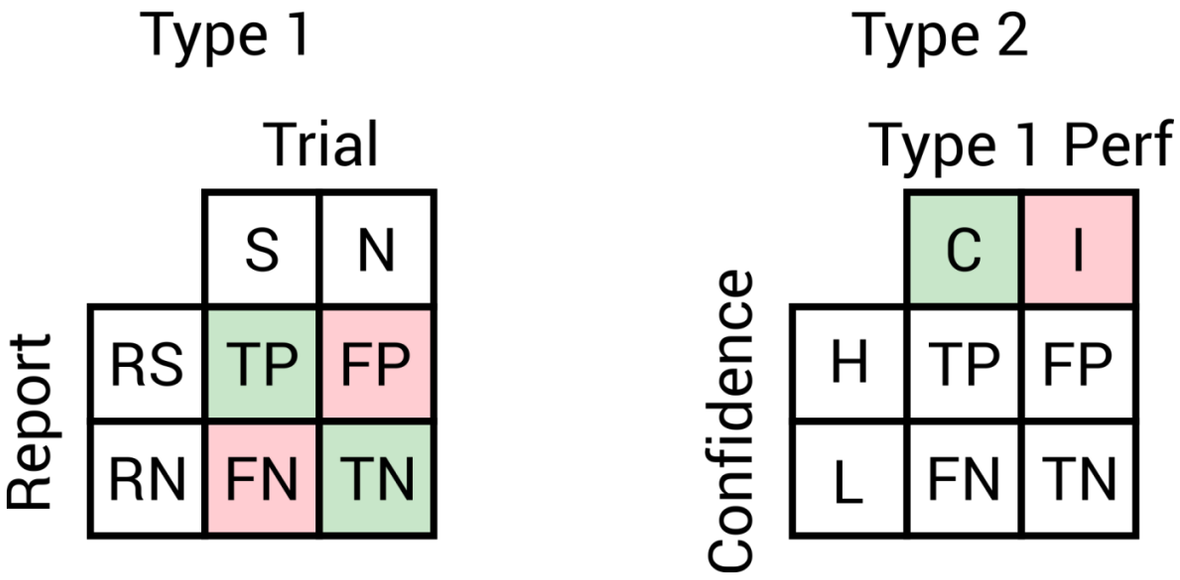


Figure 11) Different outcomes from the Type 1 signal/noise identification task, and the Type 2 confidence task.TP = True positive, FP = False Positive, TN = True Negative, FN = False Negative. In the Type 1 task, trials either have signal (S) or noise (N) present, and participants respond with either signal present (RS) or noise present (RN). In the Type 2 task, participants are judging the correctness of their type one response with either High Confidence (H) or Low Confidence (L). Correct trials (C) are those for which the presence or absence of the target was correctly judged (TP and TN) and incorrect trials (I) are those where the presence or absence of the target was judged incorrectly (FP or FN).

Given this setup, there are four different possible outcomes from any trial. A signal is present (S) and the participant reports a signal (RS), a True Positive (TP); a signal is absent (N) and the participant reports a signal (RS); a False Positive (FP); a signal is present (S) and the participant reports noise (RN), a False Negative (TN); and a signal is absent (N) and the participant reports noise (RN), a True Negative (TP). This is illustrated in Figure 12A). From this model we can observe two things. Firstly, that since $TP = 1 - FN$ and $FP = 1 - TN$, we need only consider two of these measures at any time. We will stick with the TP and FP

rate for consistency. Next, while our ideal case is one in which our TP rate is maximal, and our FP rate is minimal, we can trade these measures off against one another by shifting our criterion c. In this model, bias is simply the overall propensity to give one type of response or another – the positioning of the c criterion. How well we can discriminate between S and N depends not on the choice of c at all, but how close together our S and N distributions are, which can be quantified by the measure of *discriminability* d':

$$d' = \frac{\mu_S - \mu_N}{\sigma_{SN}}$$

That is, the difference in the means of the two normal distributions in units of their mutual standard deviation. This d' can be calculated empirically from TP and FP rates gathered from human data. This general framework can be applied both to Type 1 data and Type 2 data, and a Type 2 d' calculated empirically in the same way. In Type 2 data, Type 2 TP and Type 2 FP rates are calculated analogously, see Figure 11.

One problem with this framework is the assumption that the data is normally distributed across the decision axis described. When the data is not normally distributed, the d' measure described can be contaminated by bias. One resolution is the use of Receiver Operating Characteristic (ROC) curves. ROC curves plot the True Positive/False Positive trade-off across a range of conditions. These conditions should be set up such that they reflect a range of different metacognitive *biases,* while maintaining the same metacognitive *sensitivity*. The ROC curve as a whole then embodies, in a sense, metacognitive sensitivity independent of any individual bias. In fact, when the assumption of normality holds, the SDT framework and ROC curves are equivalent. In order to extract a measure of discriminability of these ROC curves when these assumptions do not hold, it is common to use the area under the ROC curve. This effectively captures the sensitivity given by the curves, and provides a good statistic of the overall TP/FP trade-off.

Metacognition and SDT

With this introduction to SDT in mind, we now discuss how measures of metacognition have evolved. In the previous section, one approach we put forward for evaluating metacognition is to examine the correlation between Type

1 performance and confidence, say, for example, with Pearson's r. Intuitively, this would seem to provide an excellent method of evaluating how well confidence ratings discriminate Type 1 performance. Unfortunately, as we have discussed, such methods are known to be subject to bias. Subjects with a different overall propensity to report high or low confidence (type 2 c) will have necessarily different correlation coefficients r, even if the ability to discriminate between Type 1 performances remains constant (type 2 d') (Nelson 1984). An alternative to the r measure (Nelson 1984) that some have advocated for solving these problems is the Goodman-Kruskal coefficient G, which provides a similar measure of correlation. This G measure has the advantage of not making any assumptions about the distributions of the data, and can be easily extended to use a rating scale rather than a high/low design. Unfortunately, despite these merits, it has ultimately been shown (Masson, Rotello 2009) that, similarly to r, G is sensitive to changes in metacognitive bias.

In light of these problems, another class of methods has been proposed. These methods are based upon comparing observed metacognitive sensitivity with a theoretical metacognitive sensitivity given under some condition. One proposed method of doing this (Galvin, Podd et al. 2003) is to compare the observed type 2 sensitivity to expected type 2 sensitivity for an ideal observer, which can be calculated from the type 1 data. This method is theoretically sound, but the computation of both required sensitivities is difficult (Maniscalco, Lau 2012), making implementation impractical. An alternative is the meta d′ approach that has been proposed by (Fleming, Lau 2014). The meta d′ method attempts to solve the same problem, but avoids running into the difficulties of Galvin et al's solution by specifying observed Type 2 sensitivity in terms of the Type 1 sensitivity that would have led to (metacognitively) ideal observer to reach the type 2 performance demonstrated. This gives us a *meta d′* that is then directly comparable to the actual d′, the Type 1 sensitivity. Since the two measures are in the same units, they are highly interpretable. Meta d′ = d′ for example indicates that a participant is making use of all the information available to them from the Type 1 task, while Meta d′ < d′ (the typical case) would indicate that they are making use of less information than an ideal observer has available, and (in theory) Meta d′ > d′ would indicate using more information than an ideal observer has available (Fleming, Lau 2014).

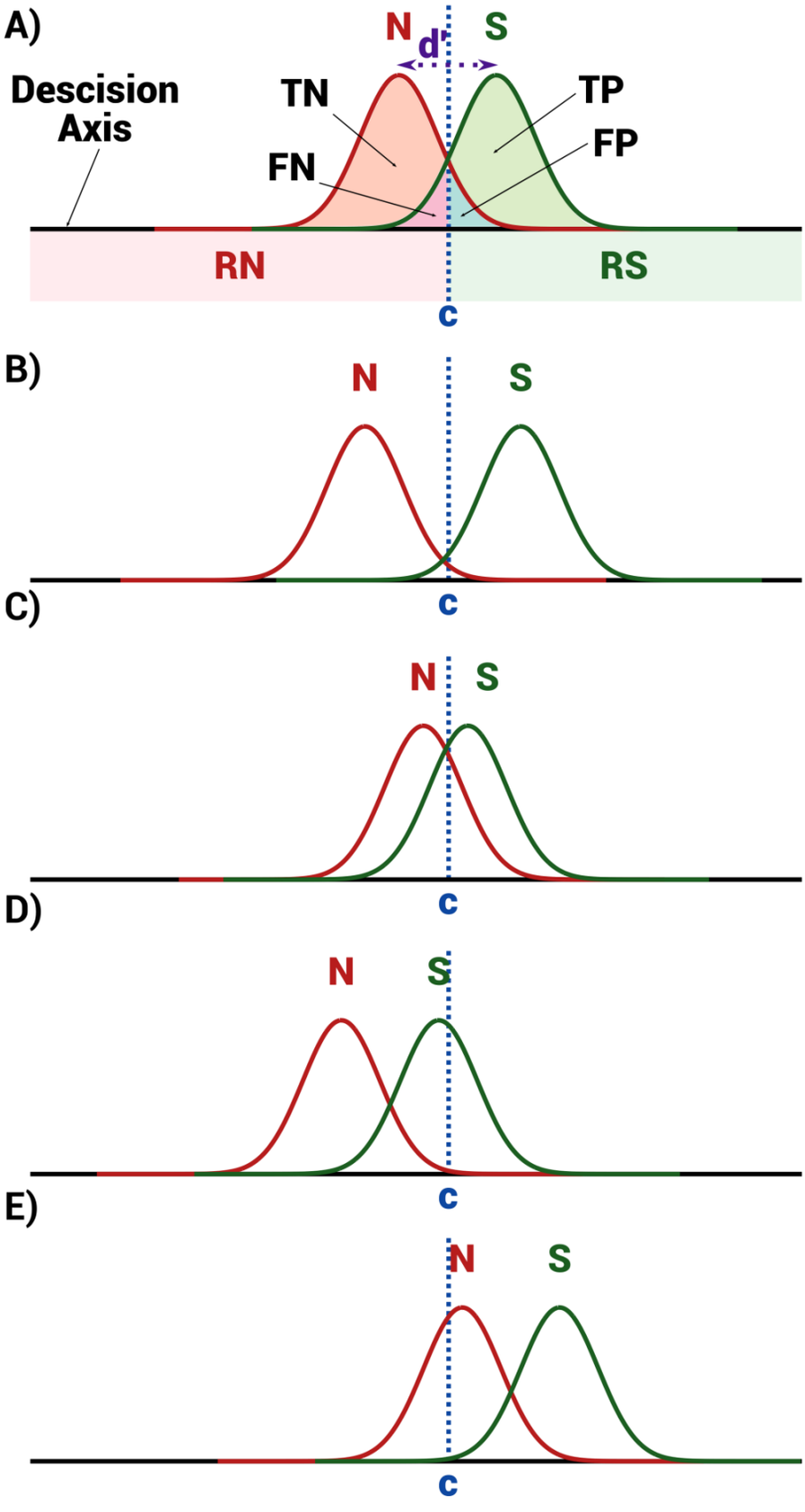


Figure 12) A) An illustration of the classic SDT framework demonstrating how each of the four possible outcomes arise. B&C) Examples of tasks with high and low discriminability respectively.

In B) the distributions for the two tasks are far apart, thus it is easy to distinguish the outcomes and d' is high. Conversely, in C), the distributions for the two tasks are close together, it is difficult to distinguish the two outcomes and d' is low. D&E) Examples of tasks with high and low bias. In D), the criterion is placed in such a way relative to the distributions that almost all responses are negative – as such there is a negative bias to responses. Conversely in E), the criterion is set up such that almost all responses are positive and there is a positive bias to responses.

3. Methods

In the previous chapter, we provided a literature background for the proposed research question. In this chapter, we cover in more detail the specific methods we will make use of. In particular, the previous chapter was about placing the research in a broader theoretical context, but in this chapter we attempt to provide more details about the methods that we will directly use throughout the thesis.

Data acquisition

Electroencephalography

The general attentional blink paradigm and behavioural measures that might be taken from it have already been introduced, but a method that we will be making use of that bears brief introduction is electroencephalography (EEG). EEG measures brain activity through electrodes placed on the scalp. The signal is a sum of the activity of various neural sources, though since there is no reason to believe that the dipoles of each source line up, the respective orientations of these dipoles may determine the signal as much as the strength of activation. Despite this, the method is very effective in identifying large groups of neurons firing at the same time in the same direction.

Unfortunately, neural sources are not the only sources that make up EEG signal. Neuronal activity inside the brain is often dominated by noise from muscle artefacts, especially those from the muscles around the eyes (Teplan 2002). In order to observe the neuronal signal among that noise, researchers often make use of filtering and averaging techniques to improve the signal-to-noise ratio of the data. In particular, to observe the signal from a certain stimulus, we average many trials with activity time locked to a certain event related to the stimulus. Assuming noise is distributed with zero mean, the noise in these trials will cancel out, improving the signal. The resulting signal after this filtering and averaging is called an Event Related Potential (ERP), and reflects neural activity related to the presentation of a certain stimulus.

Different components of these ERPs have been identified to relate to different stimuli or tasks. One that is particularly pertinent to this thesis is the P300 (P3) component. The P3 is a positive deflection peaking around 400ms after stimulus

presentation in a range of tasks. It has long been thought to represent higher cognitive functions and has been linked to working memory encoding, conscious access, and subjective experience (Vogel, Luck et al. 1998, Kranczioch, Debener et al. 2007, Pincham, Bowman et al. 2016). Notably, in the dataset we will make extensive use of throughout this thesis, from (Pincham, Bowman et al. 2016), the P3 was sampled both for working memory encoding and subjective experience, and it was found that the P3 was modulated by both, though to different degrees.

Datasets

Two dataset we extensively make use of throughout this thesis are the set of experiments first published in (Pincham, Bowman et al. 2016). This work as a whole attempts to explore the relationship between working memory encoding and subjective experience during the attentional blink. In order to do this, the authors conduct two experiments in which both working memory encoding and subjective experience are measured during the attentional blink. The first experiment attempted to examine behaviour, and purposely assessed a larger number of lags with fewer trials each in order to sample the full attentional blink curve. In contrast, the second collected both EEG and behavioural data, with only lags 1 and 3 sampled for 80% of trials in order to enhance the EEG signal strength at these lags. Here, we provide an overview of the experimental procedure. Since the data acquisition methods have already been published, we only provide an overview of experimental procedure here. A full treatment can be found in Appendix A – Detailed Experimental Procedure.

Materials and Methods

Targets were uppercase letters and distractors were single digits, each trial contained one or two targets - T1 occurred on every trial and was always presented in red, and T2 (if it occurred) was presented in white. Each RSVP stream contained 15 items. T1 randomly appeared as the fourth, fifth or sixth item in the RSVP stream. Stimulus Onset Asynchrony (SOA), the amount of time between the onset of each stimulus, was 90ms. At the end of each RSVP stream, participants were asked to rate the subjective visibility of T2 using a 6 point self-report scale. The numbers 1 2 3 4 5 6 were presented in a horizontal line on the screen, with the description ``not seen'' presented beneath the number 1 and the description “maximal visibility" presented beneath the number 6. Participants

then reported the identity of T1 and T2 (even if a second target did not occur). Participants were required to guess if they were unsure of the target identities.

All participants completed two experiments, spaced at least one week apart. Experiment 1 exclusively collected behavioural data and Experiment 2 collected both behavioural and EEG data. Experiment 1 consisted of four blocks, each with a different target/mask duration combination. The mask, if it occurred, was always the hash (#) symbol. In Block 1, targets appeared for 90msec with no mask. In Blocks 2, 3 and 4, the target/mask durations were 70msec/20msec, 60msec/30msec and 50msec/40msec respectively. In Experiment 1, T2 appeared at lags 1, 2, 3, 4, 6 or 8 with equal frequency. Experiment 1 deliberately sampled a large number of lags in order to examine the relationship between T2 accuracy and subjective visibility across the entire AB curve. Trials that did not present a second target (no-T2 trials) were also included with equal frequency (hence, one in seven trials did not contain a second target). Experiment 1 contained 4 blocks of 49 trials, totalling 196 experimental trials.

For each participant (of the 18 analysed), data from Experiment 1 were analysed to determine which of the four target/mask durations resulted in T2 being correctly reported on approximately 50% of lag 3 trials. Each participant's optimal target/mask duration was then employed in Experiment 2. As a result, 28% of participants received the 70msec/20msec target/mask duration in Experiment 2, 50% of participants received the 60msec/30msec duration and the remaining participants received the 50msec/40msec duration. Experiment 2 contained 5 blocks of 100 trials, totalling 500 trials. To maximise ERP signal strength in Experiment 2, T2 appeared at lag 1 on 200 trials, at lag 3 on 200 trials, at lag 6 on 50 trials and was absent on 50 trials.

State-trace analysis

Statistical methods for State-Trace Analysis

Classically, the challenge of state-trace analysis has been quantification. The state-trace framework creates an excellent tool for comparing the validity of different theoretical models of data, but practically applying it for the purpose of testing hypotheses has proved challenging (Loftus 2002, Newell, Dunn 2008, Bamber 1979). However, several authors have recently demonstrated the viability

of Bayesian statistical methods for solving state-trace problems (Prince, Brown et al. 2012, Davis-Stober, Morey et al. 2016), particularly for the *dependent variable* state-trace upon which this thesis is focused. Here, we provide an overview of those methods.

We have some state factor (our *dependent variable)* with two levels $S = \{S_1, S_2\}$, forming the *state* space over which we examine our question of interest, and some dimension factor (our *independent variables*) $D = \{D_1, \dots, D_n\}$, a manipulation we are performing across it. We are attempting to diagnose the monotonicity of this relationship, that is, whether the ordering of the levels of our dimension factor are either the same (or the reverse) of one another across each of the two axes of our state factor. If this is possible, we diagnose monotonicity, and if it is not possible we do not. As discussed, often we also introduce a trace $T = \{T_1, \dots, T_n\}$ factor, another independent variable. Overall, we must consider each combination of $Q = D \times T!$ orderings for each axis and Q^2 joint orderings. For reference, we previously gave a visual example of both monotonic and non-monotonic state-trace plots in Figure 5A) and Figure 5B) respectively.

At this point, the set of Q^2 joint orderings corresponds to the whole space of possible configurations of the state-trace graph, and currently it can be divided into two different partitions. These are the non-monotonic orderings and the monotonic orderings. With respect to our Bayesian statistics, we are attempting to choose between the monotonic model consisting of all monotonic orderings, and our non-monotonic model consisting of all other (non-monotonic) orderings. To do this, we calculate a Bayes factor expressing how much the data has changed our preference between our two models. This is the measure of the ratio of evidence for each model. Explicitly, denoting our data as y, the prior probabilities ($P(x)\ where\ x = M\ or\ NM$) as π_M and π_{NM} for the monotonic and non-monotonic models respectively, and the posterior probabilities ($P(x|y)\ where\ x = M\ or\ NM$) as $\pi_M^{(y)}$ and $\pi_{NM}^{(y)}$, we calculate the Bayes factor as:

$$BF_{M/NM} = \frac{\pi_M^{(y)}}{\pi_{NM}^{(y)}} \Big/ \frac{\pi_M}{\pi_{NM}}$$

We follow (Davis-Stober, Morey et al. 2016) in referring to this calculation as $BF_{M/NM}$, the bayes factor comparing the monotonic versus non-monotonic models.

Implicitly so far, we have been making use of a completely uniform prior, effectively assuming all possible orderings of all combinations $D \times T$ across the levels of the state factor are equally likely. In many data sets, this is clearly not true, for example we may have strong prior expectations about the behaviour of the attentional blink. Previous work has approached this problem by using the prior to assert a-priori that certain constraints on the behaviour in the data are true. For example, in (Davis-Stober, Morey et al. 2016) the authors pre-suppose that dual task performance will always be worse than single task performance in their analysis of a data set from (Sense, Morey et al. 2017). Such constraints take the form of setting the prior belief of any orderings that contradict them to zero. Typically, these constraints are discussed with respect to the trace factor which, if introduced, is typically selected to be a factor independent of the question of interest, selected to sweep out the behaviours of the system and whose behaviour is known in advance. We denote the Bayes factor monotonicity vs non-monotonicity with a prior set such that some given set of constraints on the trace factor are true as $BF_{(M/NM)|T}$. Though constraints on our trace factor are often selected for their theoretical merits, their validity in practice, must be quantified. For this, we introduce the measure $BF_{T/N(T)}$: the Bayes factor comparing the "trace true" models for which the constraints on the trace factor are true versus the "trace false" models in which they are not (Davis-Stober, Morey et al. 2016). A similar measure can be defined for "dimension true" models that assume some constraints on the dimension factor are true ($BF_{D/N(D)}$), and, if constraints exist for both the trace *and* dimension factor, their intersection ($BF_{T\&D/N(T\&D)}$) (Davis-Stober, Morey et al. 2016).

We must also consider how to apply this type of analysis across a group of participants. As we have discussed, state-trace analysis does not work well with approaches based on averaging. In particular, it is possible both to average multiple non-monotonic datasets into a monotonic dataset, and multiple monotonic datasets into a non-monotonic one. A simple alternative analysis is the grouped Bayes factor introduced by (Prince, Brown et al. 2012). This method treats each of our participants (of which there are M) as independent from one another and calculates the group Bayes factor as the product of each individual Bayes factor:

$$GBF = \prod_{i=1}^{M} BF_i$$

As long as participants are independent samples and the results are reasonably homogeneous (all monotonic or non-monotonic), this grouped Bayes factor is a good measure of the group level effect. In the literature review, we also introduced the aggregated Bayes factor method as a method for analysing the homogeneity of the data. Though we will not end up using this in forthcoming analysis because the data is so strongly homogeneous and the method can only effectively quantify non-homogeneity, we discuss it briefly here for completeness. As we have said, approaches based on averaging do not work well with state-trace analysis because it is possible to average sets of completely monotonic results into a non-monotonic average, and vice-versa. (Davis-Stober, Morey et al. 2016) find a solution by observing that while the average of multiple monotonic datasets will not necessarily lie within the *union* of all possible monotonic datasets, it will with certainty lie within the convex hull of those orderings. The opposite is not true, lying within the convex hull is not sufficient support to demonstrate that the data is uniformly monotonic. This method then, is useful for demonstrating when monotonicity has been violated, but cannot be used as evidence to support monotonicity. In this instance, the authors propose individual level inspection of the data.

The Simultaneous Type/Serial Token model

In a previous chapter, we introduced the Simultaneous Type/Serial Token model of Attention. Later chapters will make extensive use of this model as well as the virtual ERPs it can produce; we therefore now briefly describe the models architecture and function in more detail.

Architecture

So far, we have described an STST model based on a type/token distinction that consists of two stages, a blaster that provides attentional enhancement, and a binding pool that allows types to be bound to tokens. Here we discuss some more detail about the architecture of these components, and how they relate to one another.

The first stage of the model manages the types (see Figure 13) and consists of four layers supporting different aspects of visual processing: the input layer, the masking layer, the item layer and the task-filtered layers. The second stage of the model governs the tokenisation process, and consists of the binding pool and the tokens. Items first arise in the input layer, and then pass through the masking layer, which implements masking through lateral inhibition. From here, items enter the item layer, which creates a brief, self-sustained representation. Then, the final layer of the stage: the task filtered layer, provides a salience filter that excites task relevant nodes while inhibiting others. From the task filtered layer, sufficiently active items can activate tokens through the binding pool, and become bound to them through a tokenisation process. This tokenisation process takes several hundred milliseconds, though it is shorter for more active items. In order to reach sufficient activation to achieve this binding however, most stimuli will need to benefit from the blaster. When an item becomes sufficiently active in the task filtered layer, the blaster provides a brief, powerful enhancement to the entire task filtered and item layers that allows items to reach the threshold for tokenisation more easily. During this process, a powerful inhibitory signal holds the blaster low to prevent it from re-firing and corrupting the tokenisation process: it is this inhibition of the blaster that generates the attentional blink. A walk through of how an individual item becomes encoded into working memory can be seen in Figure 13.

Virtual ERPs

The STST model can also be used to simulate virtual ERPs. Virtual ERPs are calculated from a computational implementation of the STST model, neural-STST (Bowman, H., Wyble 2007, Craston, Wyble et al. 2009). Of particular interest to us is the Virtual P3 that the model can calculate. As in the STST model described above, the neural STST model is organised as layers of nodes, connected via weighted connections. These connections are the analogue of synaptic projections in the brain. To calculate these virtual ERPs, we introduce the concept of excitatory post synaptic potential to these virtual nodes.

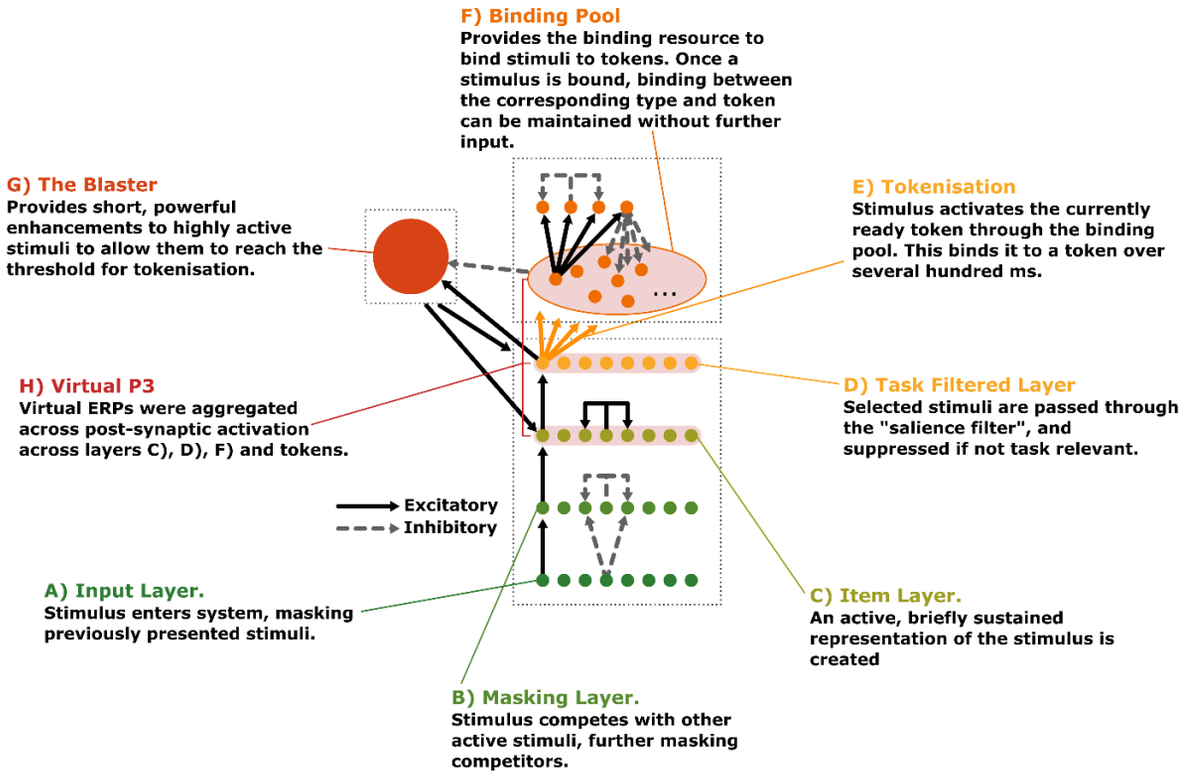


Figure 13) The Simultaneous Type/Serial Token model. A) Input Layer. Stimuli enter the system through this layer. As well as providing input, this layer implements backward masking through inhibitory connections to all other stimuli in the masking layer. B) Masking Layer. Simulates further masking dynamically through lateral inhibitory connections to all other stimuli. These lateral inhibitory connections are weaker than the forward ones from the input layer, such that backward masking is stronger than forward masking. C) Item Layer. Creates a temporary representation of a stimulus through self-reinforcing connections. D) Task Filtered Layer. Implements a "salience filter" to filter out task irrelevant stimuli, by enhancing task relevant stimuli, and suppressing others. E) Tokenisation. When a stimulus has reached an appropriate level of activation, it excites the currently ready token through the binding pool. In a process that takes several hundred ms, the token is bound to the type. Once this binding has occurred, the type-token connection can be maintained without any further input. F) The Binding Pool. Contains the binding resources that enable stimuli to bind to tokens. G) The Blaster. Provides a short, powerful enhancement to items in the item and task filtered layers when there is sufficient activation in the task filtered layer to begin the tokenisation process. While the tokenisation process is ongoing, a powerful inhibitory signal from the binding pool prevents the blaster firing again. H) Virtual P3. A virtual P3 can be generated from the STST model from the excitatory post synaptic potentials of the item layer, the task filtered layer, and a subset of the tokens and binding pool (the token gates and the binder gates).

The virtual P3 is then calculated as the sum of these excitatory post synaptic potentials across a subset of the layers. In this thesis, we follow previous work in using the 3rd, 4th, 6th and 8th layers of the neural-STST model, corresponding to the item layer, the task filtered layer, the binder gates and the token gates. These are the layers highlighted in red in Figure 13. As in previous work (Craston, Wyble et al. 2009), we also implement a retinal delay of a model equivalent of 70ms. Compared to previous works using virtual ERPs from the STST model, we selected a slightly different stimulus range over which to calculate this virtual P3 in order to provide a thorough exploration of our hypothesis. Specifically, we

sample a range of stimulus strengths with greater variability (-0.078 to +0.078 -> -0.1625 to +0.1625), at a slightly higher average stimulus and distractor strength (0.520 -> 0.570). This approach is consistent with previous simulations with the STST model, where we allow input strength ranges to vary reflecting the fact that different experiments being modelled might have quite different stimulus types and sensitivities.

4. State-Trace Analysis of the Attentional Blink

Abstract

The core thesis of this work concerns the relationship of dependency between working memory encoding and subjective experience and to what extent it is possible to dissociate these processes. Of particular interest to us is whether working memory encoding is a necessary but not sufficient condition for subjective experience or, whether a working memory encoding can occur in the absence of subjective experience. Previously, we have introduced several paradigms that show some form of working memory encoding in the absence of subjective experience. One paradigm that is particularly notable in this respect is the Experiential Blink as demonstrated in (Pincham, Bowman et al. 2016), which appears to demonstrate (among other significant results) free recall of stimuli in the absence of subjective awareness.

Demonstrating that it is possible for participants to report stimuli significantly above chance level in the absence of subjective report is a significant theoretical finding, however it does not necessarily demonstrate that subjective experience and working memory are dissociated. Without knowledge of how internal subjective experience translates into subjective report on our scale, it is difficult to come to any firm conclusions of this nature. In this chapter, we therefore make use of the State-Trace methodology introduced previously that is capable of more fully assessing these questions, and apply it to assessing the evidence for a dissociation of working memory encoding and subjective experience in the Experiential Blink paradigm. In the process, we find evidence for a dissociation that is driven by working memory encoding being a necessary, but not sufficient condition for subjective experience.

Introduction

The core thesis of this work is concerned with examining to what degree it is possible to separate subjective experience and working memory encoding, and what kind of relationship of dependency exists between the two. In particular, we are interested in the possibility that working memory encoding is a necessary (but perhaps not sufficient) condition for subjective experience. That is to say,

that it is possible to encode stimuli into working memory that have not been subjectively experienced. While we introduced several paradigms in the literature review that show this pattern of behaviour to some degree, most of these paradigms showed evidence for a weaker hypothesis, i.e. providing evidence for the *influence* of unexperienced stimuli on report. One exception to this is recent work on the Experiential Blink paradigm (Pincham, Bowman et al. 2016), an attentional blink experiment measuring subjective report, which (among other results) appears to show free recall of stimuli with minimal subjective report, a phenomenon the authors call *Sight-blind Recall*.

While it is tempting to put forward this sight-blind recall effect as evidence of a dissociation, as our literature review on functional dissociation logic has discussed, such a conclusion would likely be premature. This evidence would constitute at most a single dissociation which, as we discussed in the literature review, has the potential to be misleading (Teuber 1955). Furthermore, even if we were able to obtain the type of double dissociation typically preferred in these instances, the same discussion puts forward evidence that these may also be insufficient. In this chapter, we follow up on these findings that may amount to a demonstration of sight-blind recall to try and determine whether the results constitute a sufficient basis to believe that working memory and subjective experience are dissociated, by using the state-trace logic introduced in the literature review as an alternative. However, implementing this state-trace analysis over our dataset requires some unique considerations, which we will now discuss.

The State-Trace Method

In previous chapters, we reviewed the theoretical framework of state-trace analysis, and the statistical methods by which one can quantify its effects. However, our data does not quite neatly fit this framework. One notable difference to previous analysis in the literature is that our experiment lacks a trace factor. Theoretically there is nothing wrong with this, in fact in many ways it simplifies the analysis. Typically, the trace factor is introduced to sweep out underlying behaviours of the system in question, but in our case, we already have the attentional blink to achieve this. Furthermore, dimensionality is not an issue – the number of levels of our dimension factor (6) already matches or exceeds the

number of *combined* Dimension × Trace factor levels in other analyses (Tulving 1981, Sense, Morey et al. 2016, Heathcote, Freeman et al. 2009).

Accounting for this difference and submitting the data to a cursory examination, it appears on the face of it that our data has quite a strongly non-monotonic relationship between accuracy and subjective report Figure 14. The behaviour at early lags (Lag 1, 2 and 3) is completely different from that of late lags (Lags 4, 6 and 8); at early lags accuracy increases while subjective report decreases, but at late lags they increase together (with the exception of lag 8, which is something of a serial position outlier, as we discuss later). However, this cursory analysis is not statistical quantification, and for this we need to turn our attention to how our prior probabilities are defined.

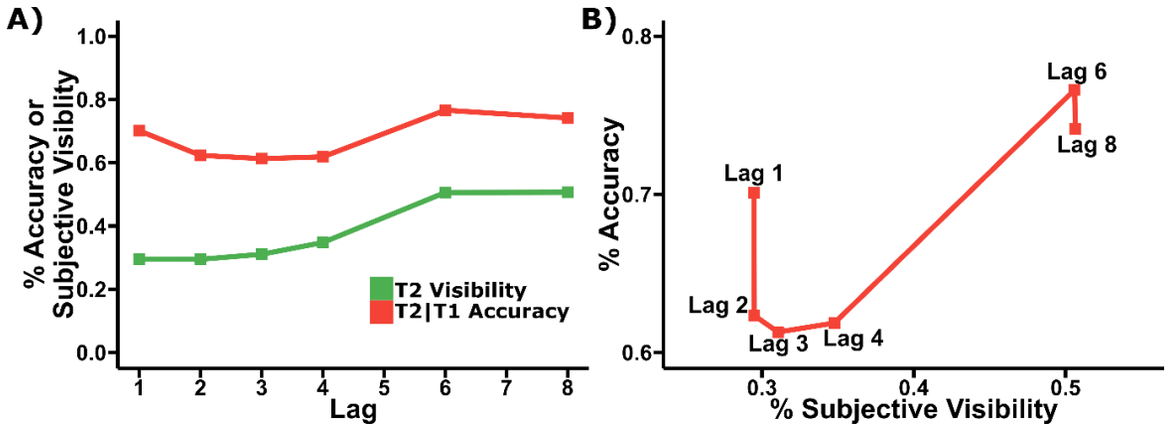


Figure 14. A) Results from (Pincham, Bowman et al. 2016), comparing accuracy and subjective visibility across lags in the attentional blink. The T2 visibility curve demonstrates what Pincham and Bowman term the Experiential blink of subjective report. B) State-trace plot comparing T2|T1 accuracy and T2 visibility from A). Note the apparent non-monotonicity of the relationship between accuracy and visibility. (Note, the T2|T1 blink curve here shows some very minor differences to that presented in (Pincham, Bowman et al. 2016). This is because T2 accuracy in the original paper was in fact presented as the accuracy of the conjunction of T2 and T1, whereas here we display the conditional probability of T2 given T1. None of our findings are impacted by the difference.

Previously, we discussed the general method for taking advantage of prior information about, for example, a trace factor. However, while we have expectations about the behaviour in the attentional blink, our situation is more complicated than previous examples in the literature. Previously, authors have been in the position to make simple statements in the prior, such as one of two levels being higher than another, or an effect being monotonic. This is not the case for the attentional blink, and setting specific ordinal qualifications of behaviour across lags in a similar manner is non-trivial. While we wish to take

advantage of as much prior knowledge as possible, the behaviour of the attentional blink is variable, and it is well established that setting a poor prior can compromise the integrity of results (Lindley 1957).

In order to make a fair analysis of the data, we therefore set the prior of our Bayesian analysis from previous literature, specifically based on the results from (Nieuwenhuis, de Kleijn 2011). This paper presents both a classic attentional blink with lag 1 sparing of report accuracy, and a similar "experiential" blink of subjective report in which lag 1 is spared a great deal less. Due to the well-established evidence for the pattern of behaviour in the attentional blink, we encoded strong expectations of behaviour, including lag 1 sparing, of the report accuracy in our data. Comparatively, the evidence for the behaviour of subjective report during the blink is less well established, so we refrained from imposing such strong constraints about it, particularly at the important lag 1 data point. We also recognise some uncertainty about the deepest point in the attentional blink: given the SOA of 90ms, we could reasonably expect either of lags 2 or 3 to be the deepest point in the blink[1]. We therefore set our prior to be consistent with several potential deepest points. Finally, lag 8 is a serial position outlier[2] in our experiment and was therefore removed from our analysis. These considerations resulted in a uniform prior subject to the following constraints across our data: for report accuracy, Lags 1, 6 and 4 would be held to be larger than Lags 2 and 3, with Lag 1 additionally being held to also be larger than Lag 4. For subjective report, lag 6 would be held to be higher than Lag 4, lag 4 higher than Lag 3, and Lag 3 higher than Lag 2. The validity of these constraints was determined by a $BF_{D/N(D)}$ calculated analogously to the $BF_{T/N(T)}$ discussed previously.

We calculate our posterior using the library provided in (Davis-Stober, Morey et al. 2016) that makes use of Laplace's method (Stigler 1986). While an analysis without a trace factor is perfectly theoretically valid, the code provided required modification to support this type of analysis. Further modification was also required to fully encode the more detailed type of priors we wished to support. A full list of changes made to the original code can be found in Appendix B –

[1] see, for example (Chun, Potter 1995, Bowman, H., Wyble 2007) in which lag 2 is the deepest point in the blink, in contrast to (Raymond, Shapiro et al. 1992) in which it is lag 3 that is the deepest point

[2] A common finding in attentional blink experiments is that in a last lag that is a serial position outlier, e.g. if there is no lag 7 and most lags in the experiment are short, participants will come to learn this regularity and optimize the allocation of attentional resources to short lags, causing lag 8 performance to be relatively low across the experiment.

Changelog to State-Trace Code Provided By Davis-Stober et al. As our results are reasonably homogeneous (all monotonic or non-monotonic) with one clear exception, we made use of the grouped Bayes factor as a measure of the group level effect.

Distribution of data

The state-trace method we are applying, based on the work of (Davis-Stober, Morey et al. 2016, Prince, Brown et al. 2012), assumes a binomial distribution of the data. This is suitable for our accuracy data which is a dichotomous variable, but not for our visibility scale that forms a multinomial distribution over 6 values. Furthermore, it is well known that asking participants to report subjective visibility on scales with more than 4 bins may lead to participants ignoring the middle bins on the scale (Sandberg, Timmermans et al. 2010). Consequently, we grouped our visibility results into two bins, a high visibility bin and a low visibility bin. To decide the fairest way of applying this split, we calculated the grouped Bayes factor comparing the validity of the constraints for each possible method of splitting the data. The results (Figure 15) clearly show that the “best" split is that of assigning the top 50% of visibility ratings to the high visibility bin and the bottom 50% to the low visibility bin.

Post-Hoc Testing for State-Trace

One facet of our state-trace analysis is that in a strict sense, it is non-constructive. Even if we should be able to conclusively provide evidence for a non-monotonic relationship, all we can infer from this is that some kind of dissociation exists - we are not able to directly infer what is driving it. However, our state-trace analysis does not exist in isolation. Figure 14 shows potential non-monotonicity is driven by differing behaviour of subjective visibility and report accuracy at early versus late lags.

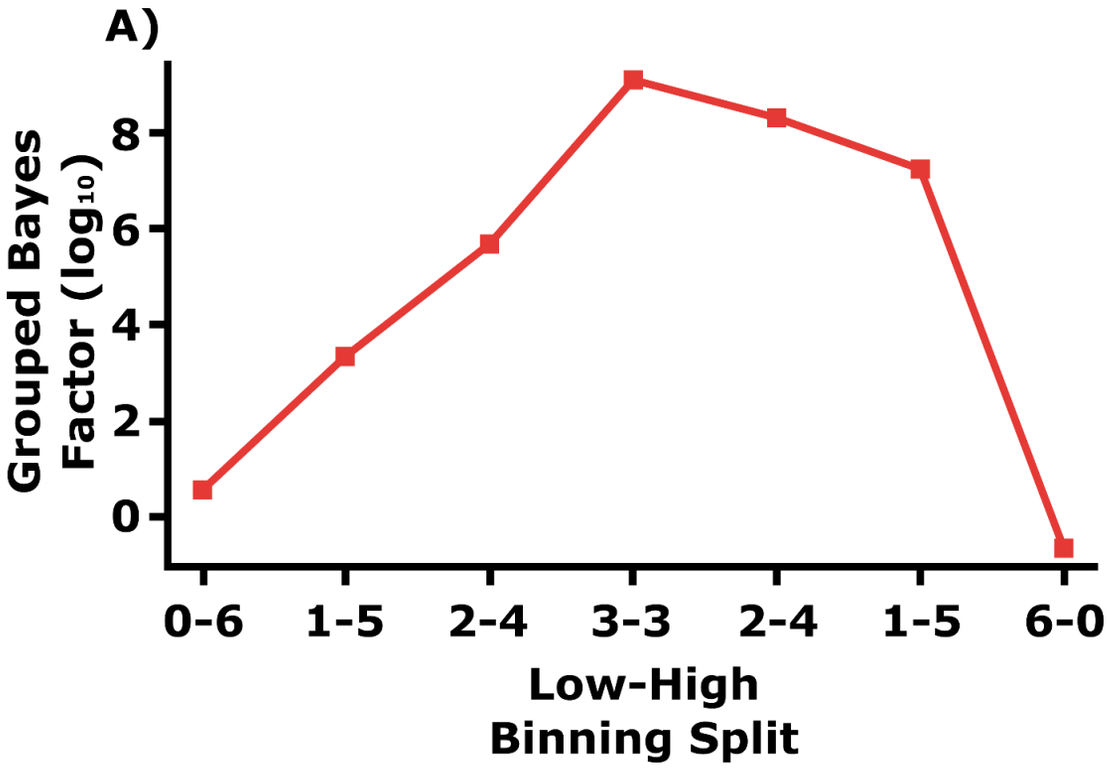


Figure 15) Grouped Bayes factor for validity across each potential binning method for high and low visibility using the set of constraints based on the data from (Nieuwenhuis, de Kleijn 2011).

At early lags, subjective visibility decreases while report accuracy increases, but from lag 3 onward, they both increase together. This would predict any dissociation as a result of working memory encoding potentially being a necessary condition for subjective experience, but not a sufficient one. We can quantify this effect by evaluating how much this differing behaviour at early lags to late lags is contributing to any overall monotonicity calculated, by rerunning the state-trace analysis with these lags left out, one at a time. This gives a measure of how much the lags are contributing to the overall non-monotonicity calculated. This method necessarily requires a change of the definition of the prior used in the Bayesian analysis. In this case, we simply exclude any conditions we placed on the prior that are no longer valid.

State-Trace Results

We first present the results of the state-trace analysis with all lags included.

State-trace results

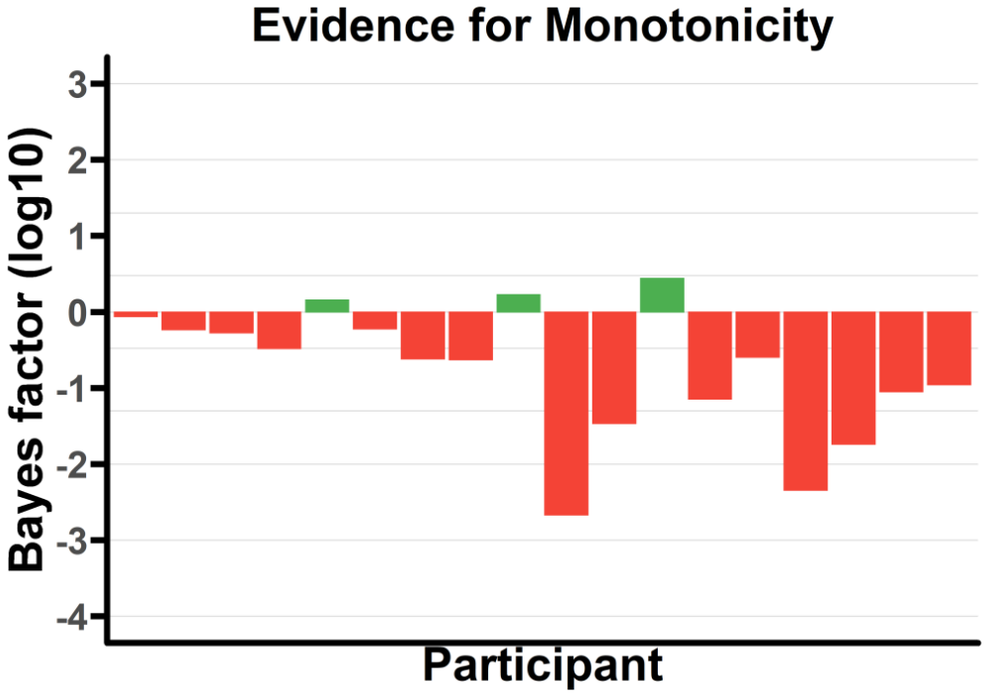


Figure 16) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) across participants. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

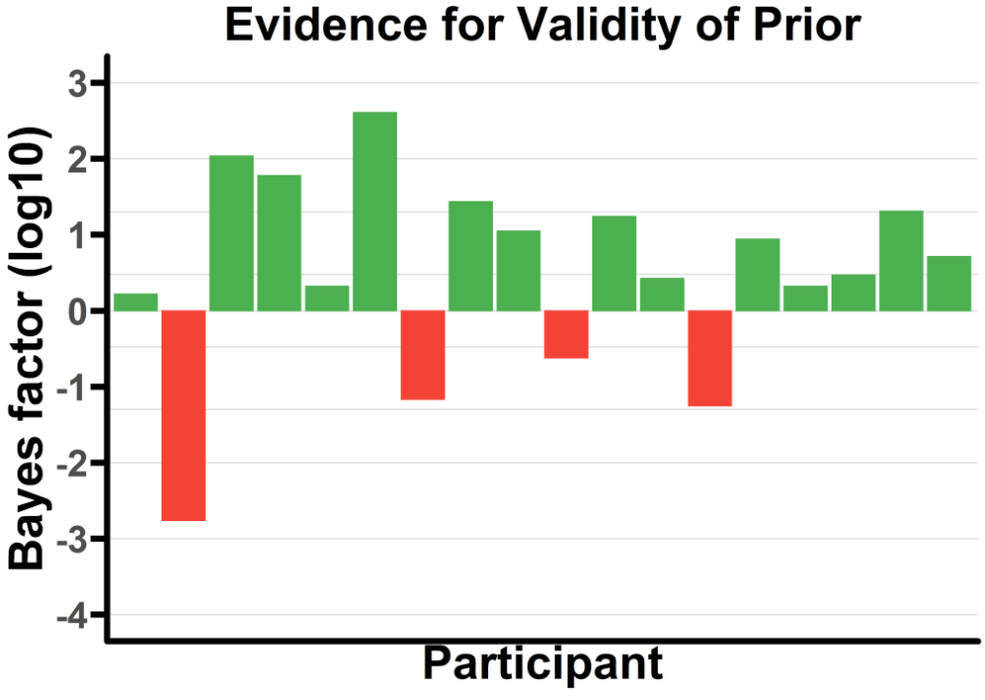


Figure 17) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) across participants. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Figure 17 shows validity for each participant for the set of prior constraints derived from (Nieuwenhuis, de Kleijn 2011). At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 1.22 \times 10^{9}$. However, we note that while the group validity is strong, four participants show the opposite pattern. Figure 16 shows the respective non-monotonicity for this set of constraints. Results are strongly and almost homogenously in favour of the non-monotonic model, with grouped (*not* log) $BF_{M/NM|(D)} = 2.25 \times 10^{-14}$.

Post Hoc Testing

In order to establish the effect of each lag on the final calculation of monotonicity, we reran the state-trace analysis with each of the lags excluded in turn, as well as lags 1 and 2 removed together. We note that in our first case our data is completely heterogeneous, rendering our GBF measure redundant, however this does not affect any of our conclusions

Lag 1

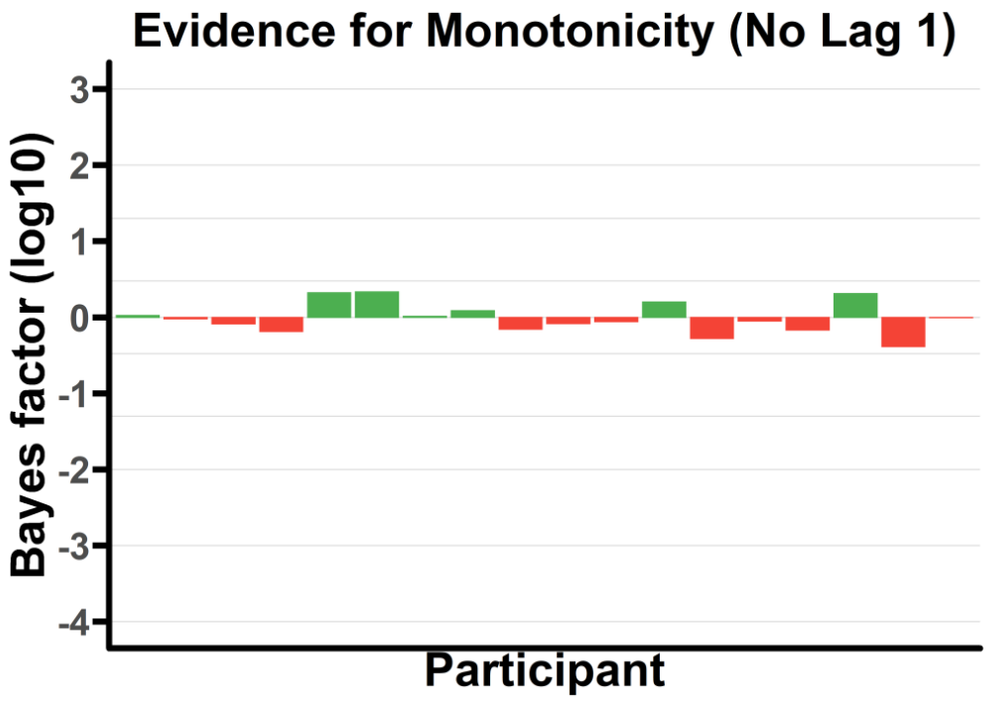


Figure 18) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) across participants with lag 1 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

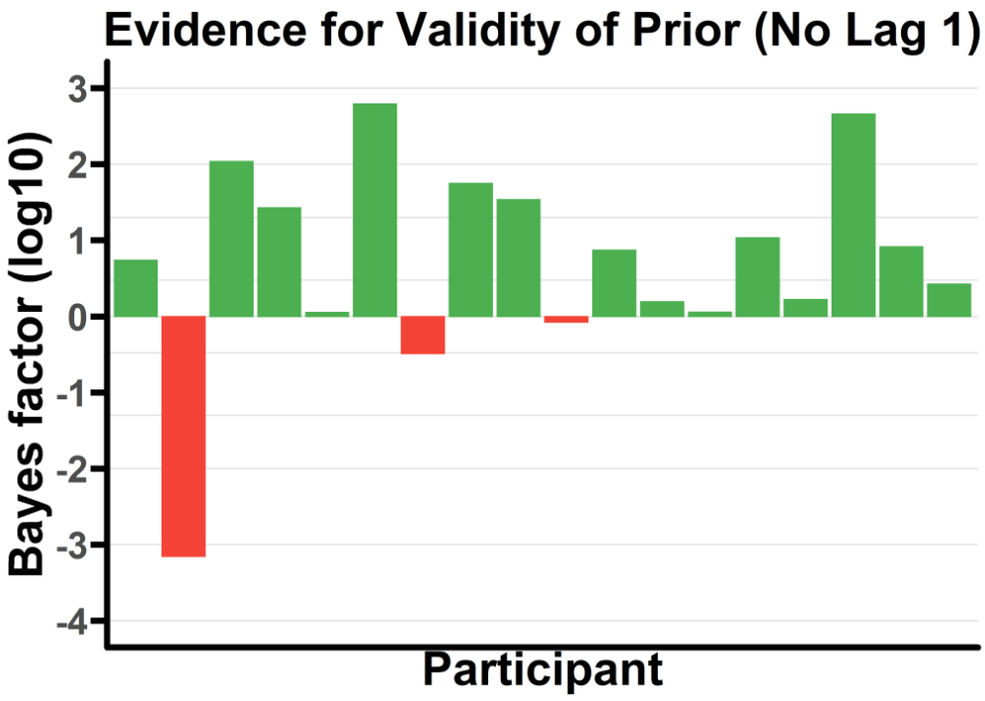


Figure 19) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) across participants with lag 1 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Lag 2

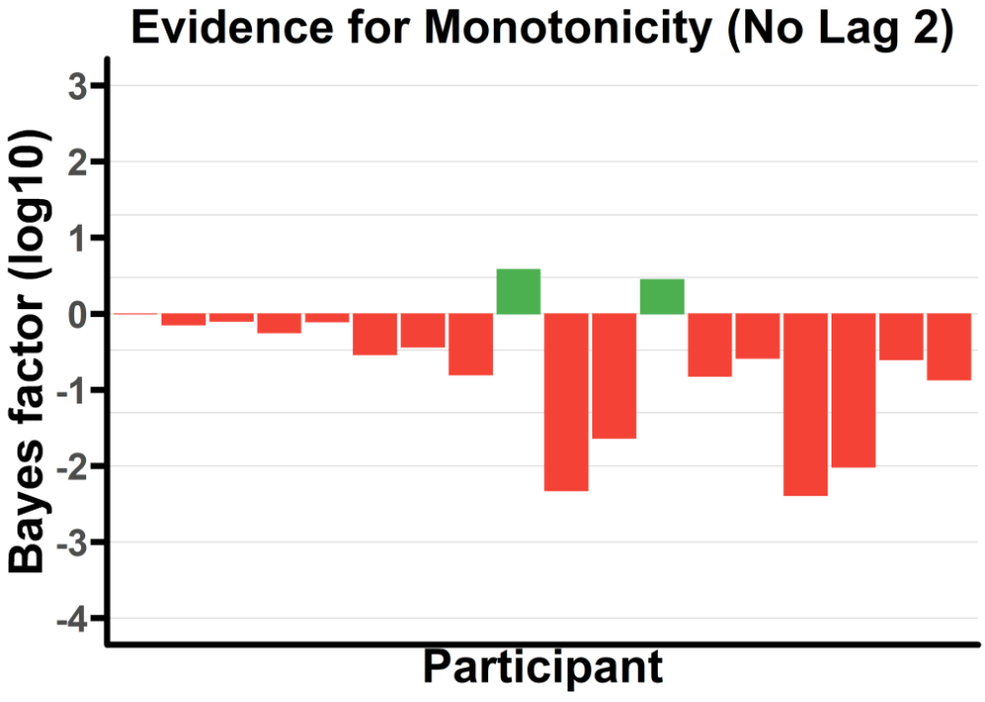


Figure 20) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) across participants with lag 2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

.

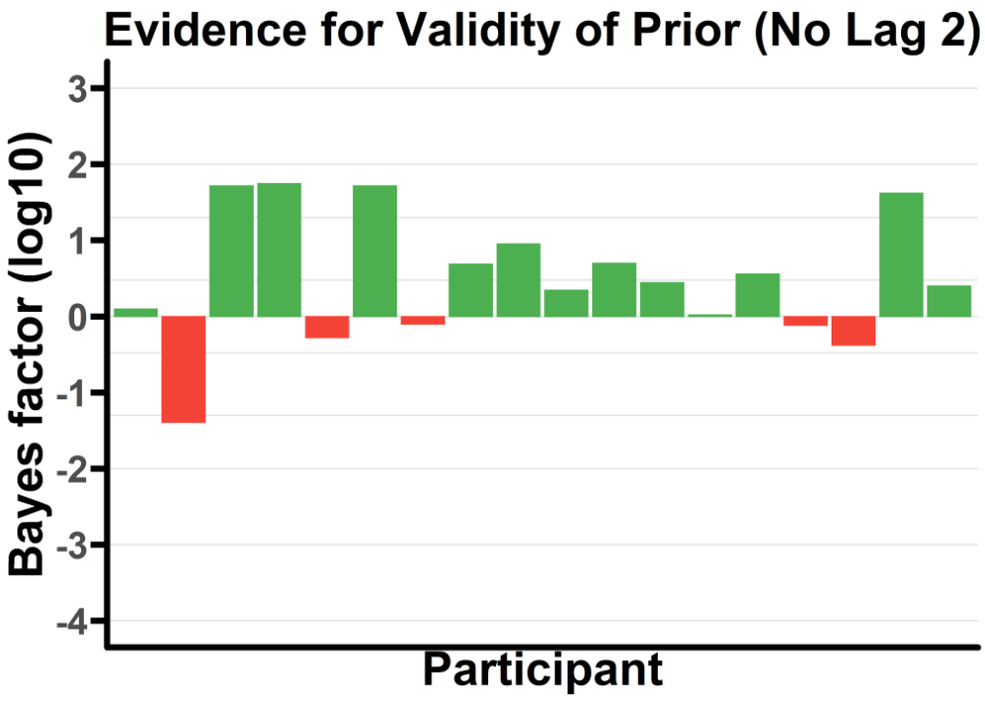


Figure 21) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) across participants with lag 2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Lag 3

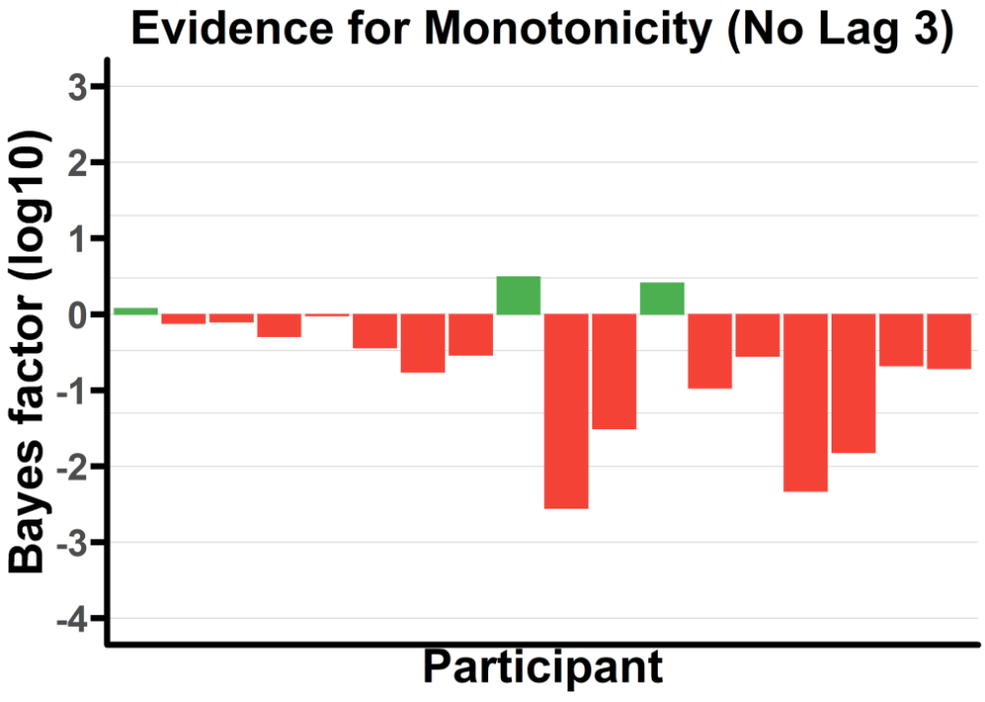


Figure 22) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) across participants with lag 3 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

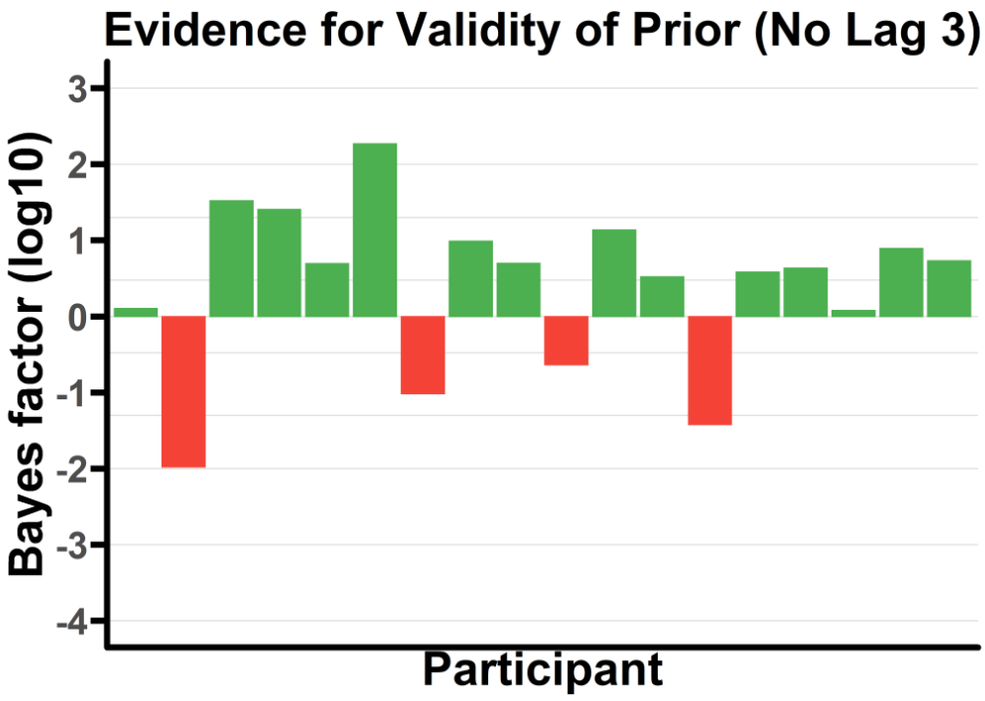


Figure 23) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) across participants with lag 3 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Lag 1&2

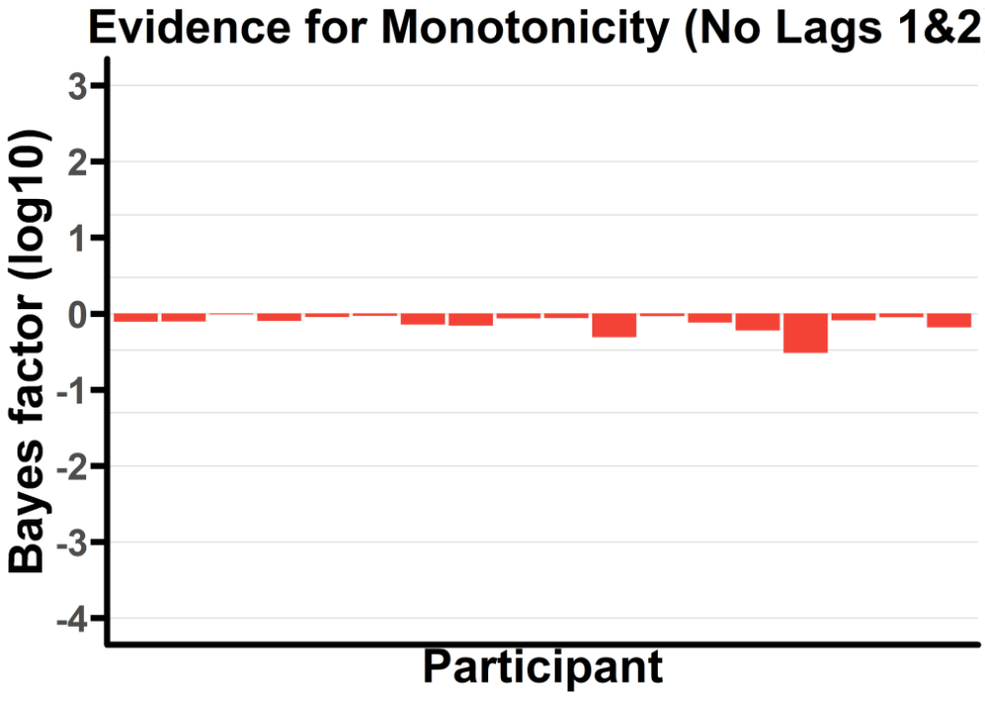


Figure 24) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) for empirical priors across participants with lags 1&2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

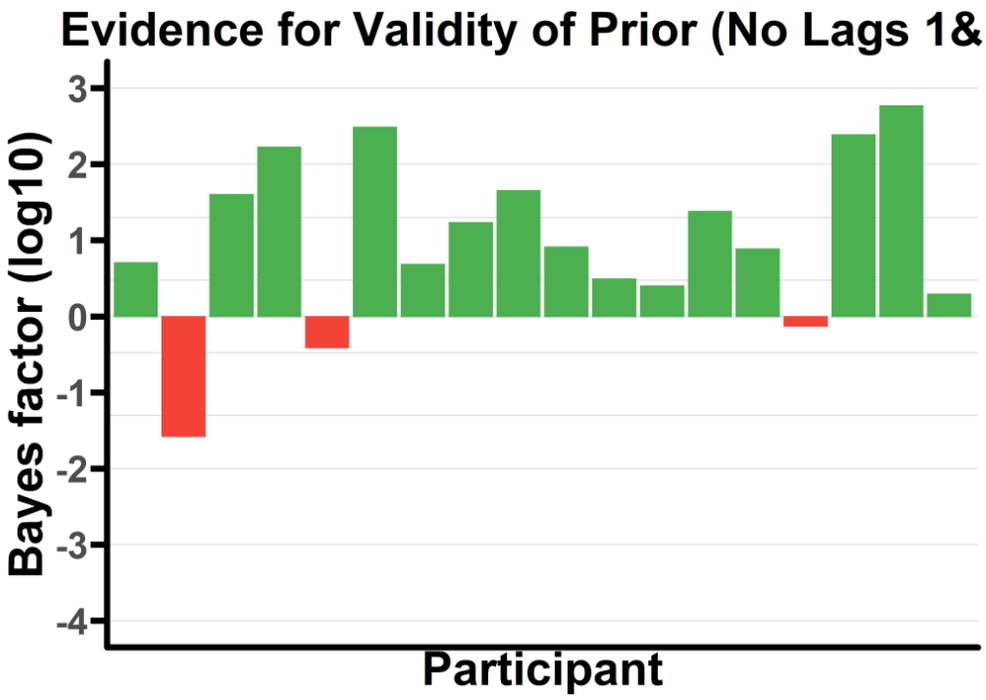


Figure 25) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) for empirical priors across participants with lags 1&2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Figure 19 shows validity for each participant for the set of prior constraints derived from (Nieuwenhuis, de Kleijn 2011) with lag 1 removed. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 1.01 \times 10^{13}$. Group validity is stronger than previously, but with more participants showing no evidence either way. Figure 18 shows the respective non-monotonicity for this set of constraints. The results show no strong preference for monotonicity or non-monotonicity and are almost completely heterogeneous, with grouped (*not* log) $BF_{M/NM|(D)} = 6.69 \times 10^{-1}$.

Figure 21 shows validity for each participant for the set of prior constraints derived from (Nieuwenhuis, de Kleijn 2011) with lag 2 removed. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 5.48 \times 10^{9}$. Figure 20 shows the respective non-monotonicity for this set of constraints. The results show a strong preference for non-monotonicity and are almost completely homogenous, with grouped (*not* log) $BF_{M/NM|(D)} = 2.65 \times 10^{-13}$

Figure 23 shows validity for each participant for the set of prior constraints derived from (Nieuwenhuis, de Kleijn 2011) with lag 3 removed. At the group level,

the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 1.70 \times 10^{8}$. Figure 22 shows the respective non-monotonicity for this set of constraints. The results show a strong preference for non-monotonicity and are almost completely homogenous, with grouped (*not* log) $BF_{M/NM|(D)} = 3.90 \times 10^{-13}$.

Figure 25 shows validity for each participant for the set of prior constraints derived from (Nieuwenhuis, de Kleijn 2011) with lags 1&2 removed. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 1.02 \times 10^{18}$. Figure 24 shows the respective non-monotonicity for this set of constraints. Results show a preference for non-monotonicity and are almost completely homogenous, with grouped (*not* log) $BF_{M/NM|(D)} = 7.02 \times 10^{-3}$.

Discussion

Monotonicity versus Non-monotonicity

Our state-trace analysis, comparing the measures of accuracy and subjective report in the attentional blink, found strong evidence for a non-monotonic model of the relationship between these two measures at both the individual participant and group level. Previous literature (Davis-Stober, Morey et al. 2016) has advocated the use of both the Grouped Bayes Factor (GBF) that we have calculated, as well as an Aggregated Bayes Factor (ABF) to confirm the homogeneity of the results, something we have not done. There seems little need to apply the ABF since its results are clear from the outset, especially since the ABF would require us to examine the individual level data to confirm homogeneity anyway (Davis-Stober, Morey et al. 2016). Our data always falls into one of two categories: complete heterogeneity (No lag 1 $BF_{M/NM|(D)}$) or substantial homogeneity (All other conditions).

One aspect of our analysis that is notable is the lack of a trace factor. However, the introduction of a trace factor is only required in the case in which there are only two levels of the dimension factor; in other cases, the introduction of a trace factor is a convenience designed to sweep out the behaviour of a system. In our case, we have 5 levels of our dimension factor, which is very close to, or exceeds the *combined* total trace × dimension factors in other state-trace experiments

(Tulving 1981, Sense, Morey et al. 2016, Heathcote, Freeman et al. 2009). There is also no need to introduce a trace factor to help sweep out behaviour; the attentional blink paradigm has already been introduced to do this with clear results.

Working Memory encoding without Subjective Experience?

Our results suggest some kind of dissociation between working memory encoding and subjective experience. Despite this, based on our initial results, we have only demonstrated that a dissociation exists and have not definitively characterised it: of itself, this finding does not enable us to say what relationship of dependency exists between working memory encoding and subjective experience, if any.

Earlier in this chapter, we hypothesised that any non-monotonicity we found would be the result of differing behaviours at early lags versus late lags. Further, we noted that this pattern of behaviour at our early lags (low subjective report, high accuracy) is consistent with working memory encoding being a necessary but not a sufficient condition for subjective experience. We therefore propose that if early lags were a substantial contributor to overall non-monotonicity, this would be indicative of a dissociation of this type. We find strong evidence for this pattern of results in the data. Removing either of lags 2 or 3 resulted in a substantial reduction of the total effect of non-monotonicity by 3-4 orders of magnitude. However, most compellingly, the removal of the lag 1 data point at which the difference between accuracy and subjective visibility is most strong causes a total breakdown of any non-monotonic effect in the data resulting in no significant non-monotonicity at all.

On the basis of these results, and the sight-recall idea that we discuss in the literature review, we propose that these findings provide evidence for a dissociation between working memory encoding that is the result of working memory encoding being a necessary condition for subjective experience, but not a sufficient one[3]. Specifically, it suggests that it may be possible to have working memory encoding in the absence of subjective experience. We follow (Pincham, Bowman et al. 2016) in referring to this phenomenon as *sight-blind recall*.

[3] Although the existence of phenomenological awareness would mean working memory encoding was also not necessary for conscious perception

However, we do note that while our results are strongly indicative, this assessment is created on the basis of the average behaviours of accuracy and visibility by subject. Strictly, to establish with certainty the relationship of dependency between working memory encoding and subjective experience, we need to assess not their respective average behaviours, but the coupling of these two measures by lag. This is addressed in a later chapter, but for now we simply note that it is difficult to imagine on the basis of our current results, a plausible pattern of coupling that is inconsistent with our current conclusions.

Importantly, *sight-blind recall* is different from more familiar notions of preconscious processing such as subliminal priming, implicit perceptual learning as well as related findings demonstrated with continuous flash suppression (Hsieh, Colas et al. 2011) and phenomena such as blindsight (Marshall, Halligan 1988), or episodic face recognition (Heathcote, Freeman et al. 2009). These experiments demonstrate only an indirect effect on a later test; in no case is the "invisible" stimulus that is not consciously perceived directly reportable. We would argue that these results are not strong enough to demonstrate the "sight-blind recall" that we have described, indicating instead *influence* without experience. In contrast to this, our results suggest the potential for *free recall* of a stimulus that has not been consciously perceived, a much stronger result that we would argue, if justified, does constitute enough evidence for "sight-blind recall" and working memory encoding without subjective experience.

There are several pieces of work that present findings consistent with our results. Firstly, evidence of working memory maintenance without conscious awareness (Soto, Silvanto 2014) sits very nicely with our results, and this is even more the case for such a demonstration with the attentional blink (Bergstrm, Eriksson 2014). Our results may help explain how these pre-conscious working memory traces arise by giving them a mechanism through which they can be encoded without subjective experience.

(Lau, Passingham 2006) also present experimental conditions in which they are able to use metacontrast masking to vary the subjective report of consciousness, while stimulus discriminability is maintained. Further, the authors find that as SOA decreases (down to around 50ms, at which point the effect reverses) shorter SOA's result in lower subjective report, consistent with our finding that subjective

report drops as T1 and T2 become closer. (Lau, Passingham 2006) is a landmark study; our results, though, move beyond their work by applying state-trace analysis rather than single dissociations, and by considering identification with free recall, rather than two alternative forced choice decisions. In this sense, our objective behaviour relies upon a significantly more complex cognitive process.

Taking our results along with those from (Block 1995, Bronfman, Brezis et al. 2014, Vandenbroucke, Fahrenfort et al. 2014) that indicate some degree of perception without reportability, it is tempting to conclude that working memory encoding and perception are highly correlated but mutually dissociable processes. However, all of the studies above provide their evidence in the form of the single dissociation, evidence that we have thus far held up as neither strictly necessary nor sufficient to conclude a separation of mental processes. Overall, the dual question to the one studied in this paper remains debated, perhaps further state-trace analysis could benefit this debate.

From a theoretical point of view, it is interesting that perception is most taxed at Lag 1. As we have discussed, (Pincham, Bowman et al. 2016) note that this pattern of behaviour is consistent with a model of the attentional/experiential blink in which stimuli are consciously perceived in a serial manner, but encoded in a simultaneous manner. This is discussed in further detail below.

Integrated Percepts

Another potential criticism of our results is that the low subjective report at Lag 1 is caused by the rather unique nature of the Lag 1 data point. Lag 1 is the only data point without any intervening distractors, and is, notably, by far the most vulnerable point to order errors (Wyble, Brad, Bowman et al. 2009), or integration of both targets into one perceptual episode (Simione, Akyrek et al. 2017). In this case, the poor report of subjective experience of T2 might be confounded by the presence of T1. Participants might report poor T2 visibility not because T2 was not vividly experienced, but because the experience of T1 in the same perceptual episode causes confusion. Part of the reason we introduced our state-trace analysis was to remove confounds of this type, however to remove any doubt or ambiguity of our results, we argue as follows.

The most compelling reason to believe this is not the case is recent results indicating that fully integrated percepts, in which T1 and T2 are perceived as one episode, show comparable subjective visibility to those in which T1 and T2 are perceived separately, and in the right order (Simione, Akyrek et al. 2017), while reversals and partial order errors score below this. This is likely to be the case even more strongly in the experimental procedure in (Pincham, Bowman et al. 2016), because it purposely gave T1 and T2 distinguishing features (by colour marking T1), and identifies the targets by these. Given this, as long as the data containing order errors is excluded from our analysis (as it has been), it is unlikely that integrated percepts can explain away our effect.

We additionally note that there are an unusually small number of integrated percepts in the experiment of (Pincham, Bowman et al. 2016). The colour marking of T1 in this experiment reduced the classical indicator of integrated percepts, order errors, from 30% in classic letters/digits tasks (Chun, Potter 1995) to approximately 10% in the task from (Pincham, Bowman et al. 2016). Further, we note that the pattern of behaviour we see at lag 1, with low subjective report and high accuracy is also visible to a lesser extent at lags 2 and 3, in which there are intervening distractors.

Conclusion

In this chapter we examined the evidence for a dissociation between working memory encoding and subjective experience, by making use of the tools of state-trace analysis to quantify evidence for this dissociation during the Attentional/Experiential Blink. Our data stands for the existence of this dissociation, and points toward this occurring as a result of working memory encoding as a necessary but not a sufficient condition for subjective experience, providing further evidence to support the hypothesis of Sight-Blind recall proposed previously.

5. Methods in State-Trace Analysis

Abstract

In our previous chapter, we made use of state-trace analysis to provide evidence for a dissociation between working memory encoding and subjective experience during the attentional blink (AB). However, the statistical quantification for our state-trace analysis was entirely based on Bayesian statistics, the validity of which can be strongly affected by the choice of prior. Notably, our prior was set based on the average results from previous experiments, but it is known that there is significant variance from this average among individual subjects during the AB. By asserting this prior, we are both in danger of precluding patterns of behaviour that would be present for an individual subject, as well as including behaviours that are not necessarily reflected by our data.

While setting a prior based on previous literature was in all likelihood a reasonable approach, which we took care to validate in the previous chapter, for the above reasons, we also explore a better method for the prior selection problem. For this, we develop a new method by which a prior extracted from previous literature can be improved to fit the data more accurately. This is achieved by using a data driven approach that makes use of a contrast independent to the question of interest to remove constraints on the prior that do not match the data. In this chapter, we define our new approach, demonstrate its good behaviour and reapply the analysis from the previous chapter based on the constraints it gives us.

Introduction

In the previous chapter, we evaluated the validity of our state-trace analysis through Bayesian statistics. However, it is well known that Bayesian statistics can be compromised by poor choices of prior (Lindley 1957). In the previous chapter, we attempted to avoid this problem by selected a prior based on constraints derived from the results from previous literature (Nieuwenhuis, de Kleijn 2011), validated by our measure of validity $BF_{D/ND}$. We found that across the group, the validity of our constraints was high and quite strongly homogeneous, indicating that the prior used in the previous chapter – and thus the conclusions based on it – were reasonable.

However, despite this positive evidence for the prior, there are clear exceptions to its validity in the data, with some subjects showing more than marginal evidence that the constraints do not fit. There are also theoretical reasons we may be uncertain about the fit of the constraints – our constraints are based on an average behaviour, however individual subjects can and do vary from this average considerably. This variation is not necessarily problematic, but constraining our data on the assumption that this average behaviour is uniformly true may be a poor choice for a prior.

While, given the positive evidence for the prior used in the previous chapter, we remain confident that its results are valid, it is reasonable that we may also wish to find a better method. If we wish to improve the fit of the prior for our specific data set, then the only way to do this is with reference to the data itself. This causes several problems. Firstly, we must be careful that our method is independent from our final measure of interest, and does not bias our final result. Secondly, we must also be careful to not “overfit" and end up setting a prior that fits the noise present in the data as well as the signal. In this chapter, we discuss these problems, and develop a new approach to address them within these bounds. We then go on to demonstrate the good behaviour of our new empirically derived prior, and reapply the state-trace analysis from the previous chapter using it.

The Problem

When approaching the Bayesian analysis of our state-trace experiments, we are required to set a prior, denoting our expectation of behaviour before we have examined the data. While Bayesian statistics are in general quite robust to variations in the prior (Liu, Aitkin 2008), it is well known that setting a sufficiently “poor" prior can influence the results. The canonical demonstration of this is Lindley's Paradox (Lindley 1957) in which it is possible to cause the Bayesian and frequentist approaches to hypothesis testing to disagree by choosing a suitable prior (LaMont, Wiggins 2016). One approach that is often taken in the absence of any conclusive prior beliefs is to set a type of prior known as an uninformative prior. Such a prior makes no strong statements about our variables, and assigns equal, or close to equal, probabilities to each of our possible outcomes. Such an approach has been proposed for state-trace analysis, on the basis that it is

simple to compute, and as long as all possible orders are assigned some non-negligible probability then the Bayes-factor estimates will be insensitive to the choice of prior (Prince, Brown et al. 2012).

However, when there do exist concrete prior beliefs about the data, this approach is suboptimal. We may for example, a-priori, believe that some orderings of variables in our state-trace analysis may *only* be the result of a measurement error. In this case, the only appropriate prior probability is one that precludes this ordering entirely, an exemption that *will* significantly affect the Bayes Factor calculated. In state-trace analysis, this has been demonstrated, for example, by (Davis-Stober, Morey et al. 2015) in their analysis of data from (Sense, Morey et al. 2017). The data from (Sense, Morey et al. 2017) examines visual working memory, and it is asserted that no theorist would believe that performance would increase alongside the number of items to be remembered. The authors therefore entirely preclude orderings that violate this by setting their prior probability to zero, and note a significant change in their calculated Bayes factors with this prior vs the uniform prior as a result.

This approach of constraining orderings is clearly effective, and in many cases justified but presents some problems in the data we analyse. In contrast to, for example, our previous example from (Sense, Morey et al. 2017), the Attentional Blink paradigm does not lend itself so well to precise ordinal assertions about our variables. There is significant subject-by-subject variation in the attentional blink to the extent that some subjects (non-blinkers) do not demonstrate the effect (Martens, Wyble 2010). Naturally, we wish to take advantage of as much prior information available to us as possible, but given this variation, we wish to do so without negatively affecting our results. Our approach previously was to set constraints on orderings based on the results from prior literature. While this is likely to be, in practice, quite a good way of setting prior expectations, the strict nature of walling off an entire set of potential orderings based on the average behaviour of previous work is likely to be an overly strong assertion. This is particularly so for our attentional blink, which we have noted exhibits variability at the subject level.

Clearly, it is desirable to seek a way to improve this situation. One approach to this is to use our own data to improve an existing set of constraints. However,

such an approach comes with potential problems. Since our data is also used to derive our contrast of interest, we must be careful that our method of improving constraints is independent from this contrast of interest and therefore does not bias the end results. This is a significant constraint. Furthermore, while we wish to benefit from having constraints that more closely match our data, we do not want to "overfit" these constraints. A set of constraints that perfectly reflects the pattern of behaviour in our data is likely to not only capture the signal in the data, but the noise as well. Whatever method we select must somehow constrain our prior selection in such a way as to preclude this problem. In light of these requirements, we propose the following approach.

The Proposed Method

We propose a method that starts with a (relatively) strict set of constraints derived from prior literature, and removes constraints (and therefore increases the space of valid orderings) based on an independent measure of validity on the data. In order to counter overfitting and the creation of models that are theoretically unsubstantiated but just happen to fit this set of data well, a certain set of constraints can be held to be "irrevocable", and as such they will never be changed regardless of the data – any violation of these can only be considered a measurement error.

Our independent measure of choice is the "trace and dimension true" measure we discuss in our literature review, $BF_{(D\&T)/N(D\&T)}$ which is equivalent in our case (since we lack a trace factor) to the $BF_{D/N(D)}$ factor used in the previous chapter as a measure of constraint validity. This measure quantifies the evidence that the data is consistent with the trace and dimension factors versus the evidence that it is not. It is a suitable measure because our contrast of interest is entirely contained inside the trace and dimension true model. Specifically, $BF_{D\&T/N(D\&T)}$ is an independent measure from the $BF_{M/NM|(D\&T)}$, since $M|(D\&T) \cup NM|(D\&T) \subseteq D\&T$, that is, the union of the monotonic and non-monotonic orderings that are permissible given some set of constraints is contained inside the set of all possible orderings that are permissible given those constraints. This implies that the changes in the balance of probabilities of the monotonic and non-monotonic orderings that are permissible given some constraints (calculated as $BF_{M/NM|(D\&T)}$) has no effect on the balance of probabilities of all orderings that

are permissible given some constraints versus all those that are not (calculated as $BF_{D\&T/N(D\&T)}$). $BF_{D\&T/N(D\&T)}$ is therefore suitable as a measure by which to select our constraints, because not only is it measuring the validity of our constraints (as per our literature review), but we can be certain that it is independent of the balance of probabilities $BF_{M/NM|(D\&T)}$.

Our method is then as follows. We first pick a set of order constraints on the state and dimension axes from prior data, $C = \{c_1, \ldots, c_n\}$. This set of constraints should be the fullest set that can be reasonably expected to fit the data, but should not contain constraints that contradict one another. We then divide this set C into two subsets, those constraints in C for which violation can only constitute a measurement error (the *irrevocable* set), and those about which we might expect variation between experiments (the *free* set). We label these $E = \{e_1, \ldots, e_l\}$ and $F = \{f_1, \ldots, f_q\}$ respectively. Next, we introduce the concept of group validity for a given set of constraints, denoted GE. This is the product of $BF_{D\&T/N(D\&T)}$ across all our M participants for the set of constraints C, specifically

$$GE(C) = \prod_{i=1}^{M} BF_{D\&T/N(D\&T)_i}$$

For each item in F, we denote the "leave one out" subset of constraints $\overline{F_j}$ as:

$$\overline{F_j} = E \cup \{f_1, \ldots, f_{j-1}, f_{j+1}, \ldots, f_q\}$$

We then calculate $GE(\overline{F_j})$ for all $j = 1..q$. For the largest evaluated $GE(\overline{F_j})$ with $GE(\overline{F_j}) > GE(E \cup F)$, we then remove f_j from F. This procedure is repeated on the new F with f_j removed until there does not exist a set such that $GE(\overline{F_j}) > GE(E \cup F)$, or until $F = \{\}$. The resulting $E \cup F$ is the "empirical prior".

Our method is justified as follows. Firstly, it is clear that setting our empirical prior based on $BF_{D\&T/N(D\&T)}$ will, on its own, converge to a prior set of constraints that best fit the data. Secondly, since we are starting from the fullest (strictest) set of constraints that are theoretically grounded and pruning from this set, it is impossible for us to introduce spurious constraints that fit the data by chance, but are incompatible with our theoretical understanding. Equally, because we

hold some constraints “irrevocable", we are protected from removing constraints that are highly likely a-priori, based on measurement errors. Finally, our condition of independence is fulfilled since $BF_{D\&T/N(D\&T)}$ and therefore grouped evidence GE is independent from our measure of interest.

Validation

To validate the method, we apply the empirical method to simulated data, with a known ground truth. Random data is simulated for two separate cases, either such that it is strongly monotonic or such that it is strongly non-monotonic. For each set of data, a random set of constraints is generated, with the property that it is at least partially inconsistent with the data. The constraints are chosen to be inconsistent with the data in order to provide the empirical method with the fair opportunity to demonstrate its improvement of them. For the monotonic case, the initial set of constraints is non-monotonic, and visa-versa. We then evaluate $BF_{M/NM}$ and $BF_{D/N(D)}$ for both this original set of constraints, and the constraints after our empirical priors method has been applied to them. If our method of empirical priors is functioning correctly, it should push $BF_{M/NM}$ toward more extremely monotonic or non-monotonic patterns that match the underlying pattern of behaviour in the data. Simultaneously, it should also increase $BF_{D/N(D)}$, since $BF_{D/N(D)}$ is the metric by which we improve constraints. We repeat this process 1000 times for each of the monotonic and non-monotonic cases, and plot the improvement in $log_{10}\ BF_{M/NM}$ from the original set of constraints to the empirically derived set. Results can be seen in Figure 26 and Figure 27.

We note that despite the strong overall improvement that the empirical method demonstrates, there are counterexamples that we will call *reversals* in which the empirical method makes things *worse* instead of better. We analyse this finding more in the discussion of this chapter, but broadly we believe it is natural to expect a small number of reversals arising as a limitation of our simulation methodology. For now, we note that these reversals occur when the Bayes Factor of the initial set of randomly generated constraints provides strong evidence for the hypothesis in question already. In the non-monotonic case, this can be seen as the average $\log_{10}$ evidence over 1000 samples for the initial set of constraints for reversals is -8.31, versus -3.73 for non-reversals. In the monotonic case, we see a similar pattern of behaviour with the average $\log_{10}$ evidence for the initial

set of constraints for reversals is 7.67, versus 0.01 for non-reversals. We also note that this effect significantly decreases as we force our simulated data to less strongly reflect out hypothesis. Allowing one order of magnitude more evidence in the generation of our initial set of monotonic or non-monotonic data (and thereby increasing the potential available evidence for the other hypothesis) reduces these values to -3.46 and -1.22 in the non-monotonic case and 3.11 and 0.25 in the monotonic case.

Results

Validation

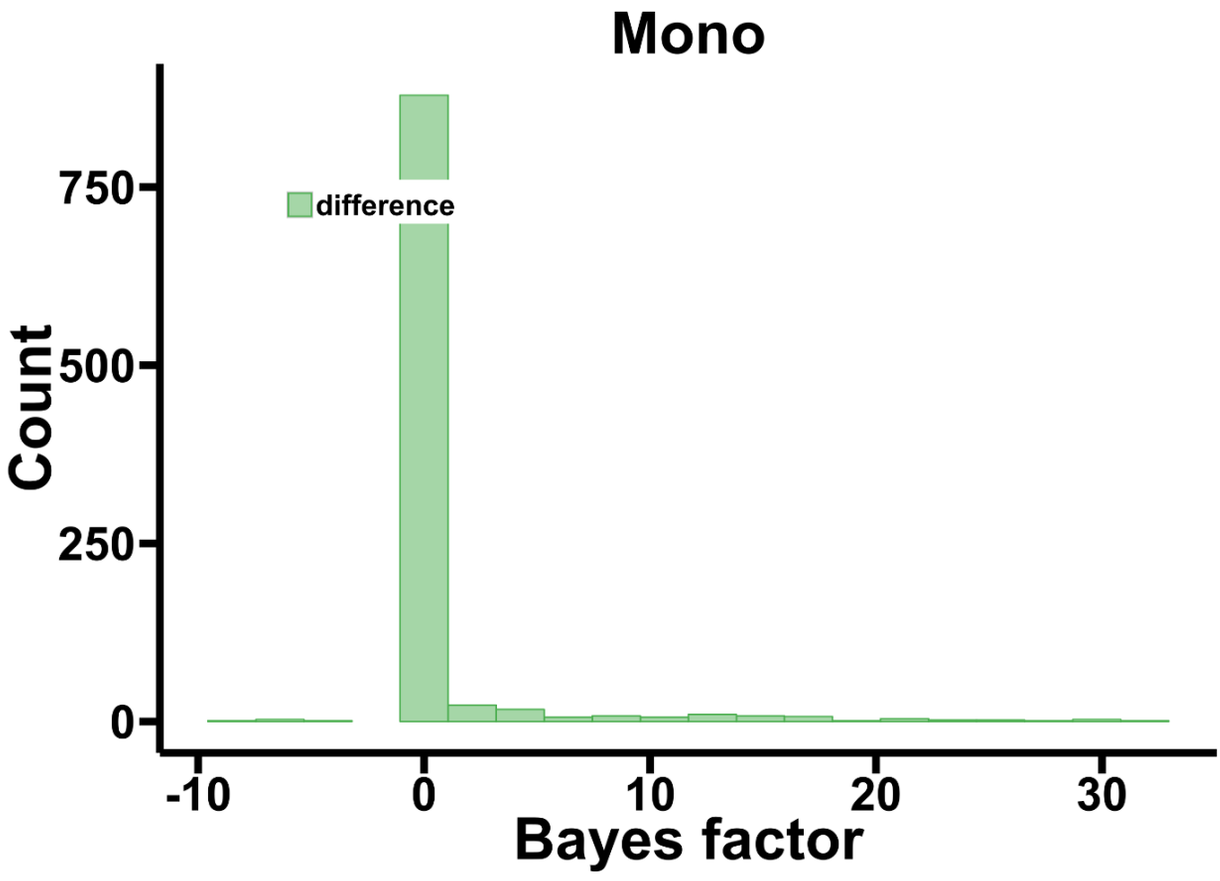


Figure 26) Difference in Bayes factors ($BF_{M/NM}$) for randomly selected constraints before and after the empirical constraints method has been applied for monotonic data.

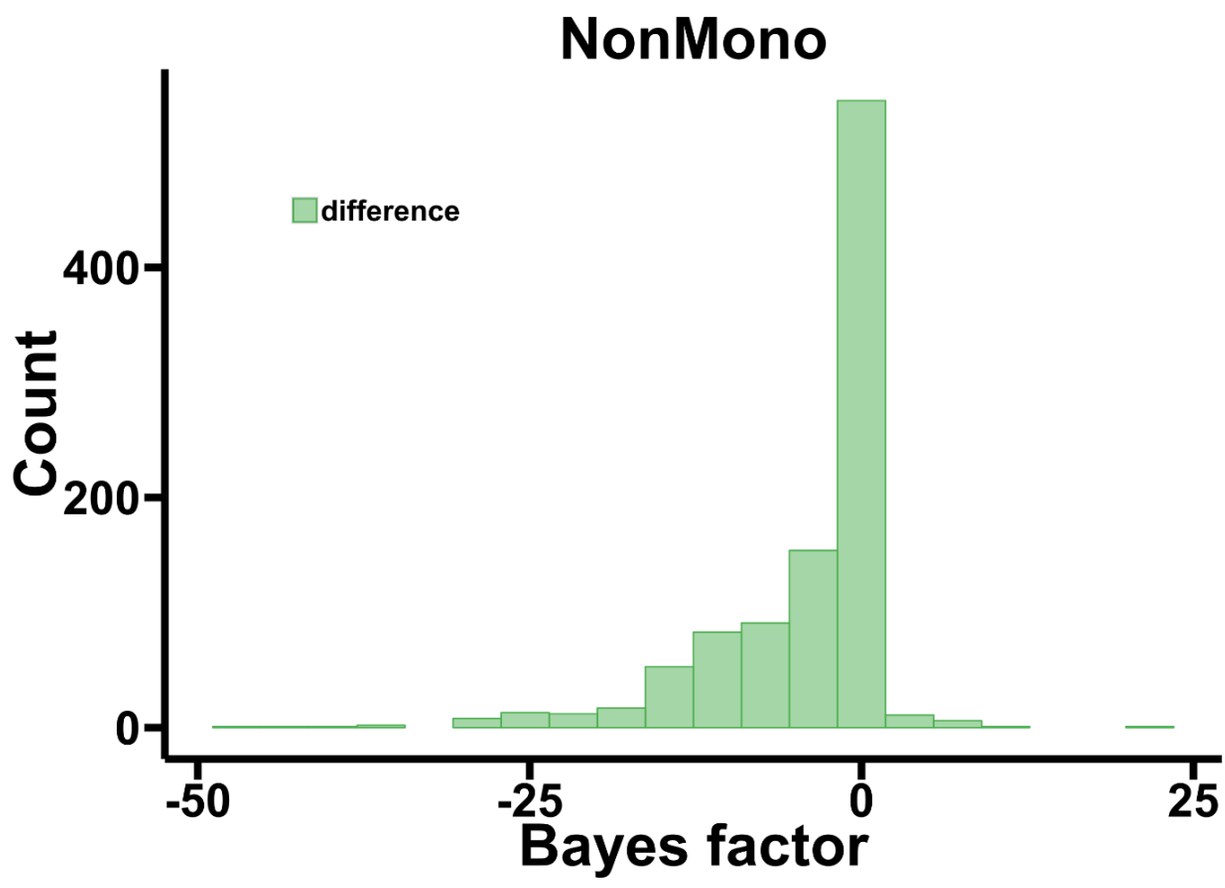


Figure 27) Difference in Bayes factors ($BF_{M/NM}$) for randomly selected constraints before and after the empirical constraints method has been applied for non-monotonic data.

Prior

We apply our method to the constraints used from the previous chapter. As previously, Lag 8 is excluded as a serial position outlier. Previously, our constraints were: for report accuracy, Lags 1, 6 and 4 would be held to be larger than Lags 2 and 3, with Lag 1 additionally being held to also be larger than Lag 4. For subjective report, Lag 6 would be held to be higher than Lag 4, lag 4 higher than Lag 3, and Lag 3 higher than Lag 2. We held the following constraints irrevocable for report accuracy: Lag 1 > Lag 2, Lag 1 > Lag 3, Lag 1 > Lag 4, and none for subjective visibility. This resulted in the removal of the constraints that Lag 4 > Lag 3 and Lag 4 > Lag 2. The final constraints after applying the empirical priors method were therefore: For report accuracy, Lags 1 and 6 would be held to be larger than Lags 2 and 3, and Lag 1 additionally would be held to be larger than Lag 4. The constraints for subjective report remained unchanged. Interestingly, we note that, had we not held the constraints concerning lag 1 irrevocable, they would have been pruned. This would have left us with the constraints for

accuracy as only Lag 6 being held larger than Lags 2 and 3, while the constraints for visibility would have remained unchanged.

State-Trace analysis

We re-ran our previous analysis using the new set of constraints. When data points have been removed and constraints are no longer applicable, the same process has been applied as in the previous chapter – relevant constraints have been removed.

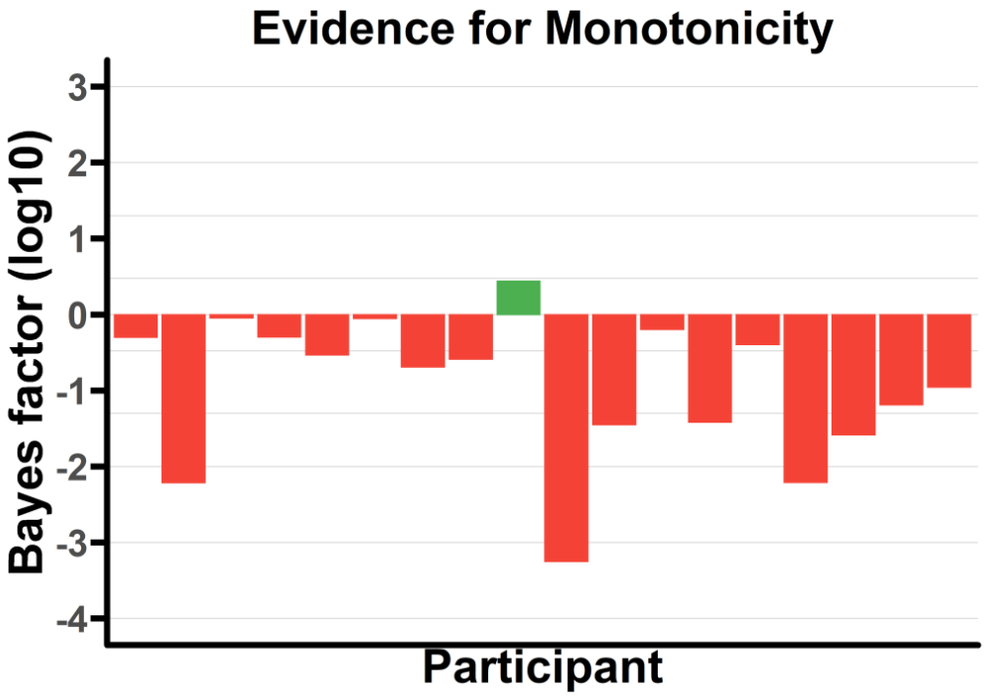


Figure 28) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) for empirical priors across participants. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to Bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

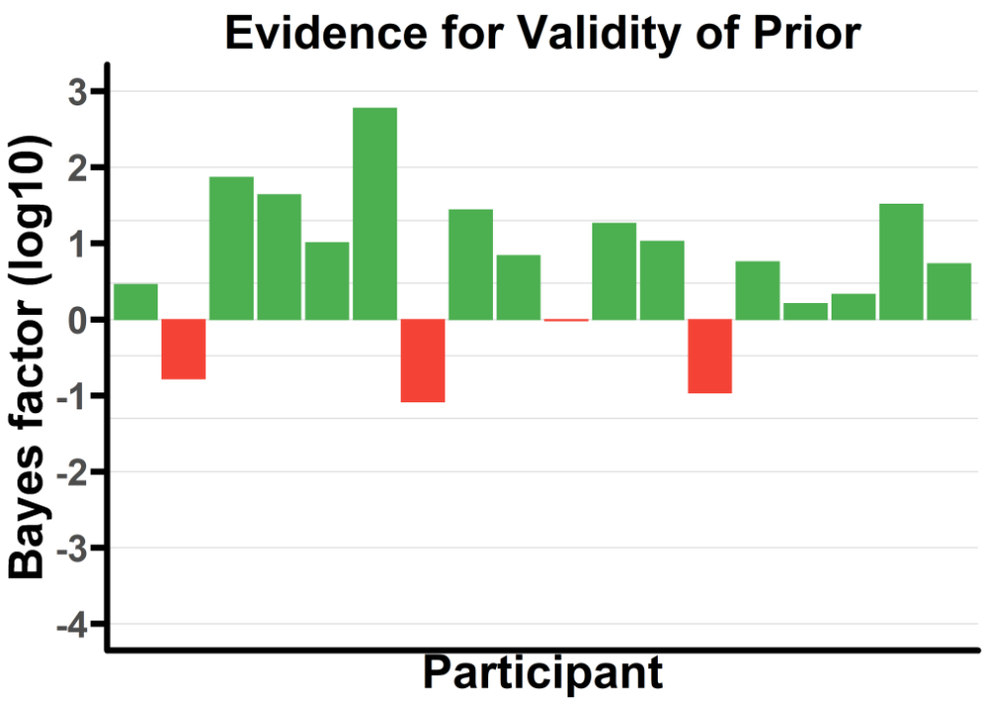


Figure 29) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) for empirical priors across participants. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to Bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Figure 29 shows validity for each participant for the empirically derived set of prior constraints. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 1.07 \times 10^{13}$. However, we note that while the group validity is strong, four participants show the opposite pattern. Figure 28 shows the respective non-monotonicity for this set of constraints. Results are strongly and almost homogenously in favour of the non-monotonic model, with grouped (*not* log) $BF_{M/NM|(D)} = 1.17 \times 10^{-17}$.

Post Hoc Testing

As in the previous chapter, in order to establish the effect of each lag on the final calculation of monotonicity, we reran the state-trace analysis with each of the lags excluded in turn.

Lag 1

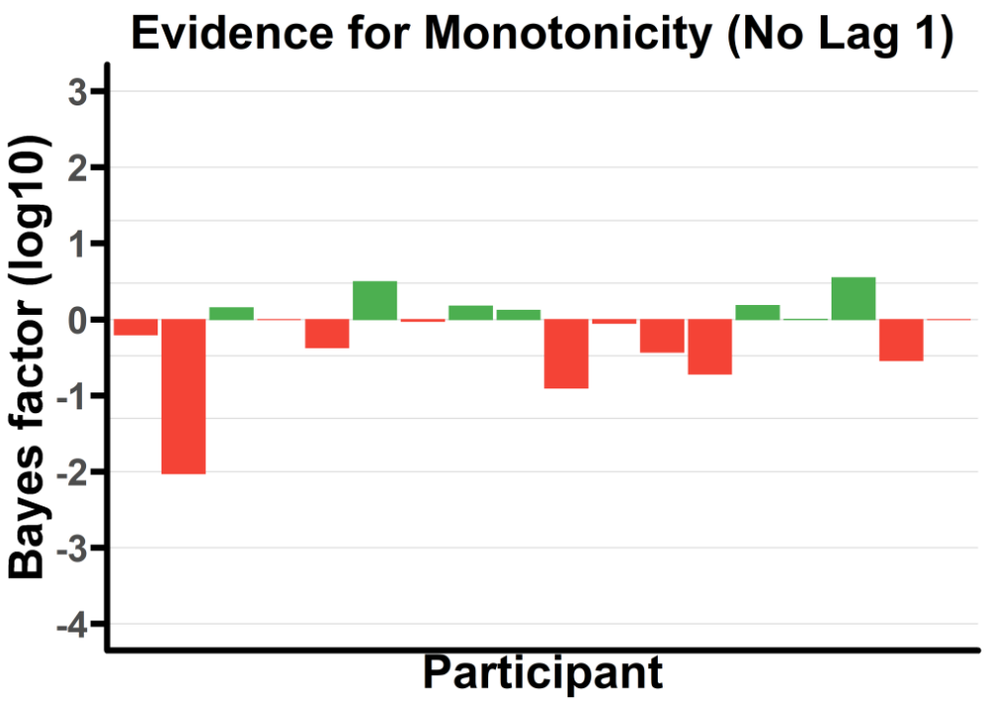


Figure 30) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) for empirical priors across participants with lag 1 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

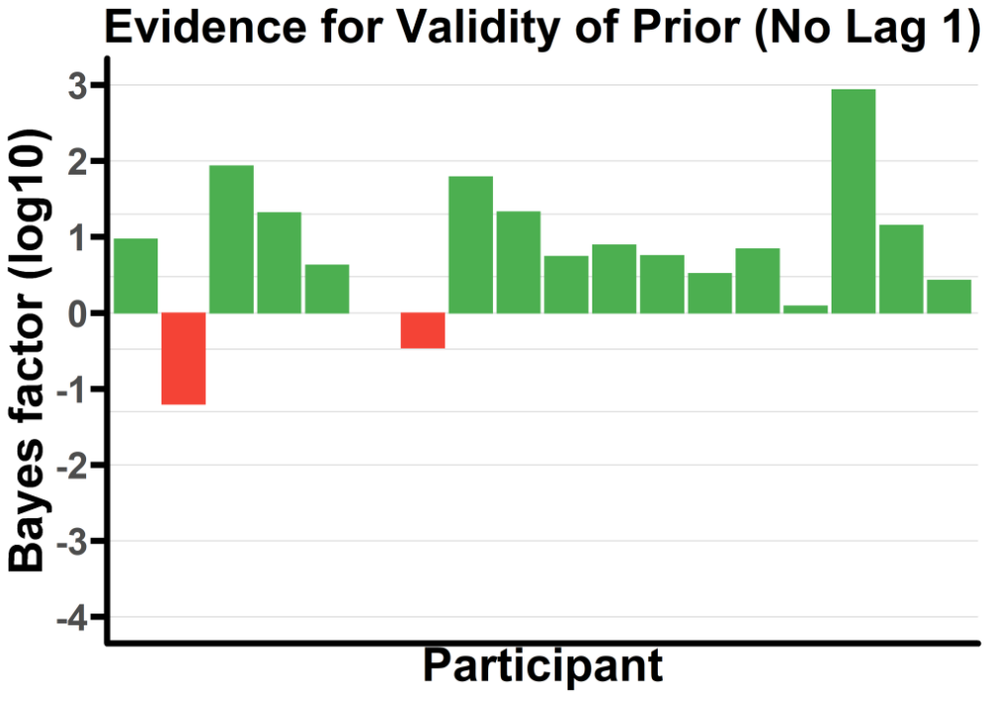


Figure 31) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) for empirical priors across participants with lag 1 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Lag 2

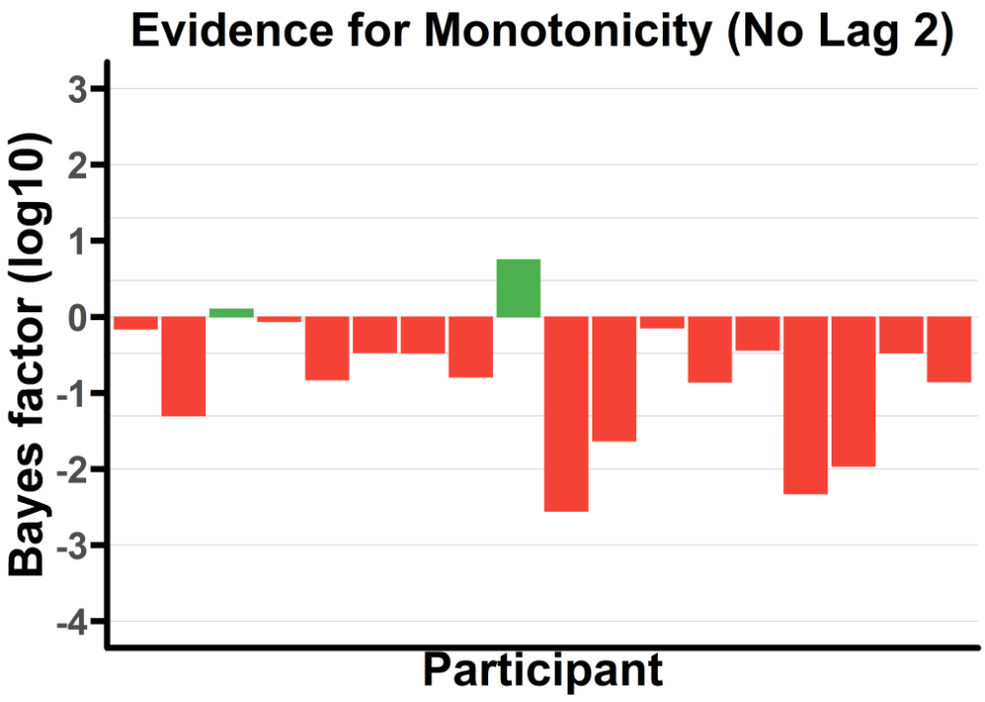


Figure 32) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) for empirical priors across participants with lag 2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

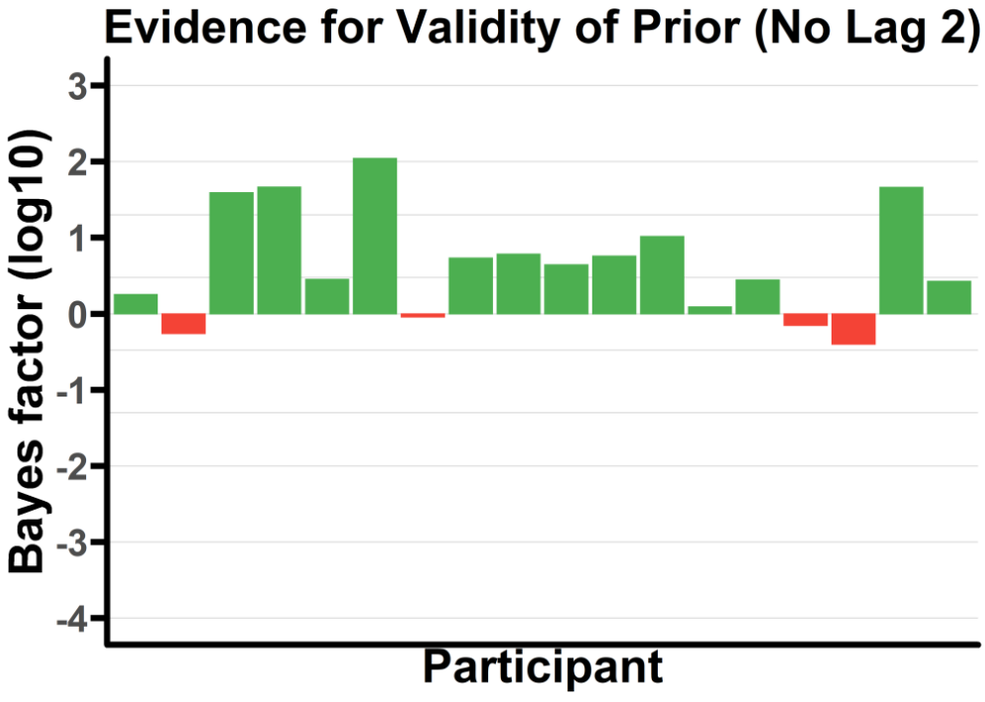


Figure 33) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) for empirical priors across participants with lag 2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to Bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Lag 3

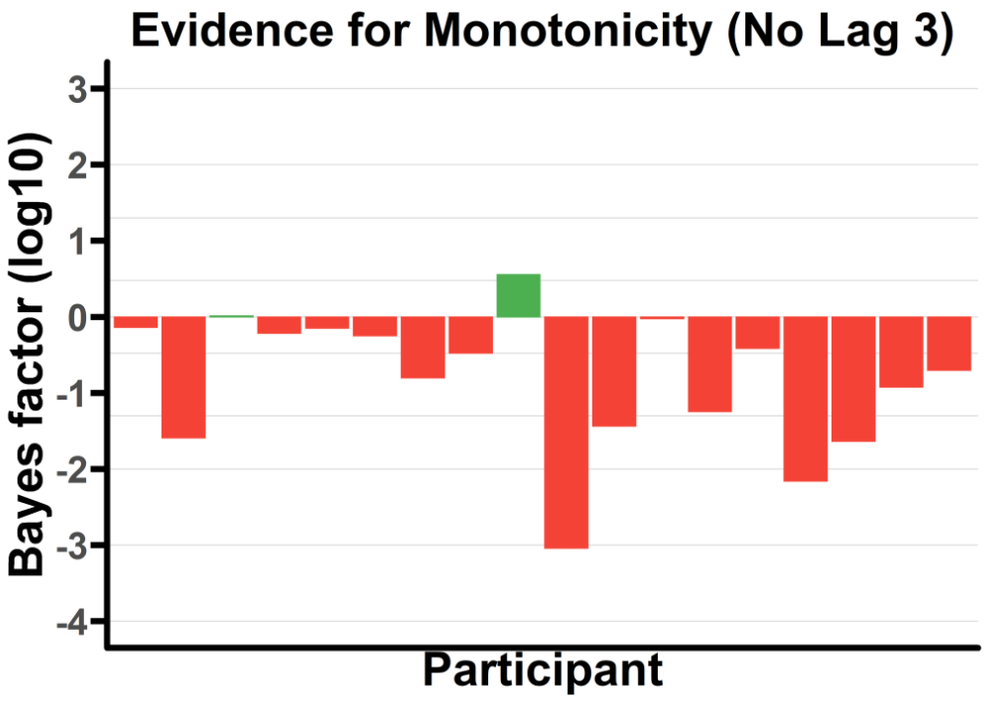


Figure 34) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) for empirical priors across participants with lag 2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to Bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

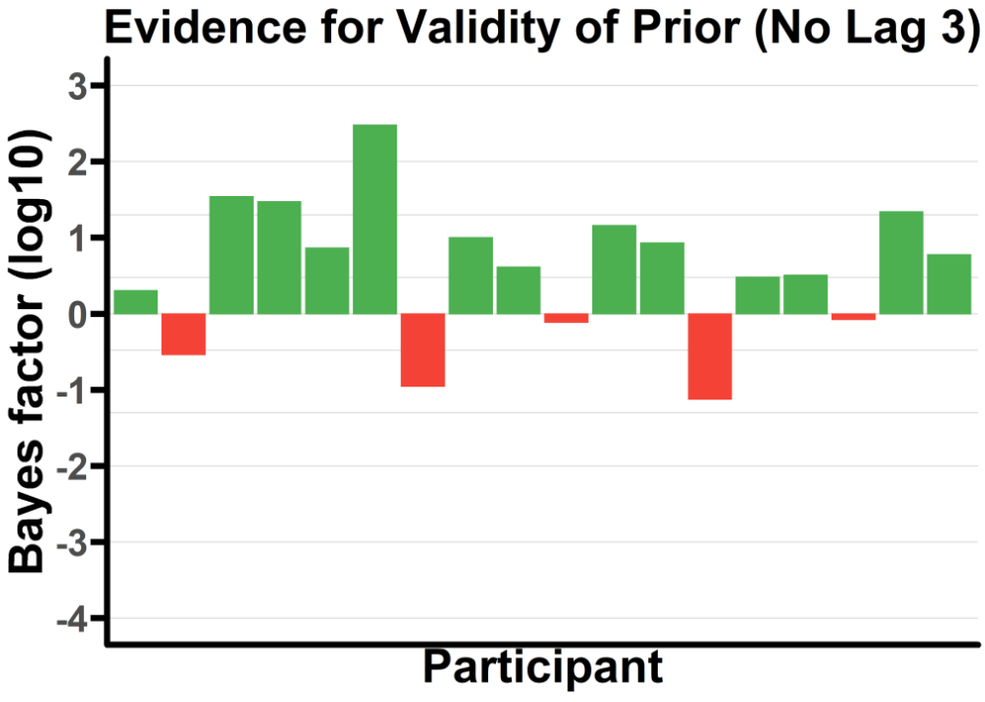


Figure 35) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) for empirical priors across participants with lag 2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to Bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Lags 1&2

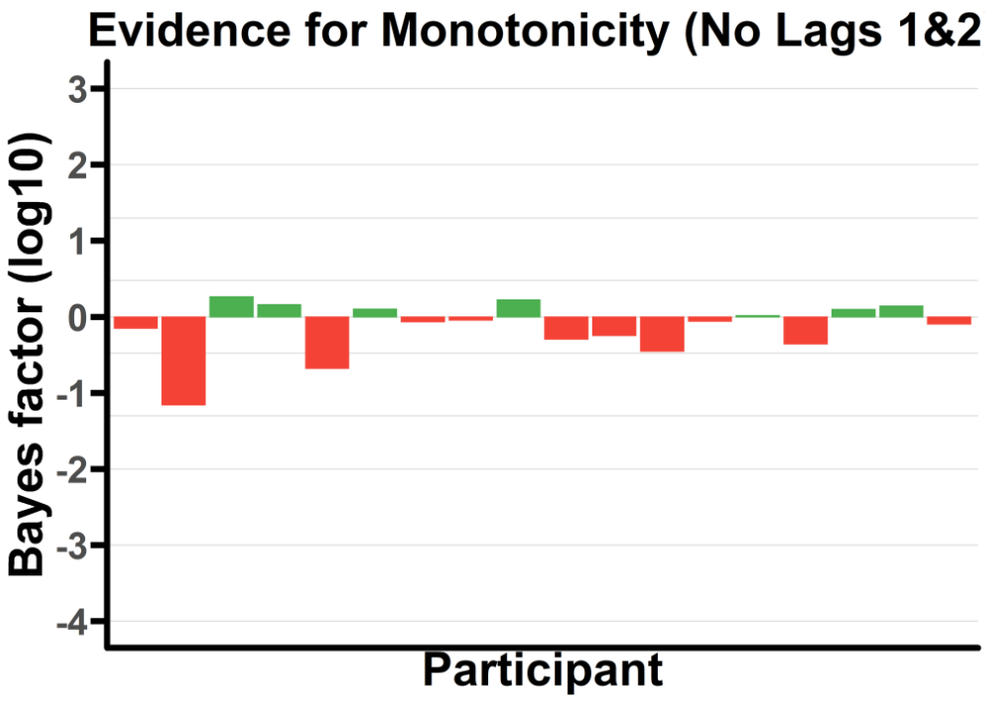


Figure 36) Log_{10} Bayes factors for monotonicity (positive, green) versus non-monotonicity (negative, red) for empirical priors across participants with lags 1&2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

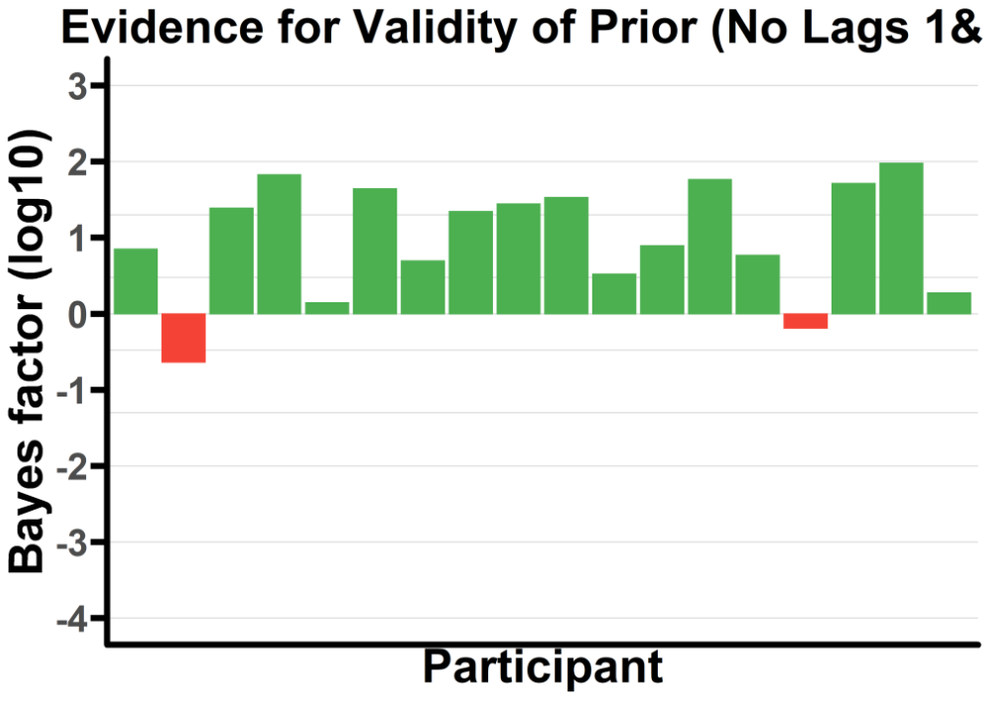


Figure 37) Log_{10} Bayes factors for positive evidence for the constraints (positive, green) versus negative evidence for constraints (negative, red) for empirical priors across participants with lags 1&2 excluded. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20 and 100 respectively.

Figure 31 shows validity for each participant for the empirically derived set of prior constraints. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 3.62 \times 10^{17}$. Group validity is stronger than previously, but with more participants showing no evidence either way. Figure 30 shows the respective non-monotonicity for this set of constraints. Results show no strong preference for monotonicity or non-monotonicity and are almost completely heterogeneous, with grouped (*not* log) $BF_{M/NM|(D)} = 2.58 \times 10^{-4}$.

Figure 33 shows validity for the empirically derived set of prior constraints. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 5.21 \times 10^{11}$. Figure 32 shows the respective non-monotonicity for this set of constraints. Results show a strong preference for non-monotonicity and are almost completely homogenous, with grouped (*not* log) $BF_{M/NM|(D)} = 3.37 \times 10^{-15}$.

Figure 35 shows validity for each participant for the empirically derived set of prior constraints. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 4.77 \times 10^{10}$. Figure 34 shows the respective non-monotonicity for this set of constraints. Results show a strong preference for non-monotonicity and are almost completely homogenous, with grouped (*not* log) $BF_{M/NM|(D)} = 2.37 \times 10^{-15}$.Lags 1&2

Figure 37 shows validity for each participant for the empirically derived set of prior constraints. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (*not* log) $BF_{D/N(D)} = 9.1 \times 10^{17}$. Figure 36 shows the respective non-monotonicity for this set of constraints. Results show a strong preference for non-monotonicity and are almost completely homogenous, with grouped (*not* log) $BF_{M/NM|(D)} = 2.67 \times 10^{-3}$.

Discussion

Validity of Empirical Prior

The validation of our empirical priors' method shows exactly the behaviour we predicted. When the data is strongly monotonic, the estimate of $BF_{M/NM}$ is pushed towards a more monotonic result (Figure 26), and when the estimate of $BF_{M/NM}$ is strongly non-monotonic, $BF_{M/NM}$ is pushed toward a more non-monotonic result (Figure 27). It is interesting to note however that this improvement does not seem to be guaranteed, there are clearly a number of instances in which the empirical priors method pushes $BF_{M/NM}$ a small amount in the opposite direction. That is, in the monotonic case, the empirical constraints are less monotonic than the original set, and vice-versa for the non-monotonic set.

However, we note that it is to be expected that a small number of cases demonstrate this *reversal* effect. In order to be able to calculate $BF_{M/NM}$ constructively, we require that the evidences for both outcomes (Monotonic and Non monotonic) are non-zero. In other words, the highly monotonic datasets must allow some small evidence for non-monotonicity in order for us to demonstrate that the balance of this evidence is improved by our method (and vice versa for non-monotonicity). However, in allowing non-zero evidence for both hypothesis there exists a small but *not* infinitesimal chance that the analysis subject to the randomly chosen constraints constitute *more* evidence than is warranted by the data on its own. In these instances, the removal of the constraints will reduce the evidence down to a level that more accurately reflects the underlying data. This can be verified by the two observations we make at the end of our validation section. Firstly, all instances in which these reversals occur are when the belief in the randomly generated constraints is already overwhelmingly strong. Secondly, the rate of these reversals can be increased by increasing the average evidence for the "alternative" hypothesis: increasing the average strength of the evidence for non-monotonicity when the data is generated as monotonic increases the rate of reversals (and vice versa if the data is generated as non-monotonic).

Monotonicity versus Non-Monotonicity

The results from this chapter further strengthen the analysis performed in the previous chapter. Evidence is qualitatively similar, and evidence for non-monotonicity has increased almost uniformly by 2-3 orders of magnitude across all conditions we tested. The presence of outliers has significantly decreased too: previously 3 subjects showed significant (>3) evidence for monotonicity in the full dataset, but with the empirical set of constraints this is reduced to none. Given that the validity of our priors $BF_{D/N(D)}$ is the measure that guides the creation of our empirical prior, it is no surprise that we see it increase after the application of our method. However, in this instance it is compelling that the evidence has improved so significantly. For this measure too, we find the presence of outliers decreased.

Lag 1

One outlier in terms of behaviour is the lag 1 data point. Both in the original set of constraints and the empirical set, the removal of the lag 1 data point increases the validity of the constraints. Furthermore, we note that had we not held several of our constraints involving lag 1 as irrevocable, that these constraints would have been removed – contrary to what we know about the attentional blink. In retrospect however, this behaviour is exactly why we specify certain constraints as irrevocable in the first place. In the attention blink, it is known that some participants are “non-blinkers” who simply do not experience any kind of “blink of the mind's eye" as the two targets approach one another in time (Martens, Wyble 2010). Indeed, we note in the data that this increase in constraints validity is the result not of a systematic change in evidence, but the sharp reversal of several participants changing from weak evidence against the constraints to strong evidence for the constraints (See Figure 29 and Figure 31). Furthermore, the lag 1 data point is by far the most constrained data point in our prior. Given the significant participant-by-participant variation in the attentional blink, it is not surprising to see an improvement in grouped evidence with its removal simply on account of natural variation. We therefore hold this behaviour up not as evidence that our empirical method is flawed, but that it is functioning exactly as intended. In this instance, despite an increase in grouped evidence, it would have been

premature to consider removing constraints on the lag 1 data point despite the pattern of behaviour in the data; we would have very likely been overfitting.

Hierarchical modelling

Another way of looking at our method is that it is an approximation of a hierarchical Bayesian model, in a manner very similar to the methods of empirical priors. Interestingly, this is not the first time that hierarchical methods (or approximations thereof) have been proposed for this type of analysis, but previous works have focused on them as achieving a better measure of the group level effect (Morey, Pratte et al. 2008, Pratte, Rouder et al. 2010, Rouder, Lu 2005) that is not subject to the averaging problem. In contrast, our method uses (a weak approximation of) hierarchical modelling in order to achieve a better fit of the prior to the original analysis. However, we note that true hierarchical models of this kind are difficult to build because of the non-parametric nature of state-trace analysis, and must be careful not to fall into the same averaging trap that normal models suffer from (Davis-Stober, Morey et al. 2016). Furthermore, without making use of Laplace's method to easily calculate posteriors (Davis-Stober, Morey et al. 2016), Bayesian analysis of state-trace problems require the use of Gibbs sampling or other computationally expensive methods to calculate the posterior (Prince, Brown et al. 2012, Davis-Stober, Morey et al. 2016). Using hierarchical models, there is a strong possibility that we will be required to use these computationally expensive methods to calculate the posterior once again.

Interestingly, though slightly different in their conception and application, both the existing proposals for hierarchical models and our new method are very similar. They are both effectively attempting to address the same problem – that the averaged behaviour across all subjects is not necessarily a good approximation of the individual level effect. This proposes some interesting directions for further research. One technique that may be well suited for assisting the derivation of a prior for the state-trace analysis for example is to borrow the Aggregated Bayes Factor method discussed in the literature review. This is a measure that can be used to confirm the heterogeneity of the data, and may be a useful basis on which to exclude a prior. Conversely, many of the difficulties of true hierarchical models for the group level effects might be subverted by using an empirical approximation instead. Overall, our empirical

method is computationally cheap and, as we have demonstrated, behaves well. This allows much of the benefit and flexibility of a hierarchical Bayesian analysis, without the computational drawbacks.

Conclusion

In our previous chapter, we used a Bayesian state-trace analysis in order to demonstrate evidence for a dissociation between working memory encoding and subjective experience. In this chapter, we re-examined the method we used in the previous chapter, and proposed some improvements to create a fairer test on our data, as well as validating these improvements empirically. We then applied this new method to the analysis in our previous chapter, and found that the results support the conclusions from the last chapter even more strongly.

6. Modelling Subjective Experience

Abstract

In previous chapters, we have taken advantage of state-trace analysis to provide evidence that working memory encoding and subjective experience may be dissociated in the attentional blink. However, while dissociations can tell us about specific effects, placing findings in larger theoretical context is pivotal to the forward progress of science, especially when the theory is encapsulated in a computational model. In this chapter, we therefore explore one particular interpretation of the data from (Pincham, Bowman et al. 2016) through computational modelling, that items are encoded into working memory simultaneously, but are only experienced in series. In combination with our state-trace analysis, this allows us to explore not only the existence of the effect, but some plausible mechanisms by which it may arise.

Our computational model of choice for implementing this hypothesis is the Simultaneous Type/Serial Token (STST) model of attention: it models data in the relevant context (the attentional blink), and naturally deals with the difference between simultaneity and seriality. However, currently the model has no mechanism by which to index subjective report, and one of the contributions of this chapter is to provide such a mechanism. In order to validate the model, we compare the behavioural predictions of the model and the virtual ERPs it generates to human data.

Introduction

In the previous chapters, we examined the relationship between working memory encoding and subjective experience during the attentional blink and found evidence that a dissociation existed between the two. However, as we have previously discussed, dissociations can only tell us so much. Even with our post hoc test, we have only demonstrated that a dissociation exists, and which lags contribute to the effect most strongly. Given the separating between working memory encoding and subjective experience seems to be largest at lag 1 when subjective report is low and report accuracy is high, we posited working memory encoding as a necessary but not sufficient condition for subjective experience. We have so far only touched upon plausible mechanisms for this result however,

hypothesising in our discussion that this could be the result of working memory encoding of stimuli occurring simultaneously, while the subjective experience of the same stimuli occurs in serial. While this does not take away from our existing findings, to fully develop this research, it is important to be able to place it in a broader theoretical context.

In this chapter, we test this hypothesis through modelling, by embodying our hypothesis in a computational model from which we can make quantifiable and falsifiable predictions, which can then be compared to our human data. To align with our goal of placing our findings in a broader theoretical context, we must consider the merits of different ways by which our model might be created. One way to do this is to integrate with an existing model. Indeed, our model may prove more compelling if it arises not out of a fresh model specifically designed to accommodate our hypothesis, but out of an existing model, or as the result of a small set of changes to one.

To be specific, our hypothesis in this case is that working memory encoding of targets occurs simultaneously and continues to function typically during the blink, but that experience of the targets occurs in serial, resulting in a decreased percept of the second target as the temporal proximity of the two targets increases. In terms of incorporating with existing models, we have previously discussed many different potential models of the attentional blink, but one model that is mature and sophisticated enough to enable us to explore this hypothesis is the Simultaneous Type/Serial Token (STST) model of attention. Of all the discussed models, only the global workspace model potentially provides any readout of subjective experience (Dehaene, Sergent et al. 2003), but STST provides a strong possibility of naturally incorporating this in a manner consistent with our hypothesis because it already deals with the difference between simultaneity and seriality. Additionally, the Global Workspace model has some difficulties accounting for some attentional blink findings, notably lag 1 sparing and spreading the sparing (Martens, Wyble 2010). Given the importance of the lag 1 data point in our existing findings, this is potentially a substantial obstacle. Furthermore, the STST model has a robust computational implementation and is capable of simulating not just predicted behavioural results, but virtual ERPs. Since the data we use in the previous chapters also has

both behavioural and EEG results, this allows us two avenues of verifying the correctness of our model. For all of these reasons, we choose the STST model over the previous models listed in the literature review. One choice we must make with respect to the STST model is which implementation to use, the original or the more recent eSTST model. For our initial exploration, we opt for the slightly simpler original model, but we propose that applying the same techniques to the eSTST model may be a fruitful area of future research.

In terms of accommodating our hypothesis in the model, we note that many, and often any, behaviours can be obtained from a model with sufficient modification and parameter adjustments (Roberts, Pashler 2000). In order to make the fairest possible assessment of the hypothesis in question, we therefore limited ourselves in two ways with respect to our modelling. Firstly, we would make no changes to the functionality or structure of the existing model, we would only build on top of it to provide a new "readout" from the model. Secondly, this readout must be simple; ideally arising from one or two principles. On this basis, we are able to generate a model for both attention and subjective report from the STST model.

Serial Experience, Simultaneous Encoding

Our proposition is that the differences in behaviour of subjective experience and working memory encoding during the attentional blink are the result of the working memory encoding of multiple stimuli occurring simultaneously, but their experiences occurring in serial. This would result in a situation in which working memory encoding proceeds as in the classical attentional blink as the two targets approach one another (i.e. lag decreases), while subjective report monotonically worsens because of the proximity of the second target to the first. We give an intuition of how this may work as follows. Suppose we have some idealised correlate of experience, a hypothetical readout from some late stage of the brain that indexes subjective experience. We say that when we receive a signal above some threshold from this readout (the threshold of subjectivity) that a stimulus is being experienced, and in all other cases it is not. We then, as per our hypothesis, assume a seriality of this readout, a second stimulus after the first cannot begin to be experienced until the first stimulus falls back below threshold.

When two targets are presented close together - say at Lag 1 – this results in a typically poor experience of the second target because, despite a reasonably strong readout, most of its readout is subsumed by its proximity to the first target. Conversely, if the two targets are far apart – say at Lag 8 – a second target with the same readout strength as before will be experienced much more strongly because its experience is no longer dominated by the proximity of the first target. An illustration of this contrasting behaviour can be seen in Figure 38. Under such a system, average visibility will increase as the proximity of the two targets decreases, up to an upper limit at which the two targets are sufficiently far apart that the second is not materially affected by its proximity to the first. This is exactly what we see in the human data (Pincham, Bowman et al. 2016).

This explains the average trend of our results, but we need to provide an explanation not only of this averaged behaviour, but in addition to be able to incorporate the significant trial-by-trial variance in subjective report seen in the human data. Over the course of an experiment, participants will make use of a large portion of the range of possible subjective reports regardless of the total averaged behaviour. Fortunately, the same system can also explain the significant variance in subjective visibility rating we see. Over the course of an experiment, participants generally make use of almost the entire visibility scale at some point regardless of report accuracy. Even in the case in which the two targets are closely presented together, a sufficiently elongated activation trace of the second target allows the visibilities of both targets to be the same. An example of this can be seen in Figure 39.

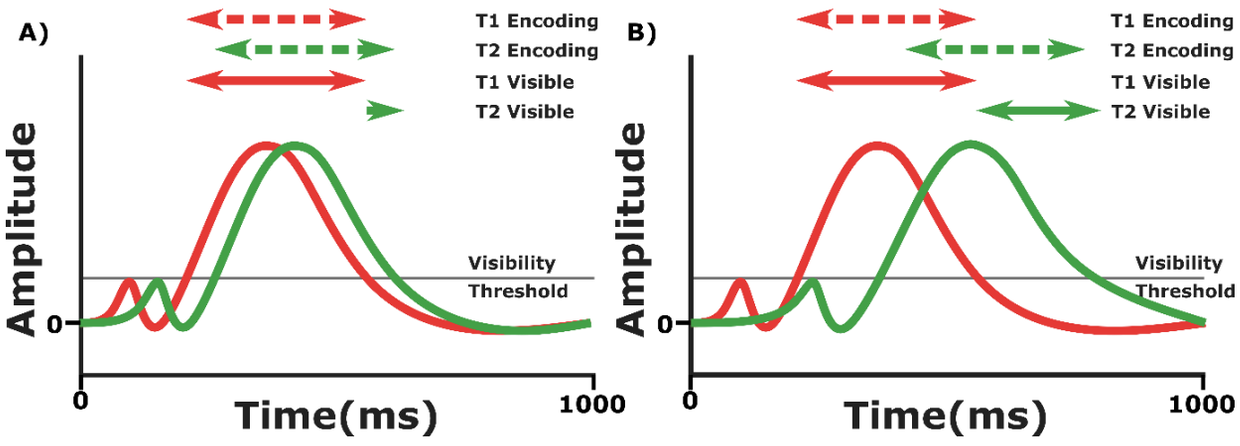


Figure 38) Seriality of experience in the modified STST model. In A), though the P3 amplitude of both stimuli is the same, the duration of the experience of the second stimulus is greatly reduced because it cannot be experienced until the first stimulus falls below the threshold. Comparatively, in B), the P3 amplitude of both stimuli remains the same, but the proximity of the second target to the first is reduced, and consequently they are both experienced for similar durations.

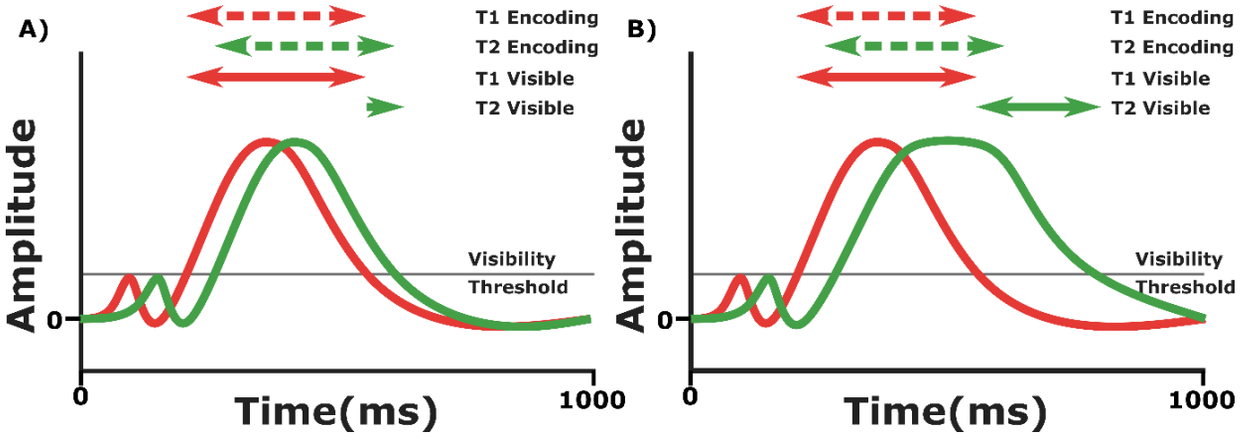


Figure 39) Seriality of experience in the modified STST model. In A), though the P3 amplitude of both stimuli is the same, the duration of the experience of the second stimulus is greatly reduced because it cannot be experienced until the first stimulus falls below the threshold. Comparatively, in B), the P3 amplitude of both stimuli is the same, although the T2's P3 is longer with a slightly delayed onset, consequently they are once again both experienced for similar durations.

These principles must be somehow encoded into the STST model. As we have previously specified, in order to make the fairest possible assessment of the hypothesis in question, this set of changes must also be minimal and must not change the functionality of the existing model. The result of these conditions is the following model to encapsulate serial experience: Subjective visibility is indexed by the strength of the P3 ERP component. This is justified as the P3 component is known to be a strong correlate of subjective experience (Lamy, Salti et al. 2009, Salti, Bar-Haim et al. 2012). Furthermore, the P3 has already

been extracted from the STST model with good results (Craston, Wyble et al. 2009). When an item is above a given amplitude (the threshold of subjectivity), it is being “subjectively experienced" and when it is below, it is not. Additionally, this experience is serial. If the individual P3's for two items are both above the threshold, then the second item cannot be experienced until the P3 for the first one falls below the threshold. The strength of an item's subjective experience is linearly related to the duration for which its P3 exceeds the threshold of subjectivity, subject to no other stimulus already being above the threshold. In this manner, a system allowing a subjective experience that is exclusively serial in manner is created, with only one addition on top of the existing model. In order to evaluate the success of this modified STST model, we will compare its behavioural output to that of human participants and the virtual ERPs it generates to human EEGs in the data from (Pincham, Bowman et al. 2016). We call our model the Simultaneous Encoding/Serial Experience (SESE) model.

Methods

Implementation

Previously, we have described the broad approach to adding subjective report to the STST model, here we detail specifically how the STST model is used to simulate ERPs, the setup of the STST model used to extract a visibility rating, and how the visibility rating was calculated. Our virtual ERPs are calculated from a computational implementation of the STST model, neural-STST. Recall that compared to previous works using virtual ERPs from the STST model, we selected a slightly different stimulus range over which to calculate this virtual P3 in order to provide a thorough exploration of our hypothesis. Specifically, we sample (uniformly) a range of stimulus strengths with greater variability (-0.078 to +0.078 -> -0.1625 to +0.1625), at a slightly higher average stimulus and distractor strength (0.520 -> 0.570). This approach is consistent with previous simulations with the STST model, where we allow input strength ranges to vary reflecting the fact that different experiments being modelled might have quite different stimulus types and sensitivities. For completeness, a full list of change to the original code can be found in Appendix C – Changelog to Neural STST.

In order to calculate subjective report from these virtual P3's, we calculate the number of time steps that a stimulus spends above a given threshold. For the

results given in this paper, this threshold is 0.01. It is necessary to normalise these time step counts into visibility ratings that can be compared to the human data. In the spirit of the simplicity that has driven the creation of the model so far, we normalise the time steps by a linear factor. To keep the range plausible and data driven, the value we selected was the most visible stimulus in the entire experiment, and divided each visibility rating by this in order to give a “percentage visibility". Additionally, although this method gives us a continuous subjective report, for the purposes of comparison with the human data from (Pincham, Bowman et al. 2016), it is necessary to be able to divide these subjective reports into the discrete cases of high/low visibility. Unfortunately, we are unable to be sure that each lag contains the full range of possible subjective reports, making setting a universal splitting point for high and low visibility across all lags potentially problematic. Fortunately, what we are interested in is only the respective behaviour of higher and lower visibility trials – for these purposes, it does not especially matter where we make this split, as long as this comparison can be robustly made. In order to facilitate this, we therefore split our data into high and low visibility for each lag separately by splitting around the mean average visibility by lag. This does mean that the meaning of high and low visibility bins changes by lag, but since our comparison of interest is a qualitative comparison with the behaviour of the human EEG data, this is acceptable. In this way, we provide a simple index of both continuous and binned subjective report that requires no changes to the original model.

Gaussian noise

The model, as we have described it so far, assumes that subjective experience is deterministically defined by the number of time steps – a given number of time steps always corresponds to the same subjective visibility report. For the purposes of testing our hypothesis about the respective behaviour of high and low visibility ERPs, this is suitable because it provides the best possible contrast of these behaviours. However, the brain is a noisy biological system; consequently, we add noise to the visibility report of the model. Specifically, we set the visibility rating of the model to be additionally dependent on a Gaussian distributed error term. This error is initially set by randomly sampling from a Gaussian distribution with mean of zero time steps and a standard deviation of

30 time steps. Since adding such an error may potentially lead to post-noise visibility ratings that are negative or exceed the unnoisy maximum visibility, we set a floor and ceiling on post-noise values. Any data point that would have its visibility reduced below 0 will be set to 0, and any that would have its post-noise visibility increased above the unnoisy maximum will be set to that maximum. This is a sensible constraint: whatever uncertainty one maintains about their percept, their report will never fall short of an unseen report, or exceed a maximal report.

A significant advantage of adding this noise is that we also no longer need to worry about each of our lags containing the full range of visibility bins – this is almost assured. For this noisy model, we therefore attempt to provide a universal definition of high and low visibility across all lags, allowing the results to be more directly comparable with one another. In order to keep the data driven nature of the point, we once again choose it from the data. Since we must now account for variation in visibility across lags as well as within lags, we instead define it as half of maximal visibility (pre-noise).

Results

In order to validate the model, we compare with data from two human experiments. The first is the behavioural data we have analysed in previous chapters. We wish to see if the pattern of behaviour in the virtual data resembles that of the human data. For this purpose, we will compare the behaviours of raw T1 Accuracy, T2|T1 Accuracy and visibility rating across the two. The second experiment is an EEG dataset, for which we will compare the virtual ERPs generated from the model to ERPs from the human data for high and low visibility results. Human EEG results are taken from the second experiment from (Pincham, Bowman et al. 2016), the follow-up to the behavioural data set we have used in the previous chapters. If our hypothesis is correct, then the pattern of behaviour that is seen in the human data, a reduced amplitude, shortened P3 for low visibility targets versus high visibility targets, should also be seen in the virtual ERPs.

Deterministic

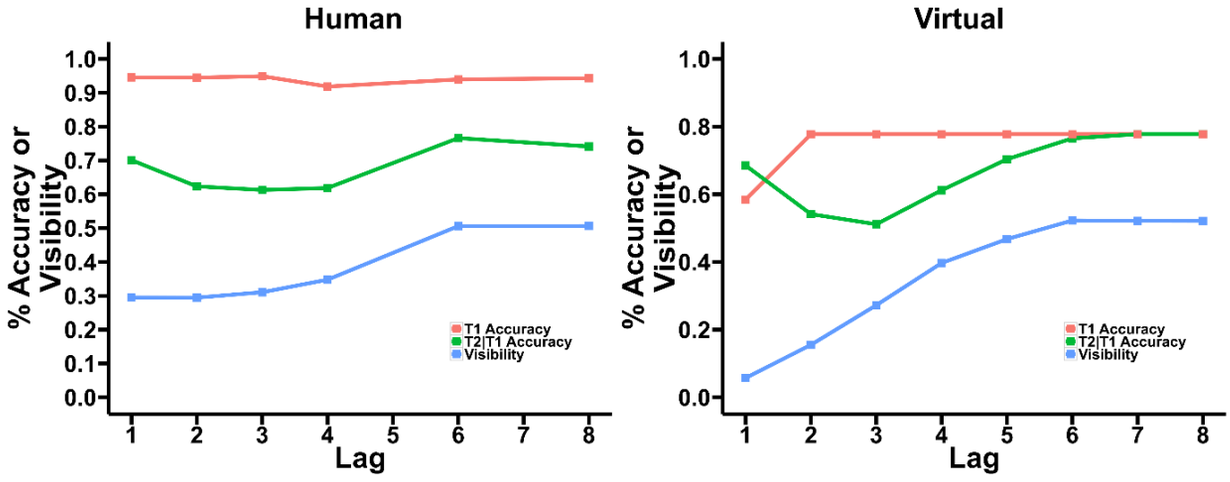


Figure 40) A comparison of T1 Accuracy, T2|T1 Accuracy and T2 subjective visibility from human data (left), and virtual, simulated results (right), for the direct readout. In this case, the length of the P3 is proportional to subjective report. (Note, the T1 Accuracy and T2|T1 accuracy show some very minor differences to that presented in (Pincham, Bowman et al. 2016). This is because T2 accuracy in the original paper was in fact presented as the accuracy of the conjunction of T2 and T1, whereas here we display the conditional probability of T2 given T1. None of our findings are impacted by the difference.

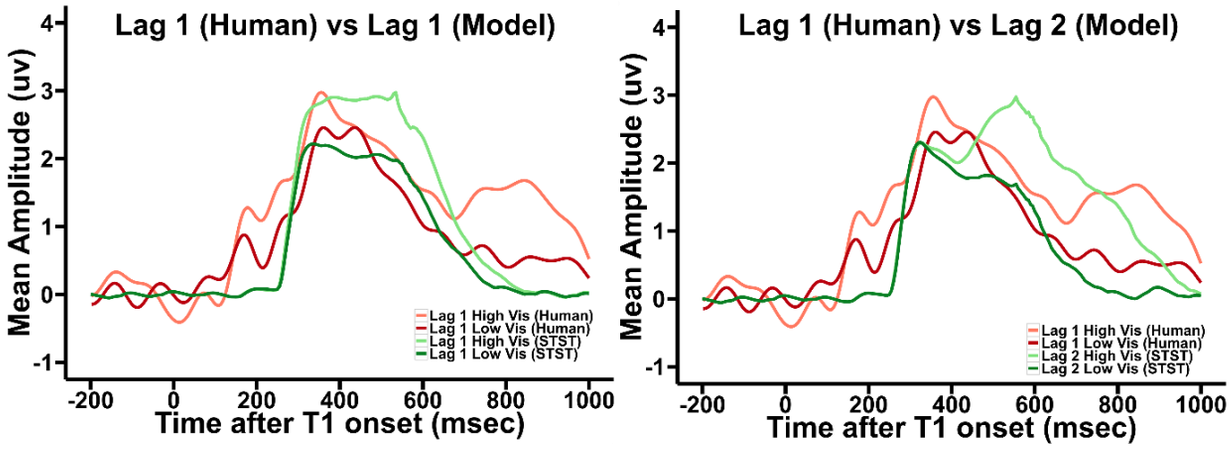


Figure 41) A comparison of human ERPs and virtual ERPs generated from the model for high and low visibility, at Lag 1 for the direct readout model. On the left, we make the comparison of human Lag 1 vs Virtual Lag 1 and on the right we compare human Lag 1 versus Virtual Lag 2. Note that the human data seems to best fit a combination of the virtual Lag 1 and Lag 2.

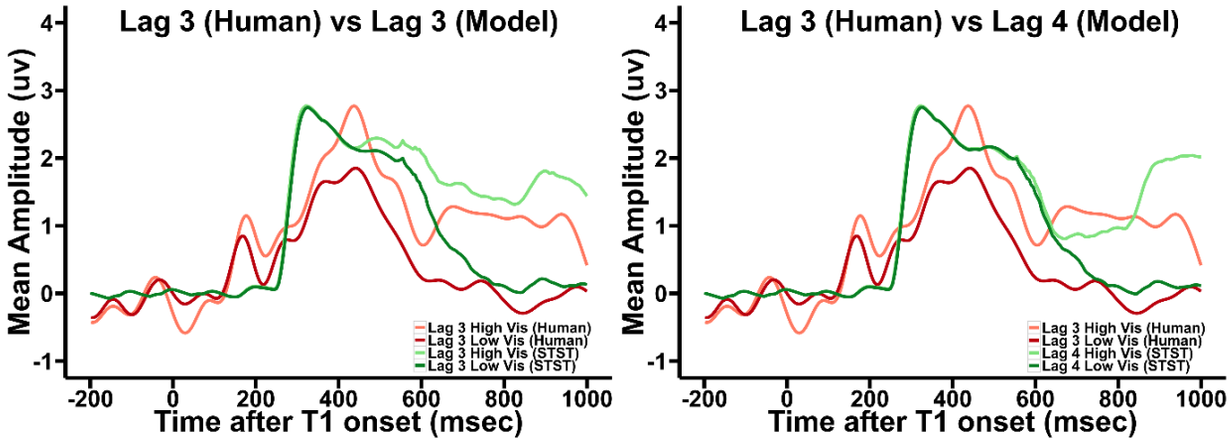


Figure 42) A comparison of human ERPs and virtual ERPs generated from the model for high and low visibility, at Lag 3 for the direct readout model. On the left, we make the comparison of human Lag 3 vs Virtual Lag 3 and on the right we compare human Lag 3 versus Virtual Lag 4.

Stochastic

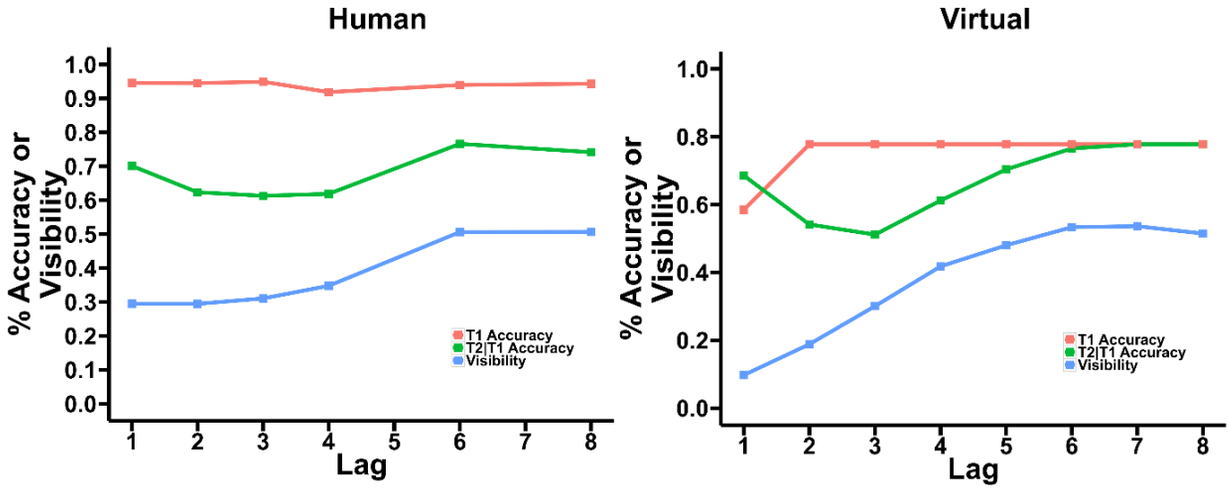


Figure 43) A comparison of T1 Accuracy, T2|T1 Accuracy and T2 subjective visibility from human data (left), and virtual, simulated results (right), for the noisy readout. In this case, the length of the P3 is not perfectly proportional to subjective report, and subject to a noise term. (Note, the T1 Accuracy and T2|T1 accuracy show some very minor differences to that presented in (Pincham, Bowman et al. 2016). This is because T2 accuracy in the original paper was in fact presented as the accuracy of the conjunction of T2 and T1, whereas here we display the conditional probability of T2 given T1. None of our findings are impacted by the difference.

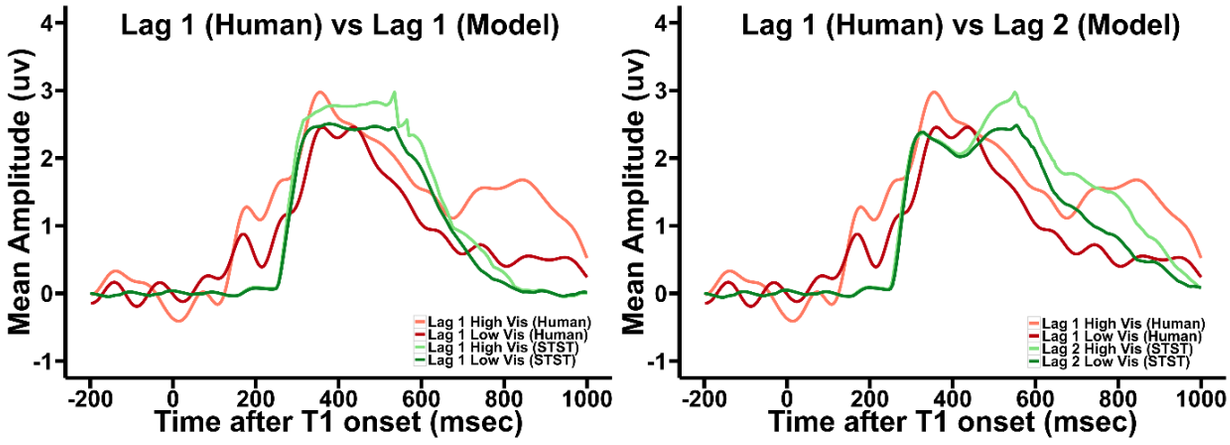


Figure 44) A comparison of human ERPs and virtual ERPs generated from the model for high and low visibility, at Lag 1 for the noisy readout model. On the left, we make the comparison of human Lag 1 vs Virtual Lag 1 and on the right we compare human Lag 1 versus Virtual Lag 2. Note that the human data seems to best fit a combination of the virtual Lag 1 and Lag 2.

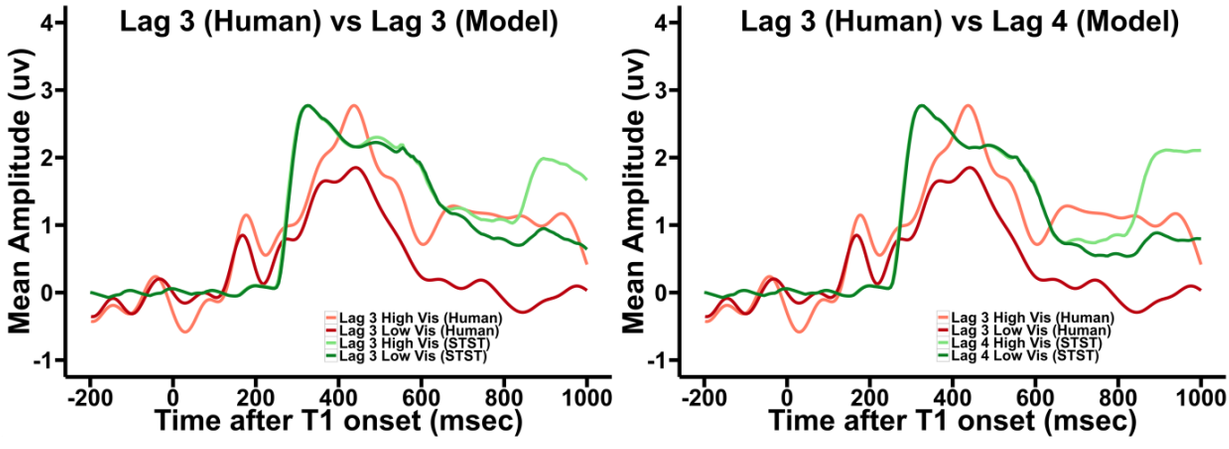


Figure 45) A comparison of human ERPs and virtual ERPs generated from the model for high and low visibility, at Lag 3 for the noisy readout model. On the left, we make the comparison of human Lag 3 vs Virtual Lag 3 and on the right we compare human Lag 3 versus Virtual Lag 4.

Discussion

Behavioural Data

The first comparison we make is between the behavioural results, specifically, we compare the respective report accuracies and subjective visibilities predicted by the read-out enhanced STST model to those from the human data. The results from this can be seen in Figure 40 and Figure 43. Overall, there is a strong similarity between the two. Though the model demonstrates more extreme behaviour than the human data, the qualitative pattern of results is very similar – the ordering of all points is almost identical. One notable difference is that the STST model is simulating a slightly more difficult task than the human data –

report accuracy lower by around 10%. Perhaps because of this, the STST model also demonstrates a more marked downturn in subjective report at earlier lags than the human data. Another difference is the behaviour of T1 accuracy between the two models. In the model data, T1 accuracy decreases below the level of T2|T1 accuracy, while in the human data it stays constant at Laasdasdg 1. For this, we note that the experimental paradigm employed in (Pincham, Bowman et al. 2016) is somewhat different to that modelled by STST in 2007. In particular, the Pincham et al. experiment contains a colour-marked T1, making identification of the first target an easier task than the second. In contrast, the STST model is set up so that identification of T1 and T2 are both equally difficult tasks.

EEG Data

We also compared the virtual ERPs generated by the STST model with the human ERP data for both of the lags sampled in the human data, Lags 1 and 3. For each of these, we present two sets of model ERPs, comparing the human ERPs to the model ERPs at the same lag (left) and the subsequent lag (right). The most significant difference between the two is the respective late dynamics of STST compared to the human data, with the STST ERPs showing differences to the human data from approximately 600ms onward. This occurs more distinctly at lag 1. This appears to be because of a fixed timing offset of the human data versus the EEG data. The human data best reflects a combination of the two model lags – features of both the identical lag and the offset lag are present in each case.

One thing we do note is the (relatively) poor match of the model at lag 1 to the human data at lag 1 compared to other results, which we argue is likely to be occurring as a result of a limitation of the STST model. In the STST model, it is the blaster that provides the strong enhancement that allows items to be bound into a temporal context (i.e. to a token). After this initial enhancement, the blaster is shut off by a strong inhibitory signal to prevent multiple items being bound to the same temporal context. Unfortunately, this shut down also has the effect of creating a hard limit on the length of the P3. In this sense, the STST model provides strong constraints on the upper limit of the length of the P3, and the effect is most noticeable at lag 1 because of the close proximity of the two targets. An interesting future direction to take this research may be to implement

the same measure in the eSTST model, which, due to differing blaster functionality, may not be constrained in the same way.

Regardless of this, there is still a strong qualitative fit between the STST data and the human data. It is also important to note that we have taken the STST model exactly as it was formulated 10 years ago, i.e. in Bowman & Wyble, 2007. Most notably, we have not refitted the parameters of the model in order to improve the match to the experimental data presented in this paper. This surely means that the match between model and experimental data is not going to be quantitatively perfect. In this respect, it is perhaps only reasonable to just expect a qualitative match between model and experimental results. In this context, the quality of match to the empirical data is, we would argue, impressive. Most importantly, the simulations we have run with STST have provided a proof of principle that the explanation presented in Figure 38 and Figure 39 for why report accuracy and subjective visibility diverge at lag-1 is tenable. This explanation rests on the concept that encoding into working memory can proceed in parallel, but subjective experience cannot. The natural electrophysiological correlate of this is a time-extended P3 when both T1 and T2 are consciously perceived, as opposed to just T1. This is what we observe in our data (See, e.g. Figure 41 and Figure 42).

There is also the question of the use of the P3(b) as a correlate of subjective experience. This has been addressed quite extensively (especially with respect to this dataset) in (Pincham, Bowman et al. 2016), however we replicate the argument here (albeit largely quoting the original) because of its importance.

Overall, in our specific dataset, there seems little evidence that the P3 is responding to report accuracy (Pincham, Bowman et al. 2016). However, there are a wide range of findings that would dispute this conclusion in the literature (Pitts, Padwal et al. 2014, Shafto, Pitts 2015, Squires, Hillyard et al. 1973). Our experiment though has a number of differences to these previous studies, which presumably explain the difference in findings. We discuss these in turn.

1) The stimuli that were being assessed for perceptual experience in these previous studies were typically drawn from a very small set, e.g., a detection task in Squires et al. and two shapes in Pitts et al. Consequently,

the perceptual judgement/encoding into WM processes were much more informationally rich in our study (i.e., identify a letter).

2) The elegant demonstration by Pitts and colleagues (Pitts, Padwal et al. 2014, Shafto, Pitts 2015) that the P3 in their experiment was modulated by task set, does not naturally carry over to our setting. This is because task set, certainly in respect of instruction, is constant throughout our experiment. What modulates the P3 in the current study is behaviour (specifically High vs Low visibility).
3) (Pitts, Padwal et al. 2014) highlighted a number of confounds associated with studies of conscious experience that mean that post-perceptual processes are not equated across conditions. These confounds are, at least to a large extent, resolved in our experiment. In particular, our key P3 comparison contrasts subjective visibility levels, while report accuracy is controlled, i.e., the “same" identity report (although not, of course, the same subjective visibility report) is made for both High and Low subjective visibilities.

Additionally, the P3 in RSVP may present differently to theP3 in paradigms without a rapid sequence of repeated onsets. Importantly, the key P3 finding in RSVP previous to this paper was that a P3 is present when an item breaks through into awareness, and is reported. But to all intents and purposes, the P3 is not present at all when it is not correctly reported in the current study. Our specific claim in this paper is about the P3 as it manifests in this “fringe of awareness" context.

Stochastic vs Deterministic

We created two different interpretations of readouts from the model, one in which the number of time steps corresponds directly to visibility, and one which subjects this readout to noise. The direct readout from the model provides the clearest possible examination of the high vs low visibility hypothesis, while the noisy readout creates a potentially more realistic situation. As expected, the direct readout does provide a much clearer contrast between high and low visibility ERPs. The noisy readout, in comparison, has this sharp distinction reduced. This is entirely expected; not only does the noisy model reduce discriminability between high and low visibility trials through the added noise, but

the direct readout model benefits from setting a high/low visibility split for each lag individually. Perhaps slightly less expected, the noisy readout does however produce behavioural results that more strongly resemble the human data. Though both readout methods show much more polarised results than that of the human data, this effect is less severe in the noisy readout model.

While the direct readout was a useful model for informing our hypothesis, in general, we advocate the use of the noisy model. Specifically, the choice to set different thresholds for the high/low visibility for the direct readout at each lag was informative in this specific instance, but it makes any generalisation of the results difficult. Furthermore, the deterministic, perfect translation of P3 length to subjective report is a strong assumption that may not fit well into a wider theoretical context. The noisy readout model fixes both of these issues, and displays similar behaviour.

Dissociations and sight-blind recall

In the previous chapters, we demonstrated evidence for a dissociation between working memory encoding and subjective experience during the attentional blink. Here, we have provided one theoretical interpretation of the results, which we have attempted to validate through computational modelling. It is easy to see that the model results are consistent with the sight-blind recall hypothesis. One only needs T2 to be entirely subsumed by the presence of a large T1 for a report of zero subjective visibility to occur for an item that the model has encoded into working memory. In terms of dissociations, the same result implies that a dissociation exists, and that it is of working memory encoding as a necessary, but not a sufficient condition for subjective experience. We are able to come to this conclusion more strongly because the relationship between internal cognition and report in the model is not subject to uncertainty in the same way as it is for human data. Instead, it is a parameter under our control, not an uncertainty to be accounted for.

Interestingly, the model as described would also seem to be able to theoretically accommodate the dual effect, high subjective experience in the absence of working memory encoding. This effect is unlikely to be large, but so far as it is possible for the model to generate a P3 for targets that are not encoded into working memory, so too is it possible to generate a subjective report of them, and

so too is it possible to generate a subjective experience of an unencoded target. Though we have not examined it specifically in this chapter, a theoretical situation that might cause this to arise in the model would be a strongly active target that somehow fails to make it into working memory. This may arise if a highly active target occurred while the blaster was being held low. This target would have to be sufficiently active to cause a significant and elongated P3, but not active enough to push through the blaster. Such a situation may arise at a very short SOA with closely placed targets. This is an interesting prediction of our model that bears further study.

Though there is no direct evidence in our own data that encoding-absent experience can occur, such a finding would sit very well with the phenomenal consciousness and associated findings that we discussed in the literature review. If one were able to find evidence of working memory encoding absent experience alongside sight-blind recall, this would indicate that as well as working memory encoding being an insufficient condition for subjective experience, it is also an unnecessary condition. Put another way, this would be evidence working memory encoding and subjective experience would be highly correlated but mutually independent processes.

Conclusion

In this chapter we have attempted to explore one possible interpretation of the data from the previous two chapters – that working memory encoding of targets occurs in serial, while their subjective experience occurs simultaneously. Our tool of choice for this exploration is computational modelling, and we achieved this by building on top of an existing model of the attentional blink, the Simultaneous Type/Serial Token model. The results from our model strongly match those from the human data, indicating that there may be some truth to the hypothesis we have put forward.

7. Meta-Experience in the Attentional Blink

Abstract

In previous chapters, we made use of state-trace analysis to demonstrate evidence for a dissociation between working memory encoding and subjective experience in the attentional blink, and explored one theory of these results through computational modelling. However, all of the work we have so far is based upon an implicit assumption that the averaged behaviour of our data is representative of the trial-by-trial behaviour in the data. That is, it has been assumed that the general trends of the data and the proposed sight-blind recall effect have been the result of a more generalised "metacognitive" (or "meta-experiential" in our case) failure than the statement of sight-blind recall we have currently. This would amount to participants at early lags becoming *systematically* worse at subjective report when they correctly identified stimuli. To avoid confusion of nomenclature, since our data measures "subjective experience" instead of "confidence" of the second target, which a metacognitive measure would normally make use of, we refer to metacognition measures calculated over *our* data as "meta-experience", and "metacognitive" effects as "meta-experiential".

In this chapter, we examine this assumption of a "meta-experiential" failure, both because it is theoretically interesting in its own right, and because it is an opportunity to critically appraise the computational model of subjective report we developed in the previous chapter. One of the tools we use to achieve this is a measure of metacognition that allows us to quantify the accuracy/subjective visibility coupling across lags. While there exist several methods for this calculation, none are applicable outside of *detection* tasks. One of the contributions of this chapter is therefore the development of a general method by which metacognition can be indexed for a more general class of *identification* tasks. Further,

Introduction

Previously, we have studied the possibility of a dissociation between working memory encoding and subjective experience during the attentional blink. In

previous chapters, we concluded that there was evidence that such a dissociation exists, and that it potentially arises from a phenomenon we call sight-blind recall, in which participants report minimal subjective visibility despite correctly reporting targets at a rate significantly above chance. Furthermore, based on these results, we built a computational model of subjective experience during the attentional blink that predicted the behaviour of the human data. However, our current analysis is not conclusive. Implicitly, so far, we have been making the assumption that the general trend of subjective report decreasing at early lags while report accuracy increases implies that the coupling between report accuracy and subjective report is worse at early lags. Further, consistent with the sight-blind recall interpretation, we have also been assuming a direction for this effect – items in correct trials are experienced less well as lag decreases.

Discarding briefly our prior expectations, our results are not necessarily logically consistent with this conclusion. An equally valid (if perhaps unlikely) alternative hypothesis on the basis of our evidence is that, despite a shift in average behaviour, subjective reports actually match better with objective reports at early lags, and are completely decoupled at late lags. This may also produce a dissociation that could be identified by state-trace analysis, and would not necessarily be distinguishable at the level of average behaviour. Few would argue this probable, but this thought experiment demonstrates that we have been making assumptions about how our data is behaving without first confirming them.

Theoretically, this behaviour is also of interest. It is potentially highly informative for theories of the attentional blink and indeed, a generalised meta-experiential failure of the type we have been implicitly assuming exists in the attentional blink would be a compelling finding. We may also find, for example, a meta-experiential impairment occurring not just as an impairment of subjectively experience correctly identified stimuli, but also as a consistently high visibility report of stimuli that are incorrectly identified. Depending on the results, such behaviour might be argued to indicate either some kind of illusory percepts (Maniscalco, Lau 2012), or the sight blind recall previously discussed. This also provides us an excellent opportunity to test the computational model that we developed in the previous chapter. If the model is valid, then it should be able to predict all the

subsequent meta-experiential behaviours that we observe in this chapter as well as the results in the previous chapter, with no further additions or modifications to its code.

One of the problems we have described as causing our uncertainty about the parity between accuracy and subjective report is the problem of averaging – based on averages we only know the mean behaviour, we do not know how accuracy and subjective report correspond on a trial-to-trial basis. One simple approach that responds to this problem is to assess the average visibility ratings for correct and incorrect trials separately. If our assumptions about generalised "meta-experiential" failure existed along the lines of sight-blind recall, we would expect to find that correct trials have lower visibility ratings as targets approach one another, while the average visibility rating for incorrect trials might remain unchanged. Equivalently, we might look at accuracy for high and low visibilities across lags. If the same hypothesis is true, we would expect the rate of correctness for low visibility trials to increase as lag decreases.

Ideally, we would like a measure that quantifies coupling, and can be examined across lags. Fortunately, what we are describing has already been studied, and corresponds to the measures of metacognition described in the metacognition section of our literature review. In the literature review, we discussed several different approaches to the calculation of metacognition that are prominent in the literature, and they are (almost) ideally suited to being applied in our case. Unfortunately, despite the strength of these approaches, they are near uniformly united by having a basis in Signal Detection Theory (SDT) (Fleming, Lau 2014, Maniscalco, Lau 2012). This is problematic; SDT based approaches are only suitable when the Type 1 task is either currently a detection task, or can be broken down into an analogous binary choice. This is not the case in our attentional blink paradigm, which is an identification task with 21 different stimuli and 21 different outcomes. We therefore develop our own approach to the calculation of metacognitive sensitivity that captures the benefits of these existing approaches, but without the limitation to detection (binary choice) tasks.

Metacognition – A General Approach

As discussed, while existing SDT-based approaches work well for tasks to which they are applicable, they are difficult to generalise. A large part of this problem is

that the SDT framework is so dependent on the true-positive/false-positive rates that are calculable in a binary task in order to evaluate sensitivity. Generalising these measures to $k > 2$ outcomes is theoretically possible inside the SDT framework (Ashby, Soto 2015), but rapidly loses its intuitive meaning (Thomas 1999). Furthermore, it causes the problem to explode in complexity – discriminability becomes dependent on $\frac{k(k+1)}{2}$ different quantities and most importantly this causes us to lose any simple analogy to the key d' measure that embodies sensitivity (Thomas 1999). It also calls into question the key assumptions underlying a simple SDT model – is it really plausible to model 21 different letter stimuli as choices along a single decision axis? If not, our problem rapidly becomes yet more complicated.

In light of this, we propose that whatever measure we use for discriminability, it must fundamentally be simple - the complexity of calculation cannot grow in the way described above. Equally, it must be a measure of discriminability uncontaminated by bias. Our method of choice is *Discrete Mutual Information* (MI). MI calculates sensitivity (or discriminability), in a sense, bias free; it is clear that MI is not concerned with the criteria participants use to distinguish stimuli or how well responses correctly correspond to stimuli, rather it assesses how well participants' responses discriminate between the stimuli presented. Even further, it measures it in units with intuitive meaning and high interpretability – bits of information.

Mutual Information

Unless otherwise specified, we will use the term mutual information to refer to *discrete* mutual information, which is an information theoretic measure that quantifies the amount of information that one (discrete) probability distribution X, tells us about another (discrete) probability distribution Y. In this section, we will describe, from the ground up, what information means in this context, and how this measure of *mutual* information can be constructed from this definition.

We begin with the concept of *self-information* or *information content*, which defines the amount of information gained from any particular outcome x_i from the discrete probability distribution X. More formally, let X be a discrete

probability distribution with x_i a result of sampling X with $P(x_i) = p_i$ with $0 \leq p_i \leq 1, \forall\, p_i$, then the *self information* of x_i is:

$$I(x_i) = -\log p_i$$

Intuitively, the rarer the chance of x_i occurring, the more information is conveyed by its appearance. Self-information quantifies this, in units of information. The specific unit of information used depends on the base of the logarithm chosen in the calculation. In this paper, we opt for the use of *bits* of information based on a logarithm of base 2. For an event with two equally likely alternatives e_i $(p_i = 0.5)$, $I(x_i) = 1$. This means that one *bit* of information corresponds to the ability to distinguish two equally likely alternatives.

This defines the amount of information from a single event, but what we are interested in in mutual information is the amount of information a system of events carries about another system of events. In order to progress toward this, we next define the amount of information transmitted from not just a single event, but across sets of events. This measure is known as *entropy*. To formalise, given a set of events X, the entropy of this set of events is the expected value of the self information across *all* events, i.e.

$$H(E) = -\sum_{\forall i} p_i \log p_i$$

An interesting facet of entropy is that while it is, by construction, a measure of expected information transfer, it can also be thought of as a measure of disorder. In particular, entropy is maximal when the underlying distribution of p_i is uniform, and minimal when it is concentrated into one single point. It is also worth noting that entropy can be defined for joint probability distributions, and is defined identically in this case.

With the definition of the information carried not just by a single event, but by a set of events as a whole, we can now go on to define the *mutual information* that one set of events carries about another. Let X and Y be discrete probability distributions with joint distribution (X, Y), then the mutual information of X and Y is defined as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

That is, the amount of information carried by each distribution on its own, *less* the information that can be gathered from the joint distribution of the two on its own.

For example, consider a participant presented with one of four letters chosen uniformly at random, with the task of identifying them. We wish to calculate the information carried by the responses (R) about the stimuli (S), $I(S;R)$.There are four letters that are all equally likely to be presented, so the stimulus distribution carries 2 bits of information($\sum_{\forall i} p_i \log p_i = -4 \times (0.25 \log_2 0.25) = 2$). If the participant correctly identifies every letter, the response and joint distributions will also carry 2 bits of information, so $I(S;R) = 2 + 2 - 2 = 2$ bits. This is an intuitive result – if responses perfectly match stimuli then by knowing 2 bits of information about the responses, we know 2 bits of information about the stimuli. Consider the converse, our participant responds completely at random. In this instance, our stimulus and response distributions would still carry 2 bits of information, but this time the joint distribution will carry 4 bits of information $\sum_{\forall i} p_i \log p_i = -16 \times \left(\frac{1}{16} \log_2 \frac{1}{16}\right) = 4$. Mutual information will then be $I(S;R) = 2 + 2 - 4 = 0$ – another intuitive result. If responses are completely uncorrelated with stimuli, then knowing the responses gives us no information about the stimuli. We could also construct examples in which participants correctly identify the solution some of the time and in this case $I(S;R)$ would lie proportionally between these values.

From Mutual Information to Metacognition

In the context of our experiment, we apply mutual information to stimulus/response correspondence. However, it is unclear how to go from mutual information as we have set it out now to a true measure of metacognitive sensitivity. Somehow visibility must be brought into the equation. We could approach this in a similar way to the SDT methods discussed previously, for example calculating a "Meta MI" analogous to the "Meta d'" we reviewed in the literature review. However, we feel that such methods are likely to be overcomplicated; while such methods were justified in the Meta d' case, it only came to using these complex measures as an alternative to simple measures because of complicated facets of SDT itself (Barrett, Dienes et al. 2013), particularly the violations of the strong assumptions that SDT makes (Fleming, Lau 2014); for example the normality of data. We have no reason to believe any

similar problems apply to our measure of MI, so we therefore opt for the simplest option possible; we divide our visibility report into high/low bins (NB: there are other good reasons for wanting to do this, see below) and take our Type 2 metacognitive measure for any given condition as the difference in Type 1 mutual information between the two.

Metacognition - Challenges

Challenge 1 – Accurate Mutual Information Estimation

The application of this Mutual Information measure is not without its challenges. A problem with our analysis as it stands is that it relies on an *estimation* of mutual information based on (biased) estimated entropy calculations. The calculation of true, unbiased discrete entropy and therefore mutual information between two variables is simple, *assuming that one has perfect knowledge of the probability distributions of those variables*. In practice, we can rarely claim such knowledge and must instead rely on approximating these probability distributions by repeatedly taking a finite number of samples of them (Paninski 2003). Consider our case of stimuli and responses. In order to calculate the mutual information between these two measures perfectly we would have to have perfect knowledge of the probabilities of giving each response to each stimulus. Clearly we do not have this, and must estimate these imperfectly from the data we collect, leading to our estimate of the entropy or mutual information of this distribution to be imperfect. To quantify this imperfection, we introduce the measures of bias (β) and variance (v). Bias measures how far our entropy estimate is from the true value, it is the expectation of the deviation of our entropy estimate from the true value. Variance measures how much our entropy estimate varies, and is defined as the expectation of the squared deviation of the entropy estimate from its mean.

Clearly, by the law of large numbers, one way we can improve our estimate of entropy (and decrease these measures) is to improve our estimate of the probability distribution from which it is calculated by basing it on a larger number of samples from the true distribution. In this instance, we could simply set some acceptable bias and variance levels for our entropy calculation, say δ_β and δ_v, and increase n until both $\delta_\beta > \beta$ and $\delta_v > v$. However, in practice, n is often fixed, so if

we wish to approach some specific level of acceptable deviation, we must rely on other approaches to reduce the bias and variance.

Before we discuss these approaches however, we need to establish a practical method of testing them. We've established the metrics we'll use to compare our different approaches, β and υ, but need to be able to calculate and compare these between methods. In order to do this, we need a distribution X that we can create random samples $\hat{X}_n^i$ from, and whose entropy $H(X)$ can be calculated analytically. From this, we can calculate the deviation of our entropy estimator $\widehat{H}(\hat{X}_n^i)$ from the true entropy as $\widehat{H}(\hat{X}_n^i) - H(X)$, from which we can calculate estimates of our β and υ by looking at the expectation for over all samples. For our X, the normal distribution suits this role perfectly: creating random samples from the normal distribution is trivial, and both entropy and mutual information can be calculated for normal distributions knowing only their mean and variance (Misra, Singh et al. 2005). There is a small obstacle for this method, in that these calculations strictly only apply for the calculation of entropy over the *continuous* normal distribution, and our interest is in *discrete* entropy over discrete distributions. However with a discrete distribution defined over a reasonable number of bins, the difference is acceptably small (Beirlant, Dudewicz et al. 1997). In our testing we set the number of bins to be the same number of bins we use in our analysis – 21.

Our testing procedure for a given approximation method $\widehat{H}(\hat{X}_n^i)$ is then as follows. We start with two normal distributions X and Y with known $\mu_x, \mu_y, \sigma_x, \sigma_y$, and correlation ρ. We then calculate our known marginal and joint entropies for normal distributions, as well as our MI as follows (Misra, Singh et al. 2005):

$$\mathrm{H(X)} = \log(\sigma_x\sqrt{2\pi e})$$

$$\mathrm{H(Y)} = \log(\sigma_y\sqrt{2\pi e})$$

$$H(X,Y) = 1 + \log(2\pi) + \frac{1}{2}\log(|K_{XY}|)$$

$$MI(X;Y) = -\frac{1}{2}\log(1-\rho^2)$$

$|K_{xy}|$ *is the determinant of the covariance matrix of the joint distribution of* X *and* Y

We then set a range of sample sizes $N = \{n_1, \dots, n_L\}$ over which we will evaluate our bias, and for each n_i, we take 10 random samples of size n_i from our X and Y for each of 10 different correlations $\rho = \{0, 0.06, 0.12, \dots, 0.6\}$. This range of ρ is selected to provide a range of conditions under which to evaluate MI. Conspicuously, it is missing large values of ρ, but this is a requirement borne of our use of a continuous estimator of discrete mutual information: a large value of ρ demands that a support size of the joint distribution is proportionally shrunk across one diagonal axis. With a finite support size, this means that increasing ρ will eventually force the joint distribution to such a small support on one diagonal that it is no longer a good approximation of a normal distribution. An illustration of this can be seen in *Figure* 46. Empirically, we find ρ exceeding 0.6 begins to cause this, and thus restricted our ρ a commensurate amount. From these values, we create our estimated discrete distributions for X and Y, $\widehat{X_{n_i}} = \{\hat{X}_{n_i}^1, \dots, \hat{X}_{n_i}^{10}\}$ $\widehat{Y_{n_i}} = \{\hat{Y}_{n_i}^1, \dots, \hat{Y}_{n_i}^{10}\}$ and evaluate the following:

$$\varepsilon_{H(X)} = \widehat{H}\left(\hat{X}_{n_i}^j\right) - H(X)$$

$$\varepsilon_{H(Y)} = \widehat{H}\left(\hat{Y}_{n_i}^j\right) - H(Y)$$

$$\varepsilon_{H(X,Y)} = \widehat{H}\left(\hat{X}_{n_i}^j \otimes \hat{Y}_{n_i}^j\right) - H(X,Y)$$

$$\varepsilon_{MI(X;Y)} = \left(\widehat{H}\left(\hat{X}_{n_i}^j\right) + \widehat{H}\left(\hat{Y}_{n_i}^j\right) - \widehat{H}\left(\hat{X}_{n_i}^j \otimes \hat{Y}_{n_i}^j\right)\right) - MI(X;Y)$$

We can then create scatter plots out these errors for each of our sample sizes, for each of our calculations, and examine how the pattern of behaviour changes between our methods. We can also estimate the bias and variance for each of our sample sizes, for each of our comparisons. However, since we do not have any information about any underlying distributions of errors for entropy we favour scatter plots for visual presentation.

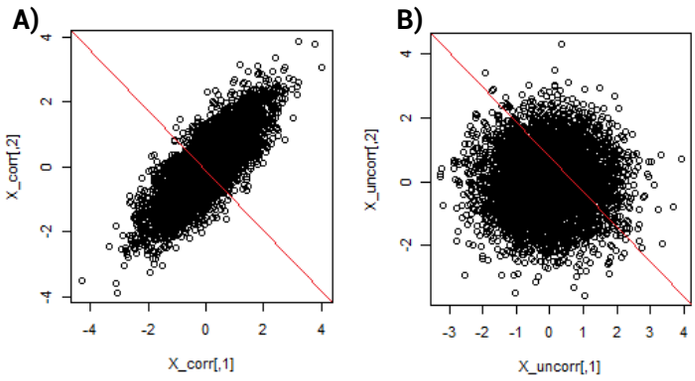


Figure 46) A) A normal distribution with $\rho = 0.5$, B) An (uncorrelated) normal distribution with $\rho = 0$. Notice that in A), the support size of the distribution across the red diagonal is effectively shrunk compared to B). With a discrete and finite number of bins, as ρ increases and the support size shrinks, the joint distribution will eventually be supported on so few points on the red diagonal that it is no longer a good approximation of a continuous normal distribution.

The first method for which we perform this examination is the so-called *plugin estimator* of entropy. For this, we simply assume our estimates of X and Y are accurate, and simply "plug in" the values from our estimated distributions to the entropy formula i.e.:

$$\widehat{H}_{plugin}(\hat{X}_n^i) = H(\hat{X}_n^i) = -\sum_{\forall j} p_j \log(p_j)$$

Applying our testing procedure to this, results in the following:

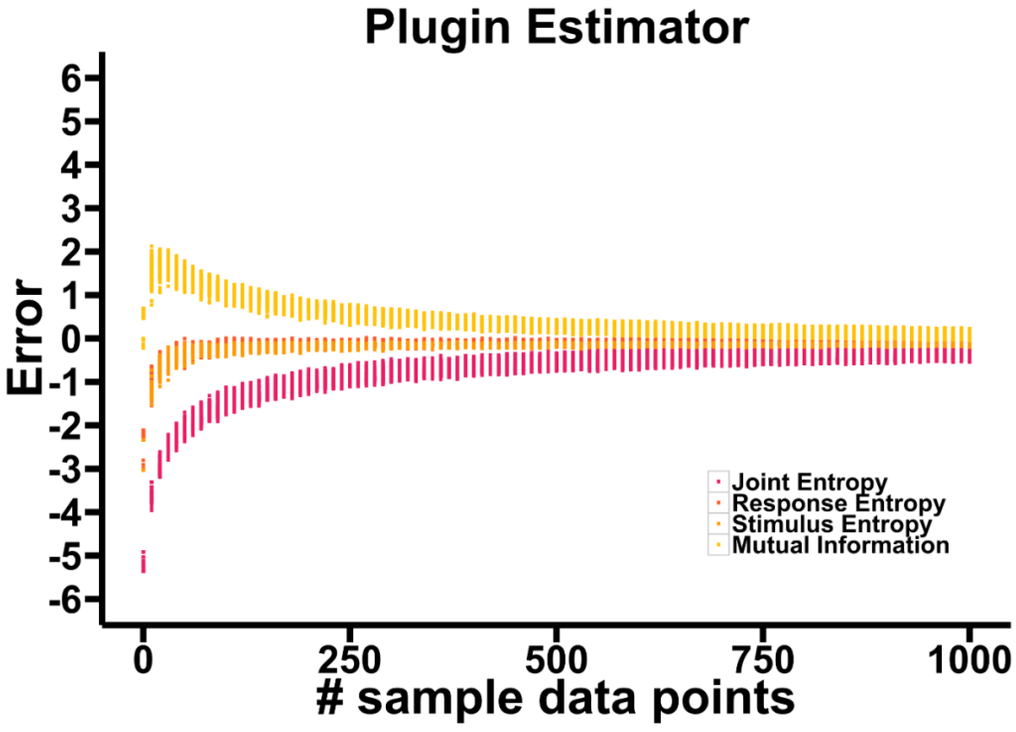


Figure 47) Error of the plugin estimator of stimulus, response and joint entropy, and mutual information for a range of different sample sizes from 1 to 1000. Response and stimulus

entropies are effectively identical, and so stimulus entropy (orange) overlays response entropy (red) on the plot.

This brings us to the demonstration of the fundamental problem we have been discussing in this section. Entropy estimation based on finite samples is not only biased, but *badly* so for small sample sizes. In this case, we are evaluating samples of between size 8 and size 1200, and in the worst case of our joint probability distribution with 441 bins, even at the limit of 1200 samples, our joint distribution and MI estimators are still noticeably biased.

If we consider the practical applications of this for our own work, it is clear we have a problem: we have a maximum of 196 trials *per participant*, that are spread across 7 lag $\times$ 6 visibility bin combinations. Even worse, our trial counts and therefore our bias will not be evenly distributed just as trials are not evenly distributed across lag $\times$ visibility combinations. It is clear that if we were to use this naïve estimator in our own work, then our results would suffer from noticeable bias.

One step we can take to improve the situation is to reduce the number of bins the data will be distributed across. In the previous chapter, we, for various reasons, merged our 6 visibility bins into 2 bins – high and low. Doing this improves our situation slightly, averaging around 14 data points per lag $\times$ visibility combination. However, it is clear from our previous plot, and our next, that examines our testing procedure across a more realistic set of sample sizes for our data, that our plug-in estimator is still not viable.

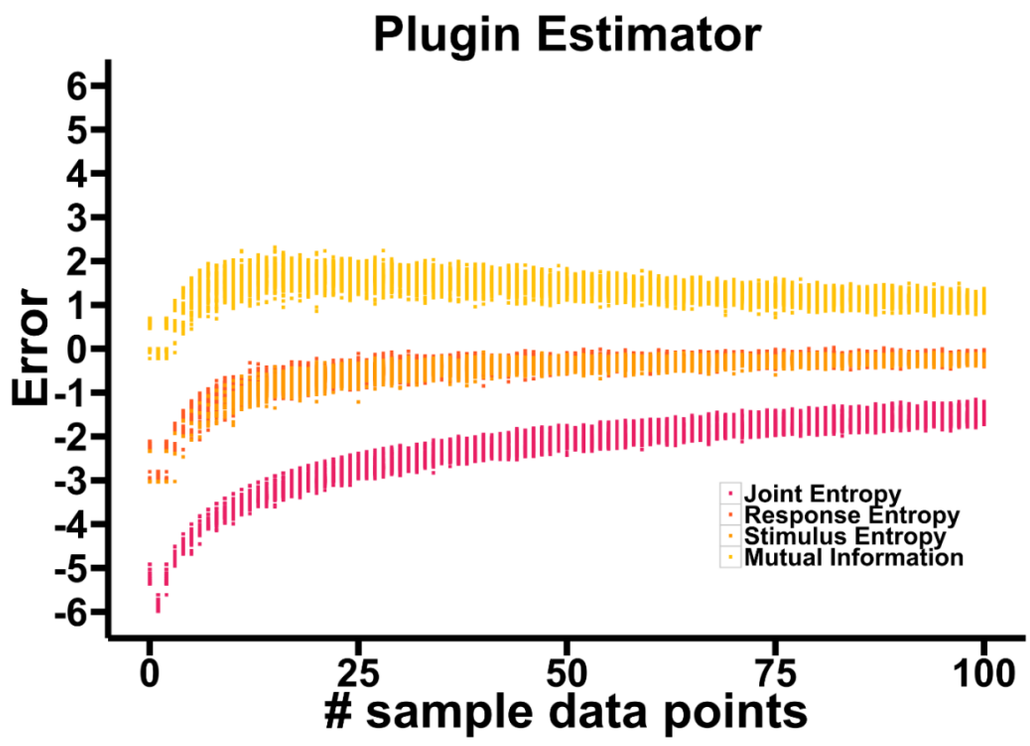


Figure 48) Error of the plugin estimator of stimulus, response and joint entropy, and mutual information for a range of different sample sizes from 1 to 100. Response and stimulus entropies are effectively identical, and so stimulus entropy (orange) overlays response entropy (red) on the plot.

In fact, previous results quantify this problem for us. Let S be our number of discrete bins and let n be our sample size, then for our plugin estimator, bias is $\beta = O(\frac{S}{n})$, and variance $v = O(\frac{\log(S)^2}{n})$ (Paninski 2003).

One approach to improve our estimator, is simply to subtract this bias out. This is called the Miller-Madow correction (Miller 1955). Bias is more precisely

$$\frac{S-1}{2n} + O(n^{-1})$$

And so the formula for Miller-Madow corrected entropy is (Paninski 2003):

$$\hat{H}_{MM}(\hat{X}_n^i) = \hat{H}_{plugin}(\hat{X}_n^i) - \frac{S-1}{2n}$$

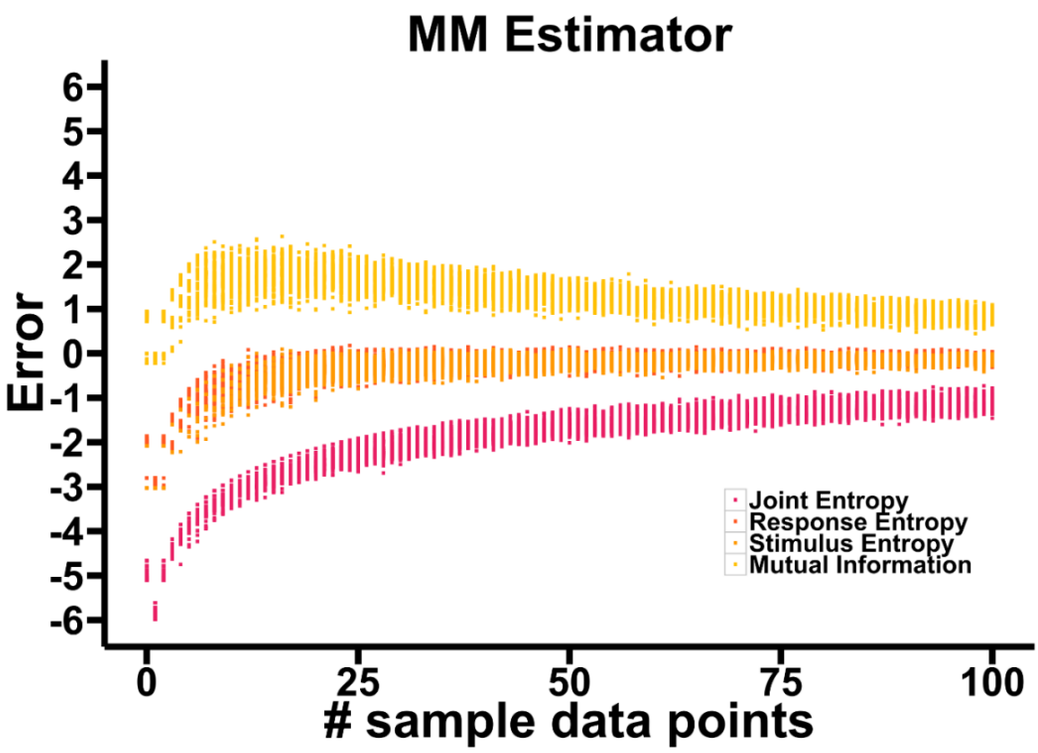


Figure 49) Error of the Miller-Madow estimator of stimulus, response and joint entropy, and mutual information for a range of different sample sizes from 1 to 100. Response and stimulus entropies are effectively identical, and so stimulus entropy (orange) overlays response entropy (red) on the plot.

Unfortunately, while this correction might, empirically, provide a small improvement over the plugin estimator (Figure 49 vs. Figure 48), bias is still far too high to be useful, and analytically, bias and variance still remain $O(\frac{S}{n})$ and $O(\frac{\log(S)^2}{n})$. A very large class of estimators of entropy suffer from the same problem, and have the same asymptotic behaviour. We provide some more examples in the appendix, but suffice to say that we need a substantial improvement to make any kind of analysis based on entropy viable.

One result that bears mention at this point is that however we try to improve our estimator; there is no "best" estimator of entropy (Paninski 2003). For any entropy estimator we may design, there is no universal rate at which the errors go to 0; there always exists at least one "bad" distribution for which the convergence rate is infinitely worse than our general convergence rate (Antos, Kontoyiannis 2001). In terms of bias, this places a bound on the bias of *any* entropy estimator we might design: $O(\frac{S}{\sqrt{n\log(n)}})$ (Valiant, Valiant 2011, Han, Jiao et al. 2015, Jiao, Venkat et al. 2017).

When it comes to selecting an alternative entropy estimator that reduces bias to acceptable levels, we must also bear in mind that despite merging our visibility bins, we are still under-sampled. Specifically, for us, $n \ll S$ – even after merging bins: in the worst cases, we have considerably less samples than we do bins. This is known to make accurate entropy estimation particularly difficult (Valiant, Valiant 2011, Han, Jiao et al. 2015, Jiao, Venkat et al. 2017) and limits our choices of viable algorithm.

Given that there is no best estimator, and that we are working in a regime that is known to be difficult, in order to evaluate the best estimator, we opt to test several estimators that are known to estimate entropy well (particularly for $n \ll S$). For this, we consider three different estimators: the JVHW estimator (Jiao, Venkat et al. 2017), the NSB estimator (Nemenman, Shafee et al. 2002) and the HS estimator (Hausser, Strimmer 2009).

The JVHW (Jiao-Venkat-Han-Weissman) is the closest to a minimax estimator that exists for entropy. It has been shown that a generalised minimax estimator for entropy does not exist (Paninski 2003) for entropy, but the JVHW algorithm avoids conflicting with this result by breaking the problem down into two domains: an easy “minimax solvable" domain and a hard “minimax unsolvable" domain. When the problem is solvable, optimal minimax rates are achieved, and when it is not, results are as close to minimax as possible (Jiao, Venkat et al. 2017).

The NSB and HS estimators take another approach, and are based on very similar ideas. Classically, entropy estimators are biased downward because, with fewer data points, it is difficult to capture the underlying “smoothness" of the probability distributions from which entropy is being estimated. Both of these methods attempt to counter this by performing “shrinkage" toward a uniform distribution, and decreasing this shrinkage as n grows relative to S and estimates of the probability distribution become more accurate. Both also estimate the diversity of the underlying probability distribution based on “coincidence counting" instead of raw bin counts. That is, they count the number of occurrences of multiple samples in the same bin, versus the number that would be expected. This is particularly advantageous in the under-sampled regime because of the “birthday paradox". The birthday paradox is a counterintuitive

result in which the number of people in a room who are expected to share a birthday is unexpectedly high for even a small number of people. Similarly, we might expect datapoints to share a bin after a relatively small number of samples: we can expect coincidences will begin occurring after approximately $\sqrt{S}$ or fewer samples, allowing the capture of the diversity of the distribution with far fewer data points (Nemenman 2011).

The only real differences between the methods are how they achieve this. The NSB method approaches this in a Bayesian manner: it defines a mixture prior that is approximately uninformative over the distribution of entropy. Conversely, the HS method takes a frequentist approach to the problem, using James-Stein shrinkage to pull the estimate of the underlying distribution toward the uniform one. It works by taking a linear combination of the plug-in probability distribution and the uniform probability distribution, based on a shrinkage intensity that decreases as sample size increases. Let $\lambda \in [0,1]$ be shrinkage intensity, $t_S = \frac{1}{S}$ the probability of any bin in the uniform distribution of size S, and $\hat{p}^i_{HS}$ the estimator of the probability of bin $i \epsilon S$ for the HS entropy estimation method, then:

$$\hat{p}^i_{HS} = \lambda t_S + (1-\lambda)\hat{p}^i_{Plugin}$$

And

$$\widehat{H}_{HS} = \sum_{i=1}^{S} \hat{p}^i_{HS} \log(\hat{p}^i_{HS})$$

With optimal shrinkage intensity calculable proportional to the variance of $\hat{p}^i_{Plugin}$. An advantage of this method is that it not only calculates an estimate of entropy, but also calculates a fairly accurate estimate of the underlying distribution form which it is being calculated.

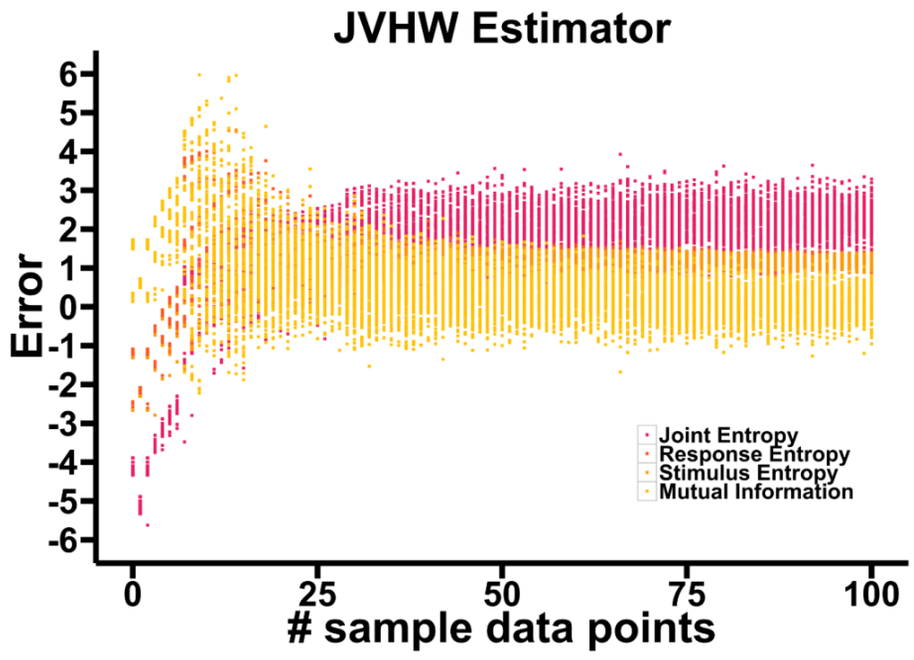


Figure 50) Error of the JVHW estimator of stimulus, response and joint entropy, and mutual information for a range of different sample sizes from 1 to 100. Response and stimulus entropies are effectively identical, and so stimulus entropy (orange) overlays response entropy (red) on the plot.

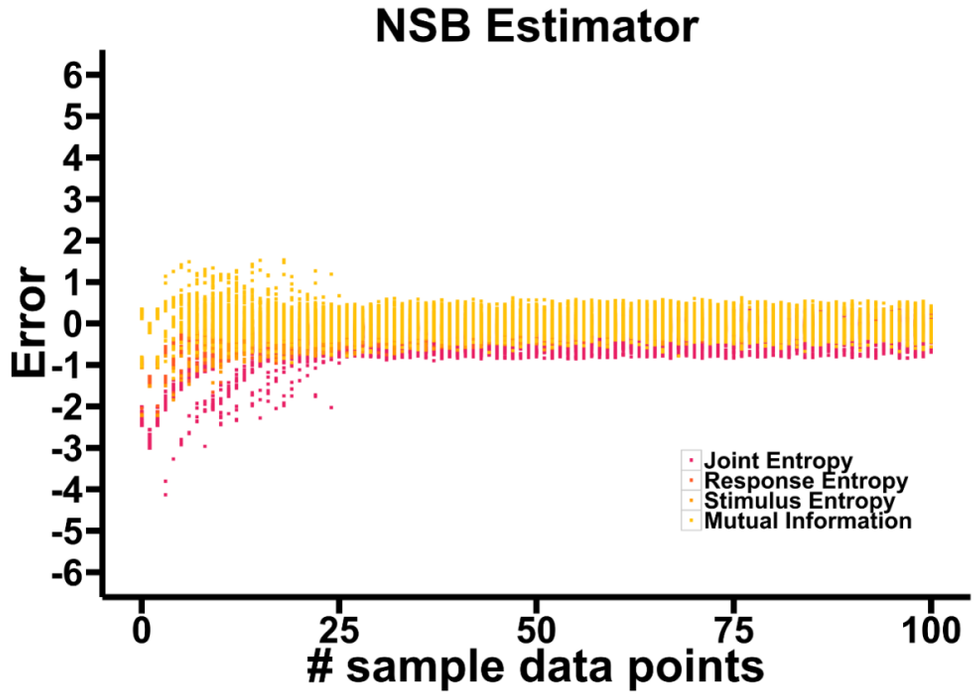


Figure 51) Error of the NSB estimator of stimulus, response and joint entropy, and mutual information for a range of different sample sizes from 1 to 100. Response and stimulus entropies are effectively identical, and so stimulus entropy (orange) overlays response entropy (red) on the plot.

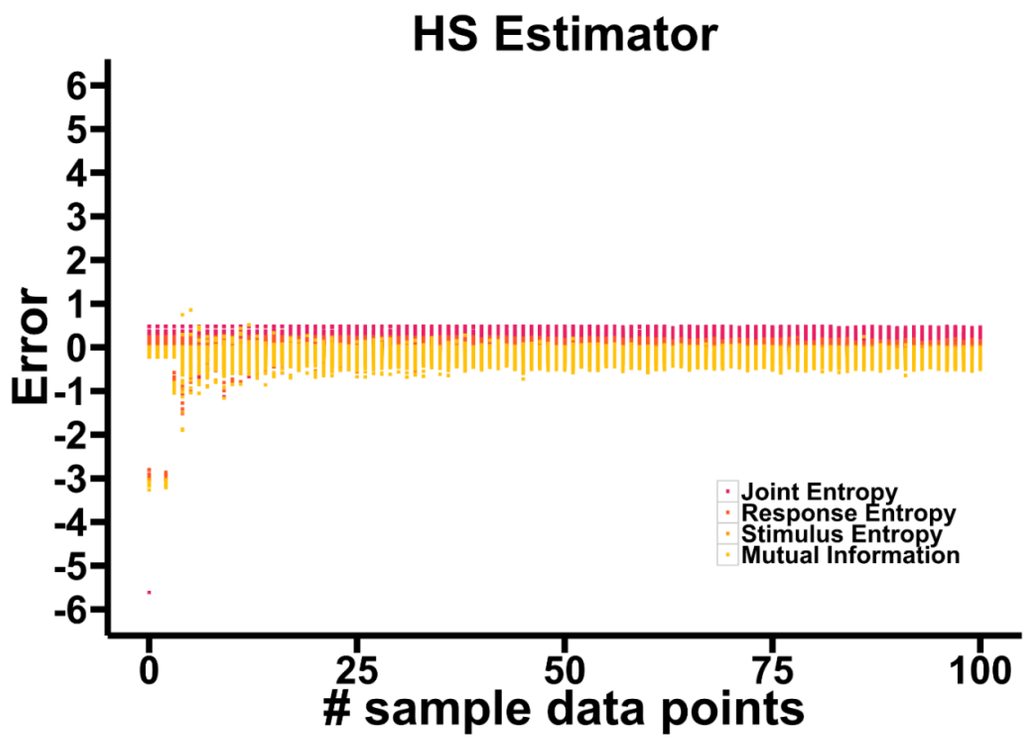


Figure 52) Error of the HS estimator of stimulus, response and joint entropy, and mutual information for a range of different sample sizes from 1 to 100. Response and stimulus entropies are effectively identical, and so stimulus entropy (orange) overlays response entropy (red) on the plot.

Overall, despite its favourable properties, the JVHW (Figure 50) estimator is clearly the worst at the sample sizes we are using. Of the other two, one might argue that the NSB estimator (Figure 51) performs slightly better as the HS estimator (Figure 52) seems to underestimate entropy consistently, but the difference is marginal. As a tiebreak, there is at least one theoretically compelling reasons to pick NSB over HS for our particular experiment: between the NSB and HS estimators, HS provides a much stronger shrinkage. This makes HS ideal in the case in which our true underlying distribution is approaching uniform and entropy is high but, given that given that our average accuracy is at least 50% at each lag, this is demonstrably not the case in our data. Furthermore, with visual similarity between the letters that are our targets (Conrad 1964, Gilmore, Hersh et al. 1979) it is likely that there will be significant orderliness between our stimuli and responses even when participants answer incorrectly. For this reason, HS is likely to overestimate entropy compared to NSB for our data, and NSB will likely prove the superior estimator for us in practice.

Challenge 2 – Statistical Methods

Having calculated mutual information, we turn to the statistical methods we will apply to it. Despite the strong improvements that our NSB entropy estimation methods demonstrated over our original plug-in method, there is both a strong empirical and analytical reason to believe it is still biased, particularly at the sample sizes we are using. This is something that we must account for in whatever statistical methods we apply to the data. Another consideration is that, for a sample size of 0, mutual information is not meaningfully defined. Depending on the situation, one might argue that either 0 or maximum mutual information may be a suitable stand-in, but neither is appropriate in our case. Furthermore, the missing data is not completely random; it is a result of, e.g. early lags being more difficult. This makes any potential data imputation highly complicated. For this reason, whatever analysis method we select, it must either be robust to missing data, or able to somehow manage this complicated imputation problem. With respect to the error of our entropy estimator, there is one saving grace: We have, from the previous section, a very good idea of how the bias changes with sample size. In particular, we are aware that in the worst case, $\beta = O(\frac{S}{n})$ (Paninski 2003) or, since our S is fixed at 21, $\beta = O(\frac{21}{n})$. With this knowledge, we can simply calculate an estimation of the error of $O(\frac{21}{n})$ for each of our samples and add this into the model as a covariate of no interest.

With this in mind and in light of these considerations, we adopt the following model; a mixed effects model in which the dependent variable of Mutual Information is dependent on the fixed effects of Lag and Visibility and Entropy Estimation Error (Referred to as Count), and the random effect of subject. This allows us to capture variance between subjects and entropy estimation error as covariates of no interest Furthermore, mixed effects is also known to be robust to missing data (Krueger, Tian 2004), allowing us to avoid complicated imputation methods or the drastic measure of removing subjects who lack data in a bin. To make mutual information calculation fairly comparable to report accuracy, we only considered trials with T1 correct.

Previously, it has been discussed that, for this analysis, it is beneficial to merge our visibility that is split across 6 bins into 2 bins. Performing either type of

analysis provides other compelling reasons to support this action. Since visibility is, strictly, a discrete measure instead of a continuous one, in this analysis it is dummy coded, and measured with respect to a reference bin. If our data consists of only 2 bins, then we can set our reference bin to be the lower of the two, and our visibility/lag interaction becomes how our meta-experience measure (High visibility – low visibility) varies across lags, allowing us a quick and easy way to determine whether our meta-experience measure is significant.

SESE Model predictions

In the previous chapter, we examined how the mean average accuracy and visibility from the model, and the virtual ERPs in the high/low conditions match those from the human data. Though the model tended to produce more extreme predictions than the human data displayed, our model results have a strong qualitative alignment with the human data. The core thesis of this chapter however, is not about the grand average accuracy and visibility across conditions changes, but how the correspondence between the two measures varies across conditions. In this section, we propose four measures to evaluate this. To provide the fairest test of these measures, we evaluate them over the more general "noisy" model discussed in the last chapter.

The first measure is of how mean visibility ratings change separately for correct and incorrect trials across lags. Similarly, we also propose to measure how mean accuracy changes between high and low visibility ratings, by lag. Though these two measures are different ways to look at the same data, it is helpful to display both. In particular, our mutual information measure is amenable to comparison with accuracy ratings plotted for high and low visibility separately. For the core hypothesis of this chapter though, it is more intuitive to compare how visibility changes for high and low accuracy separately. In this case, if our sight-blind recall hypothesis is borne out by the model, we expect the model to predict that incorrect trials will be rated with uniformly low visibility, while correct trials will be rated at higher visibility, but less visible as the two targets approach one another in time (i.e. lag decreases).

Our final measure is meta-experience. This presents a challenge as the SESE model does not have a concept of target identity; creating a stimulus response matrix as described in previous sections is therefore not possible. To work

around this, we create a loose approximation of our proposed meta-experience measure by substituting in accuracy for mutual information, so our models' meta-experience measure is High Visibility Accuracy minus Low Visibility Accuracy. Though only an approximation to our true meta-experience measure, we note that since mutual information and raw report accuracy are somewhat related, with this method, we can at least establish a loose prediction from the model.

Methods

Our dataset is the same one as discussed previously. As in previous chapters, data was grouped into high/low visibility bins in an identical manner to previously, both for the reasons discussed in the previous section, and to make it comparable to previous results.

In order to determine the effects of our factors on Mutual Information, we fitted a linear regression mixed models (using the R lme4 package (Bates, Sarkar et al. 2007)). The dependent measure in all of these models was Mutual Information (MI), calculated according to

$$MI(X;Y) = \widehat{H}_{NSB}(X) + \widehat{H}_{NSB}(Y) - \widehat{H}_{NSB}(X,Y)$$

With $\widehat{H}_{NSB}$ is the NSB estimator of entropy. Any calculations for which it was not possible to calculate MI (for example, there was no data) were excluded from the analysis. Independent measures, where applicable, were Lag (Lag), Visibility Bin (Vis), the Lag/Visibility interaction (Lag×Vis), Bin Count (Count, our estimator of bias based on sample size – a covariate of no interest), and Subject (Subject). Lag and visibility were both categorical variables, and were dummy coded with respect to Lag 1 and Low visibility respectively. Lag, Visibility Bin, Lag/Visibility interaction and Bin Count were fixed effects; Subject was a random effect on the intercept. We wished to perform three analyses: the effect of Lag, the effect of Visibility, and the effect of the interaction between the two. This necessitated the creation of 5 models, which we denote using the notation from the lme4 package:

$$Null: MI = 1 + Count + (1|Subject)$$

$$Lag: MI = 1 + Lag + Count + (1|Subject)$$

$$Vis: MI = 1 + Vis + Count + (1|Subject)$$

$$Main: MI = 1 + Lag + Vis + \ Count + (1|Subject)$$

$$Full: MI = 1 + Lag + Vis + Lag \times Vis + \ Count + (1|Subject)$$

Models were compared using a chi-square test. For the effect of Lag, we compared the Lag model with the Null model. For the effect of Visibility, we compared the Vis model with the Null model. For the interaction, we compared the Full model with both the Main model and the Null model separately.

Results

Meta-Experience

We evaluated three different effects, the effect of Lag, the effect of Visibility, and the effect of their interaction. For illustrative purposes, as we measure meta-experience by lag as the difference between mutual information for high and low visibility by lag, we also plot a similar difference in accuracy rates by lag for high and low visibility (which we refer to as meta-experience^{A}).

Our Lag model explained only 15% of the Mutual Information variance, and was not better than the null model containing only the Bin Count and a random subject effect. $\chi^2(5) = \ 10.851, \mathrm{p} \ = \ 0.05441$. The Visibility model on the other hand explained 49% of Mutual information variance, and was significantly better than the null model containing only the Bin Count and a random subject effect $\chi^2(5) = \ 105.59, \mathrm{p} < 0.001$. Finally, the Full model explained 63% of the Mutual information variance, and was better than both the null model containing only the Bin Count and a random subject effect $\chi^2(5) = \ 167.62, \mathrm{p} < 0.001$, and a model containing only the main effects of Visibility and Lag, the Bin Count and the random subject effect $\chi^2(5) = \ 36.128, \mathrm{p} = 8.9 \times 10^{-7}$. We provide illustrations of each of these main effects in Figure 53, Figure 54, Figure 55, Figure 56, and Figure 57, with a z scored comparison provided in Figure 58 to allow a more direct comparison.

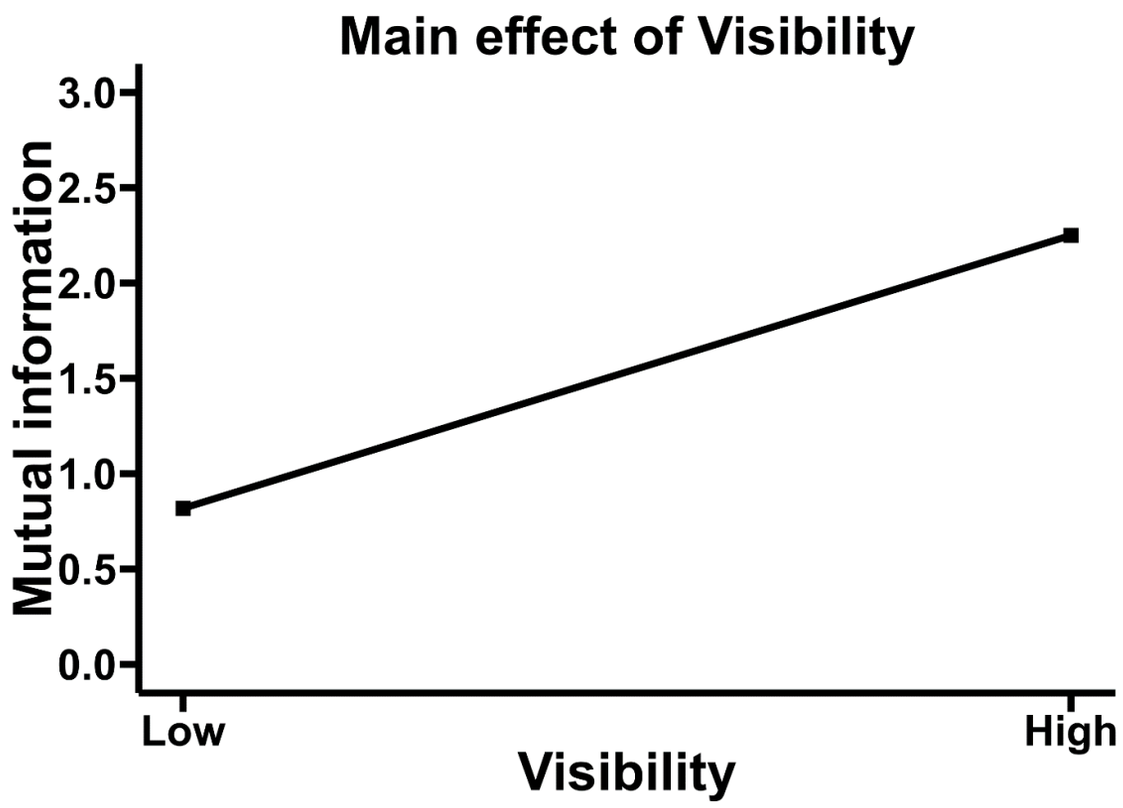


Figure 53) The effect of Visibility on Mutual Information. To illustrate this as clearly as possible, subject has been removed as a factor; MI is calculated based on a stimulus response matrix that is constructed across all Subjects.

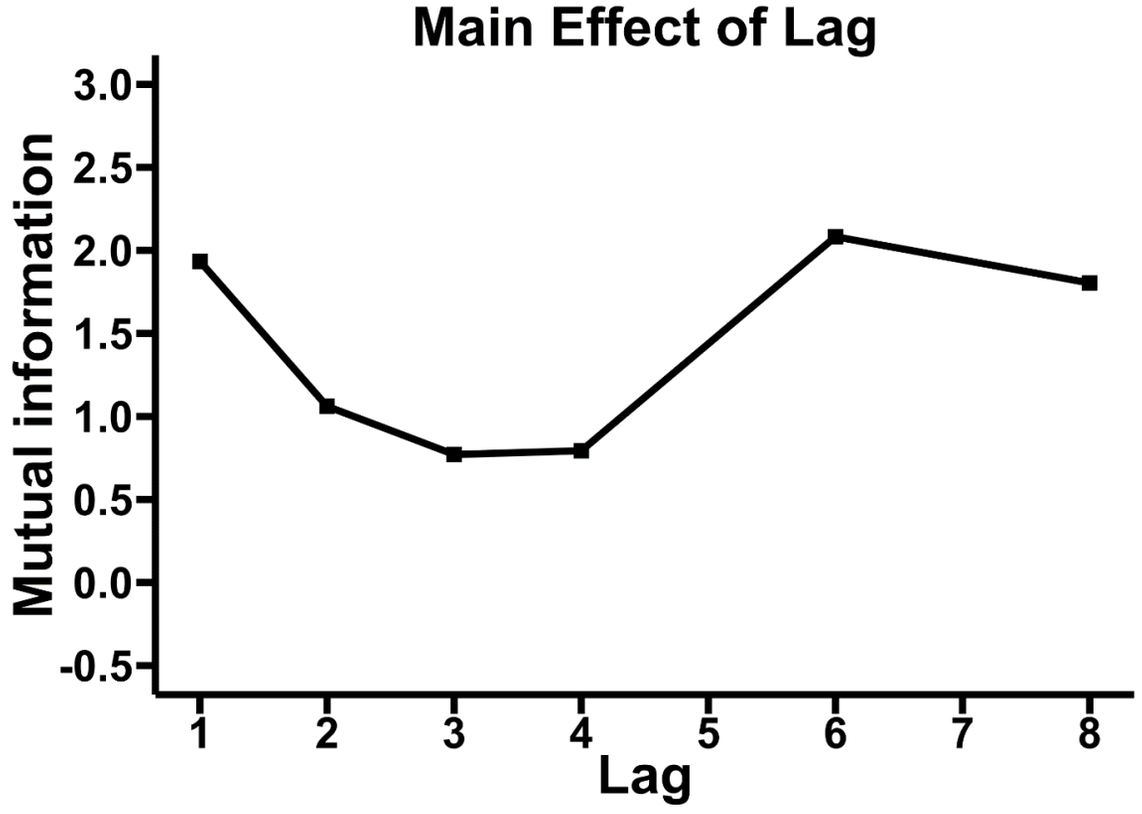


Figure 54) The effect of Lag on Mutual Information. For illustrative purposes, subject has been removed as a factor; MI is calculated based on a stimulus response matrix that is constructed across all Subjects.

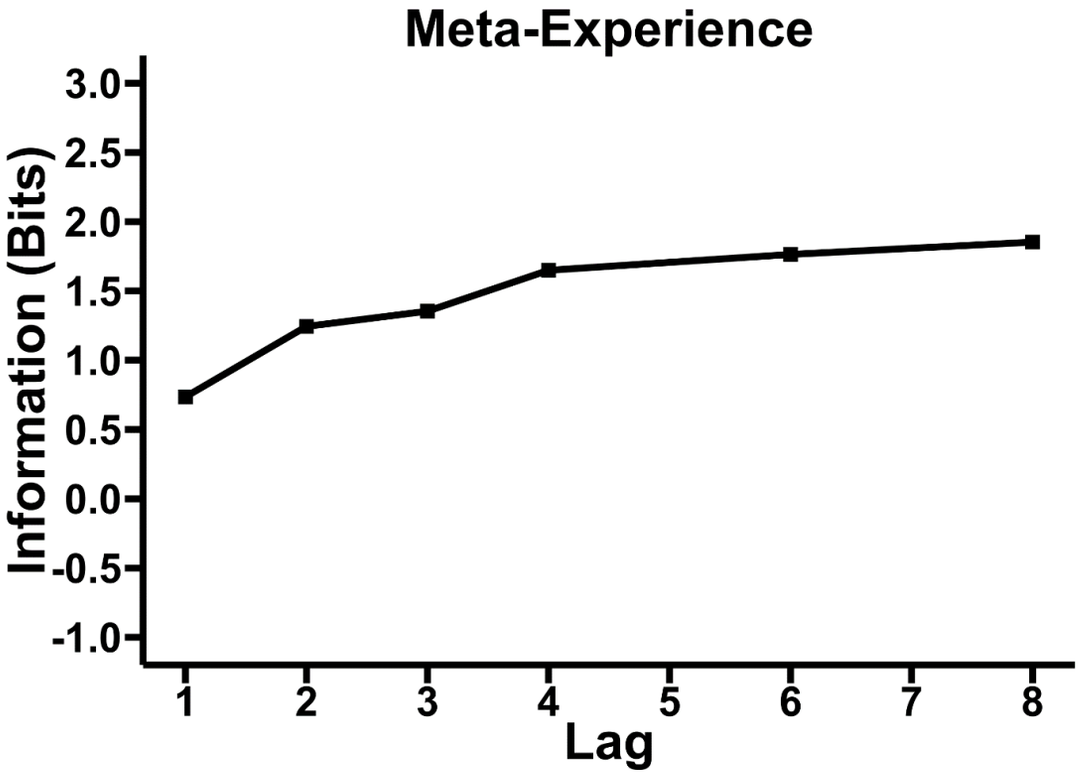


Figure 55) The effect of Lag on Meta-Experience, the difference between MI at high and low visibility ratings. For illustrative purposes,, subject has been removed as a factor; MI is calculated based on a stimulus response matrix that is constructed across all Subjects.

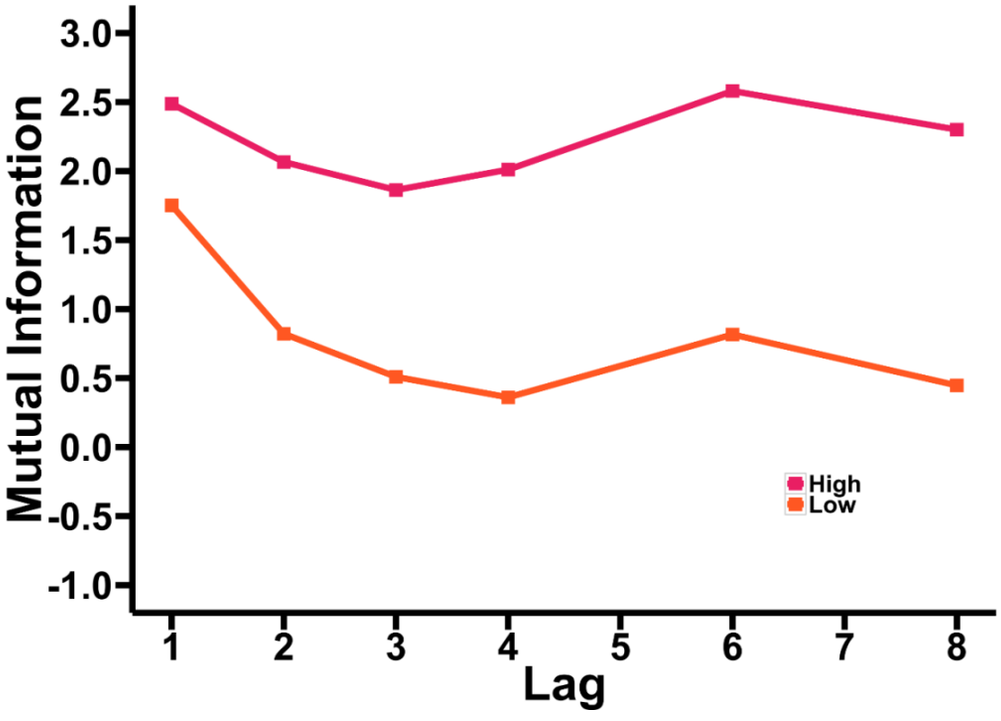


Figure 56) The effect of Lag on Mutual Information, separately for high visibility trials and low visibility trials. The difference between these two is what constitutes our meta-experience by lag. For illustrative purposes, subject has been removed as a factor; MI is calculated based on a stimulus response matrix that is constructed across all Subjects.

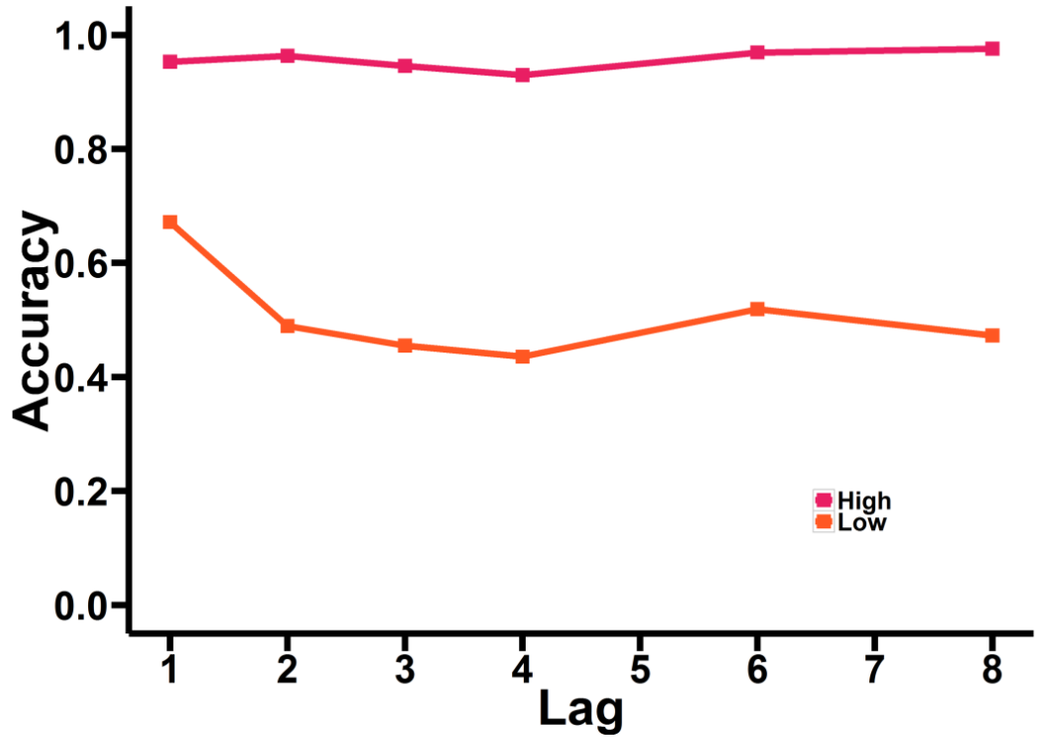


Figure 57) The effect of Lag on accuracy, separately for high visibility trials and low visibility trials. The difference between these two is what constitutes our meta-experience^{A} measure in Figure 58. For illustrative purposes, subject has been removed as a factor; MI is calculated based on a stimulus response matrix that is constructed across all Subjects.

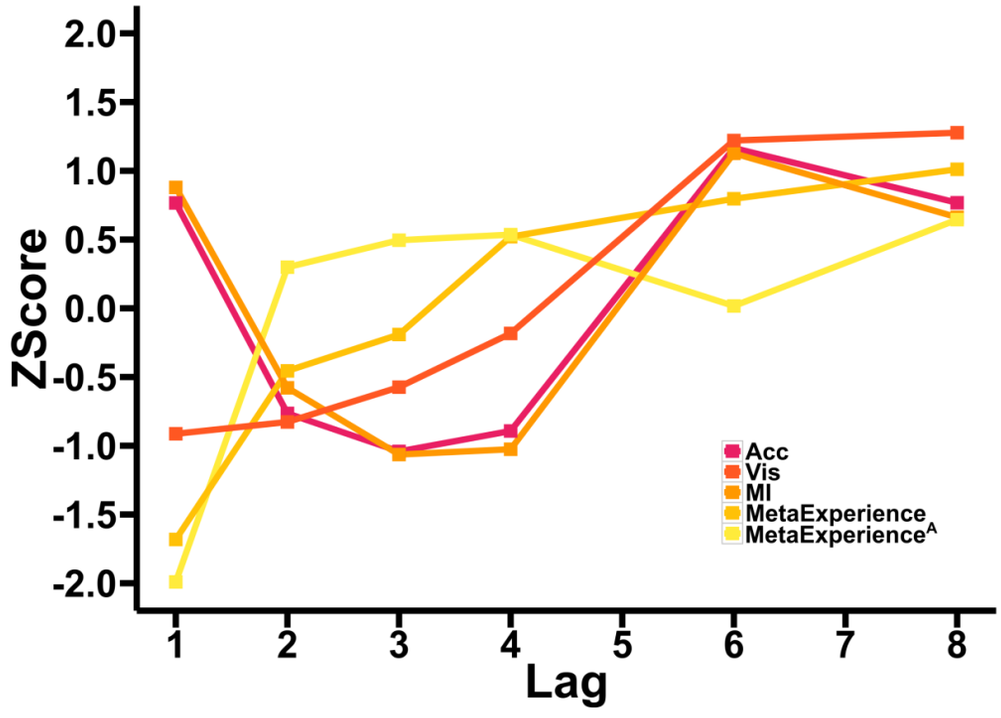


Figure 58) A comparison of Accuracy, Mutual Information, Visibility, Meta-Experience and the measure we refer to here as Meta-Experience^{A}, the difference between high and low accuracy by lag.

Model

In this section, we compare the predictions that the (noisy) SESE model from the previous chapter makes about accuracy, visibility and meta-experience, to those from human data. In the last section, we examined how report accuracy changes separately for high and low visibility trials because it bore a strong resemblance to our meta-experience measure, but for the purposes of our research questions, it is more convenient to examine a different but equivalent question: how visibility changes for correct and incorrect trials separately. We therefore present both.

Correct vs Incorrect

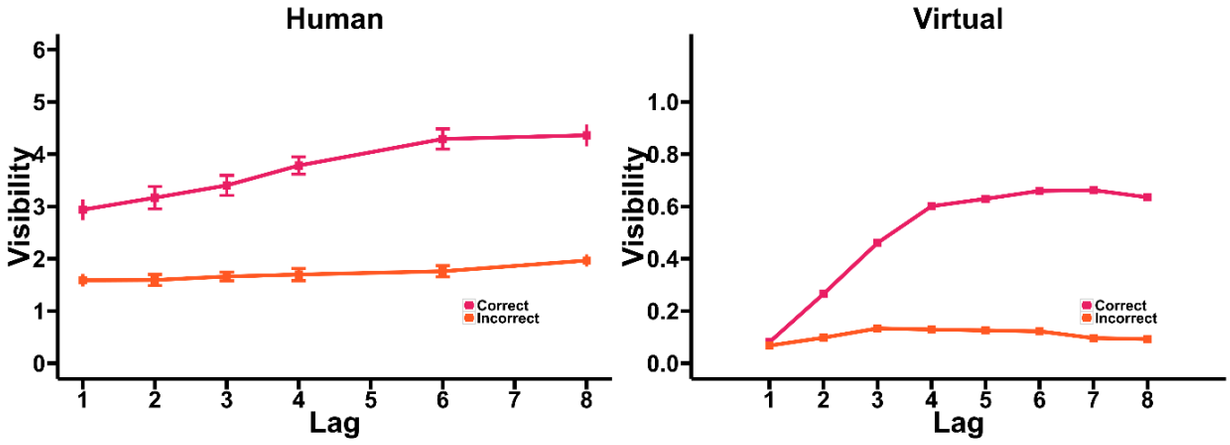


Figure 59) Mean average visibility for correct trials vs mean average visibility for incorrect trials in human behavioural data. Mean average visibility for correct trials vs mean average visibility for incorrect trials in virtual data generated by the STST model.

High Visibility vs Low Visibility

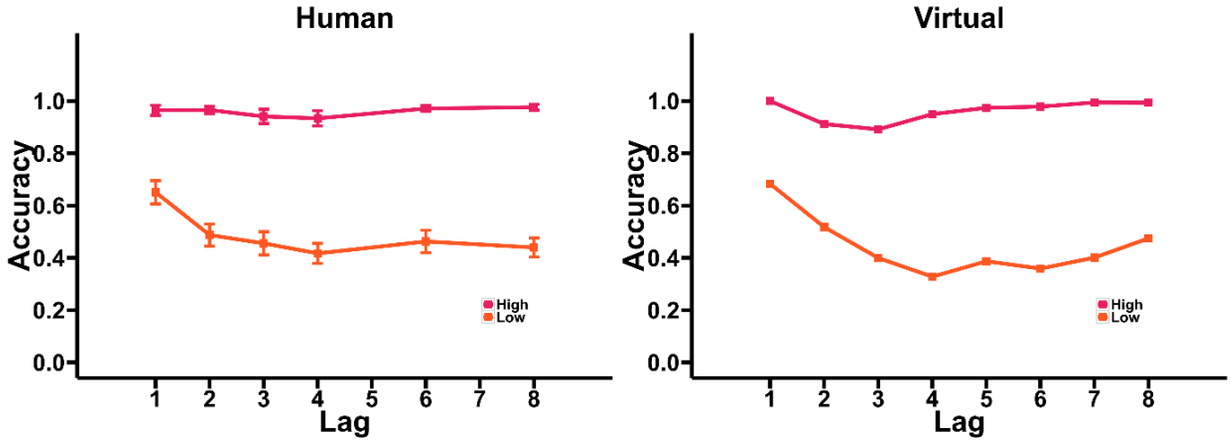


Figure 60) Mean average accuracy for high visibility trials vs mean average accuracy for low visibility trials in human behavioural data. Mean average accuracy for high visibility trials vs mean average accuracy for low visibility trials in virtual data generated by the STST model.

Meta-Experience

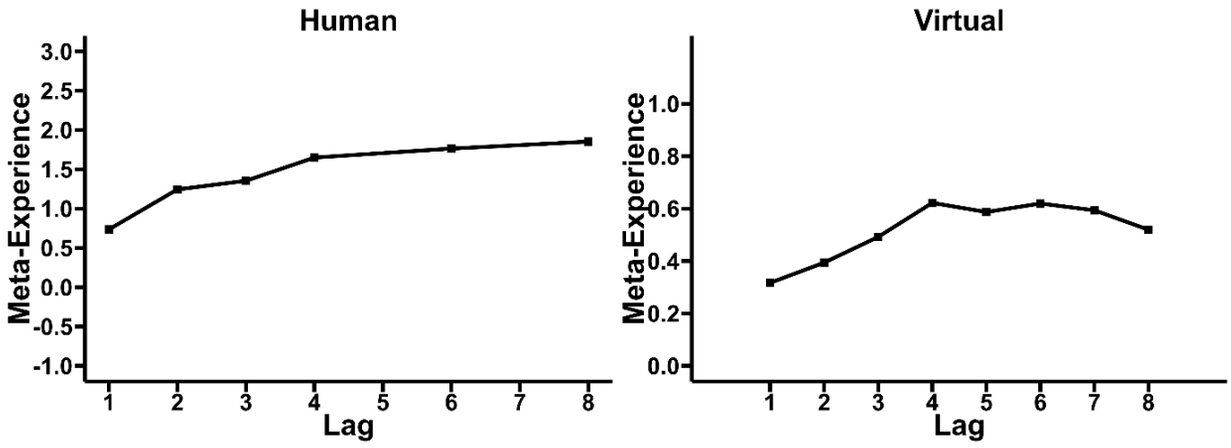


Figure 61) Comparison of Meta-Experience measures for the human data and virtual data.

Discussion

Mutual Information

Our proposed mutual information measure seems to have performed well. As a first test by which to examine our mutual information measure, it was reassuring to see that mutual information evaluates to be higher on average for high visibility trials than low visibility trials with strong significance $p < 0.001$, with a model that explained a large amount of the data's variance (49%). This showed our proposed mutual information measure is behaving as we would expect, and that our meta-experiential measure comparing high and low visibility trials was likely to produce a sensible output.

With respect to the effect of lag on mutual information, a priori, we had a strong expectation that our mutual information results would correspond quite closely to accuracy results – since improved report accuracy typically leads to improved mutual information. In terms of statistical models, we would likely have expected the effect of lag on mutual information to be statistically significantly different from the null model (since the effect of lag on accuracy is found to be significant in (Pincham, Bowman et al. 2016). Statistically, our expectations are not borne out, with Lag only approaching significance ($p < 0.05441$), and the lag model only explaining 15% of the variance in mutual information, though we do note our statistical methodology is somewhat different to the approach in (Pincham, Bowman et al. 2016).

In terms of a qualitative pattern however, in mutual information, we see exactly the correspondence with the accuracy results we expected, in a blink of bits that strongly resembles the classical attentional blink of accuracy (Figure 54). In fact, this similarity goes even further than we might have expected, as the direct z-scored comparison between the two shows. Interestingly, despite the high similarity for the total MI and total accuracy, we observe quite a noticeable difference between MI and accuracy for the high and low visibilities separately (Figure 56 vs Figure 57). We propose this is the result of accuracy being subject to a ceiling effect, whereas mutual information is not limited in the same way.

A Meta-Experiential Blink

Although our meta-experience cannot strictly be called “metacognition” because our subjective measure is about the vividness of a percept instead of confidence (Fleming, Lau 2014), in the context of our research question, it is essentially providing the same function: quantifying the coupling between our Type 1 performance, and subjective report. Examining our results, we find that meta-experience decreases monotonically with decreasing lag all the way down to lag 1, at which it shows a particularly sharp turn downwards. This suggests the general meta-experiential failure we proposed: regardless of accuracy, performance on the second target during the attentional blink, there is a generalised and increasing failure to assess this performance as the two targets get closer together.

Our generalised sight-blind recall hypothesis would predict that this is the result of participants systematically reporting poor visibility despite high Type 1 performance. However, meta-experiential results themselves are actually consistent with a range of different patterns of behaviour – they themselves only entail that there is a difference between MI and high and low visiblities. Purely on the basis of this meta-experiential measure it could, for example, be that this difference is the result of (1) correct trials being reported with low visibility while incorrect trial visibility remains constant, (2) incorrect trials being reported with high visibility while correct visibility remains constant, or (3) some mixture of the two.

Our sight-blind recall hypothesis corresponds to the first of these and, if correct, entails that mutual information for high visibility should increase at a lower rate

than mutual information for low visibility as lag decreases. We would also expect a similar effect for report accuracy, with report accuracy for high visibility increasing at a slower rate than report accuracy for low visibility as lag decreases. Effectively, another phrasing of this is that we would expect the average visibility for correct trials to decrease as lag decreases, while the average visibility for incorrect trials would decrease at a slower rate or not at all. Both these last two formulations are informative despite being very similar; one provides an interesting contrast to our meta-experience measure, while the other is more interpretable in the context of our putative sight-blind recall findings.

All of these effects can be seen vividly in Figure 59, Figure 60, and Figure 61. Notably, it is compelling that the visibility for incorrect trials consistently stays at a low but non-zero value across lags, but the average visibility for correct trials decreases all the way down to lag 1. Conversely, we see accuracy performance for high visibility trials at ceiling, while the accuracy performance for low visibility trials increases. Interestingly, despite the similarities in behaviour between report accuracy and mutual information, their respective behaviours for high and low visibility are not identical (Figure 56 vs. Figure 57). This is particularly interesting in light of our previous discussion of the similarities between the overall mutual information and accuracy behaviour. We propose that this is the result of the ceiling effect which limits report accuracy, but does not necessarily limit mutual information in the same way.

Odd perception, not poor perception

Previously, when discussing our state-trace analysis, we have discussed order errors as a potential confound for our results. While we have discussed why this is unlikely to be a confound in our case, a potential criticism that is pertinent to our current results and highly related is that participants may be reporting low subjective visibility of the second target not because they perceived it poorly, but because their percept was odd or unusual in some way. In this case, it is likely that participants may “err on the side of caution", and simply report low visibility.

To some extent, this is exactly the kind of confound we attempt to avoid by using SDT based approaches, and the same goes for the extended method we discuss here. Further, one thing that stands against this in general terms, is the low rate of order errors in this data. If one is receiving confusing percepts, we would

expect this to be reflected in reports of accuracy as well, be it in complete failures of report in which targets are not correctly identified, or partial failures in which targets are bound to the wrong context. In our data, we see neither of these things.

SE/SE model predictions

As well as analysing meta-experience, an important contribution of this chapter was comparing how the predictions made for the measures proposed in this chapter based on the (noisy) SESE model correspond to the human data. Despite no changes having been made to the model, its predictions correspond well to the human data though we note the trend, as previously, of the model tending to predict more extreme behaviour than the human data ends up reflecting (Figure 59, Figure 60 and Figure 61, virtual data vs. human data).

One outlier in terms of behaviour is that the models' mean accuracy for high and low visibility conditions does not match the human data quite so closely (Figure 60). However, in this case, it is interesting to also compare the model to not just raw accuracy, but to mutual information by lag Figure 56. We find that the model more closely matches our human MI than it does the human accuracy. In retrospect this likely occurs because accuracy performance for high visibility trials in the human data is at ceiling, while the model, by design, is not. Since mutual information is strongly related to accuracy, but also not subject to a ceiling effect in the same way, it is perhaps not surprising that it matches the behaviour of the model better. Despite our model predictions not quite perfectly predicting the human data, the “meta-experience" predictor that results from it still matches the human data quite closely (Figure 61). Perhaps the only notable difference is that the model does not predict such a sharp down turn at lag 1 as is seen in the human data. Overall, considering that we made no modifications to the existing model in order to accommodate these new measures, it is highly encouraging that it still robustly predicts these new measures.

Conclusion & Future Work

In this chapter, we attempted to validate an implicit assumption that we had been making throughout this whole thesis – that the average behaviour of our experiments was approximately reflected in the trial-by trial data too. While there existed several valuable methods for quantifying this kind of correspondence,

none were suitable for our data and so we created our own method for evaluating this. Furthermore, through these experiments, we noted an opportunity to validate the model we had created in the previous chapter.

Our results based on our new meta-experience measure showed that our previous implicit assumption that the averaged behaviour of our data is representative of the trial-by-trial behaviour in the data was not completely correct, but in a direction that further supported our previous conclusions. We found a generalised meta-experiential failure at early lags, especially at lag 1, stemming from correct trials being reported with a lower average subjective experience. In terms of validating our model, we found that without any further changes, our model predicted the results from this chapter as well as it had predicted those from the previous one.

8. Discussion and Future Direction

Over the course of this thesis we have explored the possibility of a dissociation between working memory encoding and subjective report. Here, we review the results of the previous chapters and the contributions that this thesis makes.

A Dissociation of Working Memory Encoding and Subjective Experience

Contributions of Thesis

The core research question of this thesis was to determine the relationship between working memory encoding and subjective experience, and what, if any, relationship of dependency exists between them. In the literature review, we identified the experiential blink (Pincham, Bowman et al. 2016) as an appropriate paradigm over which we might assess this question, noting both the contrasting behaviour of report accuracy and subjective report, and the possibility of a sight-blind recall effect described by (Pincham, Bowman et al. 2016). We also examined various tools by which we might identify and quantify such a dissociation, and identified state-trace analysis as the most suitable method.

Our first analysis applied this state-trace method to the data from (Pincham, Bowman et al. 2016) in order to quantify whether the difference in behaviour between report accuracy and subjective report was sufficient to conclude that working memory encoding and subjective experience are dissociated from one another. From the state-trace analysis, we found strong evidence that a dissociation did indeed exist. To inform ourselves about how this dissociation has manifested, we recalled the putative sight-blind recall effect described by (Pincham, Bowman et al. 2016) who identified a report accuracy that was very high even when minimal subjective experience was reported. Though not conclusive for the reasons we discussed in our literature review, this suggests that working memory encoding may be a necessary but not sufficient condition for subjective experience. Though we could not definitively decide this with our initial state-trace analysis, we performed a “post hoc” analysis that could identify which lags contributed the most evidence for the dissociation. The finding that the dissociation was largest at Lag 1 provides further evidence that working memory encoding is a necessary but insufficient condition for subjective

experience. We found this to be the case, as well as that removing both lags 1 and 2 together not only eliminated evidence for non-monotonicity, but provided positive evidence for monotonicity. We also note that removing lags 2 and 3 individually only contributed to reducing non-monotonicity to a small extent.

One of the challenges that this analysis provided was setting a suitable prior for the Bayesian method upon which the statistical quantification for state-trace analysis is based. Unlike previous state-trace analyses that had been in a strong position to make ordinal assertions about their variables, the behaviour of the attentional blink proved more difficult to quantify. In the analysis above, we selected a prior from previous literature (Nieuwenhuis, de Kleijn 2011), and also followed previous literature in validating this prior using an independent contrast to our question of interest that quantified how well this prior was reflected in the data. While this was an acceptable solution to the immediate problem of setting a prior in this analysis, this difficulty motivated us to search for a better method of setting a prior in similar situations, both for the above analysis and any similar analysis in future.

We achieved this by creating a method that converged to a new prior. We converged on this prior by using the same contrast we used in the previous chapter to validate our belief in the prior we had chosen from the literature. Theoretically, the method bears a resemblance to a simplified version of parametric empirical Bayes. We validated this method using simulated data before rerunning the analysis from the previous chapter using the new prior. This analysis reinforced the conclusions we had come to in the previous chapter, replicating the same results but with larger effects.

One of the goals of our research questions was not just to explore the relationship between working memory encoding and subjective experience, but to provide, if possible, a working model of any relationship we found. One theory that we put forward was that working memory encoding of multiple targets could occur simultaneously, but that subjective experience of the two targets may only occur in serial. In order to validate this hypothesis, we decided to embody it in a computational model, whose behaviour could be quantitatively compared to the results from our human data. We discussed several possible avenues we might take to achieve this, but in the end we opted to build on top of an existing model.

Our model of choice was the Simultaneous Type/Serial Token model of attention. Not only did it model attention in the relevant context (The Attentional Blink), but it naturally dealt with the simultaneity/seriality dichotomy we posed here, as well as providing a readily available computational implementation that was capable of producing not just behavioural results, but virtual ERPs.

To make our model the fairest examination of the hypothesis possible, we refrained from changing any model parameters of the existing STST model, only slightly modifying the input data stream to the model to be more appropriate, in a way consistent with how the model has been adapted to tasks in previous literature (Craston, Wyble et al. 2009). Instead, we built a new readout on top of the existing model. We compared the behavioural and electrophysiological results of this model to those from the human data examined in the previous chapters and found that there was a strong match between the two. We did observe that the human data best matched some combination of corresponding and subsequent lag from the model, but noted that this was likely to be a fixed timing offset. Interestingly, in terms of our overall research question, we found that the model as set up was consistent with both our existing hypothesis of working memory encoding in the absence of subjective report, but also subjective report in the absence of working memory encoding – though this would be likely to be a small effect.

A problem that was not completely solved by any of our analyses up to this point was that the dissociation we find from our state-trace analysis was strictly non-constructive. From our state-trace analysis, we only know that working memory encoding and subjective experience are not mutually dependent upon one another, and we must rely on other sources of information to determine any relationship beyond this. On the basis of average behaviour, the sight-blind recall effect noted by (Pincham, Bowman et al. 2016), and our post-hoc state-trace analysis, we concluded that it is possible that working memory encoding is a necessary, but not a sufficient condition for subjective experience. However, the strength of these conclusions rests on a tacit assumption that the average behaviour we have observed is also reflected on each individual trial. We described instances in which it may be possible for the general trends of our data to match up to our results but for our existing conclusions to be erroneous,

though we did reflect that such behaviour would be extremely unlikely. In order to resolve this issue we therefore attempted to quantify a measure of *discriminability*, capable of deciding this by determining the parity between our two measures. There is strong precedent for such a measure, which has been employed extensively in the field of metacognition with Signal Detection Theory. Unfortunately for us, these SDT based approaches are only applicable for binary detection tasks, while our task was a 21-way identification task. In light of this, we developed our own approach for calculating discriminability that captures the intent and spirit of the original measure but that is more widely applicable, by using Mutual Information.

This measure of Mutual Information allowed us to calculate a measure of Type 1 discriminability, which we were then able to compare for high and low visibility trials to quantify meta-experience. Though challenging to implement, this measure showed a decrease in the discriminability of correct and incorrect trials through subjective report as lag decreases. This indicated one of three situations – at earlier lags, trials were being reported with high accuracy despite low subjective report, trials were being reported with low accuracy despite good subjective report, or both. Examining the data, we determined that this was almost exclusively the first case, of high accuracy trials being rated with lower visibility at early lags. This provided evidence that, (1) our assumption that the average behaviour was reflected in individual trials was not quite supported, but that (2) our hypothesis that working memory encoding as a necessary but insufficient condition for subjective experience was correct.

A potential overarching limitation of the thesis is that it occurs over just one dataset, and a potentially slightly odd one at that – other works collecting subjective report during the attentional blink tends to observe at least some lag 1 sparing of visibility and we observe none. To solve this problem, we have replicated many of our findings over a novel dataset. The results of this can be found in Appendix E.

Limitations

A central limitation to this work is the restrictions of the state-trace analysis that we perform in the first chapter. As we have discussed at length, with state-trace analysis, we can only distinguish between “a dissociation exists” and “there is

insufficient evidence that a dissociation exists". Indeed, demonstration of monotonicity does not definitively preclude a dissociation. In a sense, much of the subsequent work we perform on later chapters is attempting to make up for this limitation by elucidating potential explanations for this dissociation.

As we discuss extensively in our literature review, the concept of phenomenological consciousness overflowing access consciousness is a topic that has been heavily debated in the literature. Though we must be careful not to over interpret our results in this context, as it is not entirely clear how our concrete measures translate to the abstract concepts, it is interesting that our results seem to point toward the dual hypothesis – access without awareness.

In terms of our specific state-trace analysis, a limitation was the setting of the Bayesian prior upon which the analysis is based. We attempted to provide an alternative method that provides a prior that better fits the data as a whole, but this method relied upon us being able to set a single prior that is suitable for all participants at the same time. While this method behaves well and is computationally cheap, it is likely that better results could be achieved with more sophisticated methodology.

While our seriality of experience hypothesis was supported well by our modelling, it will need further validation. The simplicity of the hypothesis is a strength but is is also potentially a limitation. We also note that while our choice to build on the STST model was beneficial in many respects, the mechanism of the blaster in the original STST model somewhat limits our model of subjective report by placing a hard, predetermined limitation on the maximum length of P3.

In the end, we opted for the Simultaneous Type/Serial Token model, but our hypothesis would have also worked well inside the neuronal global workspace model (Dehaene, Kerszberg et al. 1998). Indeed, the competition between stimuli that the global workspace puts forward effectively creates seriality. Had we opted for this model, our challenge would have been the formulation of a mechanism by which an attentional blink occurs that affects report accuracy without affecting subjective visibility.

Our meta-experience measure was excellent in theory, and because of the steps we took to secure our results against small sample sizes, we remain confident

that its conclusions are robust. However, there is no denying that despite the extensive steps we took in order to mitigate this effect as far as possible, the process was hampered by the quantity of data available in our dataset. With a bigger dataset, we could get a much finer grained picture of the accuracy/visibility correspondence.

While our meta-experience measure was good at quantifying our report accuracy/subjective report correspondence, it is difficult to interpret the results in the context of the wider metacognition literature because we collected reports of subjective experience instead of the confidence measures typically used. Hence, we call our measure meta-experience. Since our mutual information gives a measure of discriminability, examining the difference between this measure at high and low visibility is a reasonable way of quantifying metacognition (or meta-experience in our case). However, it is very possible that further exploration may yield better alternative measures.

Future Directions

Here, we follow previous authors in using what (Davis-Stober, Morey et al. 2016) refer to as dependent variable state-trace analysis that is the current state-of-the-art (Davis-Stober, Morey et al. 2016). However, we note that the original state-trace analysis as it was conceived by (Bamber 1979) was much broader in scope than this. The original framework was conceived as a way of differentiating many different possible 3 stage models of the data, instead of the simplified binary monotonic or non-monotonic choice we make at the moment. Unfortunately, the problem with this original conception was that it was extremely difficult to provide statistical quantification for which model was most appropriate. The dependent variable state trace analysis we perform that compared monotonicity and non-monotonicity has a significant advantage in providing a clear criterion to distinguish the models that is directly amenable to statistical methods; it is understandable that it is so widely used. However, it would be an interesting direction for future research to explore further how possible it is to go further than this simple monotonic/non-monotonic distinction, and explore more detailed models of the data that could additionally evaluate, for example, the relationship of dependency between the two variables.

Our improvements to the current state-trace methodology are clearly beneficial, and have the advantage of being extremely simple to calculate, but also suffer from this same simplicity. Notably, they are currently not flexible enough to permit different models for each participant. A fruitful direction for further research in this area may be to explore further more complicated models that make use of a parametric empirical Bayes approach or even goes so far as to use a full multileveled Bayes design by setting a hyper prior on the prior distribution of the data. This would allow a more fine grained specification of priors, as well as different models for each participant. Such research would have to be approached carefully in order to keep the methods practically applicable and from exploding in computational complexity, but could prove a significant advance. Another more immediately implementable research direction for the method we outline in this thesis specifically would be to borrow from the convex hull solution that the Aggregated Bayes Factor makes use of as another way of eliminating priors that are only suitable for a selection of participants.

While our simultaneous encoding/serial experience model of subjective experience and working memory encoding successfully explained our data, the most immediate need for future research in this area is a validation of the hypothesis with further datasets. Implementing these simultaneous encoding/serial experience findings in the eSTST model would be another excellent direction to take the research. Notably, the changes to the blaster may mitigate the effect we see comparing model lag 1 with human lag 1, where the difference is smaller than we would expect in the model data because P3 length is artificially constrained.

An interesting facet of sight-blind recall, if it were true, is what it implies about the function of consciousness. After all, if subjective experience isn't strictly necessary for sight-blind recall, what function does it perform? Under the seriality of experience hypothesis, one interpretation might be that consciousness is about encoding order information. The theory after all explicitly holds the second item out such as to achieve a unitary percept. In this instance, even items end up bound into the wrong temporal context (an order error), subjective experience makes it unambiguous what is in the brain – such order errors are an illusory

percept. Such a theory however, is currently speculative, but may provide an interesting direction for future work.

For our meta-experience measure, we would benefit a large amount from simply collecting more data. Given the amount of data that can be fruitfully collected from any given participant, this would likely involve collecting more trials for fewer lags and subjective visibility ratings. It would also be interesting to explore more directly how this measure matches up with classical SDT approaches. Though not so directly applicable to our research question, it may also be interesting to perform these same studies collecting confidence ratings instead of the subjective experience measures we made use of here. This would allow our measure to more directly translate into metacognition, allowing the results to be interpreted more broadly.

Final Observations

We set out to investigate the relationship between working memory encoding and subjective experience. At the start of the thesis, we laid out the following general research questions, which were addressed in the research chapters:

1) *Can working memory encoding and subjective experience be dissociated?*
   In our state-trace analysis of the attentional blink data in which both subjective report and accuracy were recorded, we found strong evidence that working memory encoding and subjective experience were dissociated.
2) *If working memory encoding and subjective experience can be dissociated, what is the relationship of dependency between these two processes, if any?*
   Based on a post-hoc analysis of our state-trace experiment that systematically excluded lags, we were able to discern that the lag 1 data point was contributing very strongly to this dissociation. Combined with the putative sight-blind recall effect observed by (Pincham, Bowman et al. 2016), this seemed to provide evidence that the dissociation was arising as a result of working memory encoding perhaps being a necessary but not sufficient condition for subjective experience. This was further reinforced by our meta-experience measure, which showed that not only

was there a “meta-experiential" failure at early lags, especially lag 1, but that this failure was arising as a result of participants systematically reporting low subjective experience on correctly identified trials.

3) *If we can find evidence for a particular relationship of dependency between working memory encoding and subjective experience, can we provide a model of it?*

   Based on the results of our state-trace analysis, we developed a theory to account for our results through the simultaneous encoding/serial experience hypothesis. We developed this hypothesis into a model built on top of the existing simultaneous type/serial token model of attention. We validated the behavioural and electrophysiological predictions of this model against results from human data, and found them to be a good match.

In summary, we have investigated the relationship between working memory encoding and subjective experience. We have demonstrated that the two are dissociated in the attentional blink, examined what relationship of dependency the two measures have in light of this and explored the results in the context of a metacognitive failure. We have also provided a model of this relationship during the AB. While a full understanding of the relationship between these two measures remains a broad topic for further research, in this thesis we have made small steps toward a more complete understanding of how these two similar processes can be related.

Appendix Material

Appendix A – Detailed Experimental Procedure

Participants

Initially, twenty-one young adults took part in the study. One participant was removed due to an inability to achieve 50% accuracy for T1. Two more participants were removed because more than 50% of the epochs extracted from their EEG data were rejected through the artifact detection criteria. Data from 18 participants (15 females) were therefore analysed. Participants were 19-28 years old (mean age: 21.67 years, SD 2.93 years). Participants provided informed, written consent, had normal or corrected-to-normal vision and were fluent in English. The study was approved by the Psychology Research Ethics Committee at the University of Cambridge, UK.

Stimuli and Procedure

Stimuli were presented on a Sony Graphics Display CRT monitor with a 100Hz refresh rate. Targets were the uppercase letters excluding I, M, O, Q, W. These letters were excluded because of their physical similarity to digits (I, O and Q) or because their physical size meant that they were not adequately masked by digits (M and W). Each trial contained one or two targets – T1 occurred on every trial and was always presented in red, and T2 (if it occurred) was presented in white. Distractors were single digits excluding 0 and 1, presented in white. The rationale for presenting T1 in red and all other items in white was so that the visibility question (that is, "How visible was the white letter?") would clearly refer to T2 and not T1. All alphanumeric stimuli appeared on a black screen. Stimuli subtended visual angles of 3.8˚ vertically and 2.9˚ horizontally, assuming a viewing distance of 57cm. On each trial, a fixation cue (a cross shape subtending 2˚×2˚) was presented in the centre of the monitor for 200msec. The RSVP stream began 1000msec after the onset of the fixation cross. Each RSVP stream contained 15 items that were presented one after the other in the centre of the monitor. The identities of the target letters and the digit distractors were randomly assigned on each trial with the restriction that successive items were not the same.

Distractors were presented for 90msec with no ISI. T1 randomly appeared as the fourth, fifth or sixth item in the RSVP stream.

At the end of each RSVP stream, participants were asked to rate the subjective visibility of T2 using a self-report scale: “On a scale of 1-6, please indicate how well you have seen the white letter." The numbers 1 2 3 4 5 6 were presented in a horizontal line on the screen, with the description “not seen" presented beneath the number 1 and the description “maximal visibility" presented beneath the number 6. Participants used the number keys (1-6) on the keyboard to indicate their subjective visibility ratings. Participants then reported the identity of T1 and T2 (even if a second target did not occur) using the keyboard letter keys. Participants were required to guess if they were unsure of the target identities.

All participants completed two experiments, spaced at least one week apart. Experiment 1 exclusively collected behavioural data and Experiment 2 collected both behavioural and EEG data. Experiment 1 consisted of four blocks, each with a different target/mask duration combination. The mask, if it occurred, was always the hash (#) symbol. In Block 1, targets appeared for 90msec with no mask. In Blocks 2, 3 and 4, the target/mask durations were 70msec/20msec, 60msec/30msec and 50msec/40msec respectively. In Experiment 1, T2 appeared at lags 1, 2, 3, 4, 6 or 8 with equal frequency. Experiment 1 deliberately sampled a large number of lags in order to examine the relationship between T2 accuracy and subjective visibility across the entire AB curve. Trials that did not present a second target (no-T2 trials) were also included with equal frequency (hence, one in seven trials did not contain a second target). Experiment 1 contained 4 blocks of 49 trials, totalling 196 experimental trials.

For each participant, data from Experiment 1 were analysed to determine which of the four target/mask durations resulted in T2 being correctly reported on approximately 50% of lag 3 trials. Each participant's optimal target/mask duration was then employed in Experiment 2. As a result, 28% of participants received the 70msec/20msec target/mask duration in Experiment 2, 50% of participants received the 60msec/30msec duration and the remaining participants received the 50msec/40msec duration. Experiment 2 contained 5 blocks of 100 trials, totalling 500 trials. To maximise ERP signal strength in

Experiment 2, T2 appeared at lag 1 on 200 trials, at lag 3 on 200 trials, at lag 6 on 50 trials and was absent on 50 trials.

In Experiments 1 and 2, a distractor appeared in the place of T2 on no-T2 trials. However, the experimental program still assigned a target identity to T2, and participants were asked to report the subjective visibility and identity of T2 – even when a second target did not appear. In this manner, T2 'accuracy' on no-T2 trials (trials where T2 was correctly guessed by chance) could be calculated. The no-T2 trials were included for two reasons. First, subjective visibility for T2 could be determined for trials where the second target was not present. It was therefore possible to confirm that participants were accurately using the visibility scale, because subjective visibility should be very low on trials where T2 did not occur. Second, T2 report accuracy on no-T2 trials could be calculated and compared with theoretical (chance) levels of correct T2 report accuracy.

The order of the trials within each block was randomised. Participants could take short breaks between blocks. Testing occurred individually in an acoustically and electrically shielded booth.

EEG Acquisition and Pre-processing

EEG was recorded using the Electrical Geodesics Inc. system and a 129-channel hydrocel geodesic sensor net. The sampling rate was 500Hz. An anti-aliasing lowpass filter of 100Hz was applied during data acquisition. Offline, the data were bandpass filtered between 0.01–30Hz and recomputed to an average reference. The continuous EEG was segmented into epochs between -200 to 1000msec relative to the onset of T1. Spline interpolation was carried out on individual channels if required. The mean percentage of interpolated channels was 4.60% (range: 0–8.59%). Epoched data were smoothed using a 50msec Gaussian filter.

Epochs were excluded from analysis if they met any of the following artifact rejection criteria: voltage deviations exceeded ± 100μV relative to baseline, the maximum gradient exceeded 50μV, or activity was lower than 0.5μV. Across participants, 78.02% of trials were retained after artifact rejection.

Appendix B – Changelog to State-Trace Code Provided By Davis-Stober et al.

This appendix lists the changes we made to the code provided by (Davis-Stober, Morey et al. 2016) for our analysis, with a brief discussion of why the changes are required in each case.

CHANGE[1]

Added drop=F when sub-setting trace constraints with dimension constraints. This prevents a case in which matrix(order(x),ncol=ndim)[,D.order] is a dimensionless vector if trace has only one dim

CHANGE[2]

nolap defaults to TRUE in all cases if trace has only one dim. Added a check to enforce the opposite.

CHANGE[3]

Trace vs non-trace is meaningless with only one trace dim. However, we can apply the same logic used with trace vs nontrace to our dimension constraints.

The changes mean that (hopefully) this new measure (d.nd reflects the strength of our belief in the dimension constraints. This was achieved by removing nolap (undefined for one trace dimension) and setting prior/post nontrace == 1

CHANGE[4]

Added the ability to set different constraints on each axis of the state-trace. This was achieved my defining two order vectors, D.orderX and D.orderY.

It necessitates significant changes to the way the prior and post trace+dim are calculated. If the constraints on each axis are the different then we can no longer, for example, calculate the monotonic probability as the sum of the diagonal joint probabilities - because our matrix is no longer square.

As a result of this, the monotonic effect is now calculated as the sum of the diagonal of the joint order probabilities that corresponds to the intersection of the two constraints, that is the all the ordering which are the same between the two. The non-monotonic effect is calculated as the sum of all probabilities in the joint

order probabilities that fit to the constraints, minus the sum of the diagonal of this union.

NB: the labels orderX and orderY may be a little confusing. To be explicit, orderX indexes the jointOrderProb matrix rows(that correspond to accuracy in our experiment) and orderY indexes jointOrderProb matrix columns (that correspond to visibility in our experiment)

CHANGE[5]

Added the ability to specify multiple constraints for each axis. Constraints must be the same length Dx.c or Dy.c, and the number of such constraints must be specified as Dx.r or Dy.r.

This is perhaps a little clumsy, but serves its purpose as any set of constraints can be specified in a series of, e.g. pairs.

For simplicity, it is also requires that dim.increasing is manually specified by the user. Currently, dim.increasing is assumed to be the same for each axis, but this should not be difficult to work around.

Minor changes:

Added a BF3, which is the product of all BF2.n.m. Useful as a yardstick of total evidence

Added a BF4, which is the product of all BF2.d.nd. Useful as a yardstick of total belief in the constraints

m.nl measure removed as it was causing problems with only one trace dimension

d.nd is now a part of BF2

Appendix C – Changelog to Neural STST

This appendix lists the changes made to the original STST code for the purposes of our simultaneous encoding/serial experience model.

---

bigbattery.m

Added tracking for full excitatory postsynaptic potential by neuron (as opposed to collapsed across all neurons)

global PresynapHistory ExPostsynHistory InhibPostsynHistory MembPotHistory

->

global PresynapHistory ExPostsynHistory InhibPostsynHistory MembPotHistory ExPostsynFull

save(erpfile,'MembPotBat_*', 'PresynapBat_*', 'ExPostsynBat_*','InhibPostsynBat_*','BasicAccu');

->

save(erpfile,'MembPotBat_*', 'PresynapBat_*', 'ExPostsynBat_*','InhibPostsynBat_*','BasicAccu', 'ExPostsynFull_*' , '-v7.3');

%excitatory postsynaptic output

ExPostsynBat_basic(trialcounter,:,:,:) = ExPostsynHistory;

->

%excitatory postsynaptic output

ExPostsynBat_basic(trialcounter,:,:,:) = ExPostsynHistory;

%excitatory postsynaptic output by neuron

ExPostsynFull_basic(trialcounter,:,:,:,:) = ExPostsynFull;

ExPostsynBat_basic = zeros(numTcombi,NUMSTREAMS,runlength,NUMLAYERS);

->

ExPostsynBat_basic = zeros(numTcombi,NUMSTREAMS,runlength,NUMLAYERS);

ExPostsynFull_basic = zeros(numTcombi,NUMSTREAMS,runlength,NUMLAYERS, 40);

---

modified stimulus strength variability

resolution = .012;

->

resolution = .0125;

varstart = floor(-.078/resolution)

varstop = floor(.078/resolution)

->

varstart = floor(0.1625/resolution)

varstop = floor(0.1625/resolution)

---

modified stimulus strength basevalue

baseval = .526;

->

baseval = .570;

----

runRSVP.m

Added ExPostsynFull as a global variable & initialised it

global History MPHistory OutHistory InHistory BiasHistory HebbHistory PresynapHistory ExPostsynHistory InhibPostsynHistory MembPotHistory

->

global History MPHistory OutHistory InHistory BiasHistory HebbHistory PresynapHistory ExPostsynHistory InhibPostsynHistory MembPotHistory ExPostsynFull

InhibPostsynHistory(lag,step,:) = TempInhib';

->

InhibPostsynHistory(lag,step,:) = TempInhib';

ExPostsynFull(lag,step,:,:) = ExPostsynAct;

----

st2data2eeglab.m

Updated so that upstream changes in bigbattery work with new format

Appendix D – Other Entropy Estimators

In the main chapter, we only present a few of the entropy estimation algorithms used in order to keep the discussion focused. Here, we examine the performance of several other contemporary algorithms that we might have used.

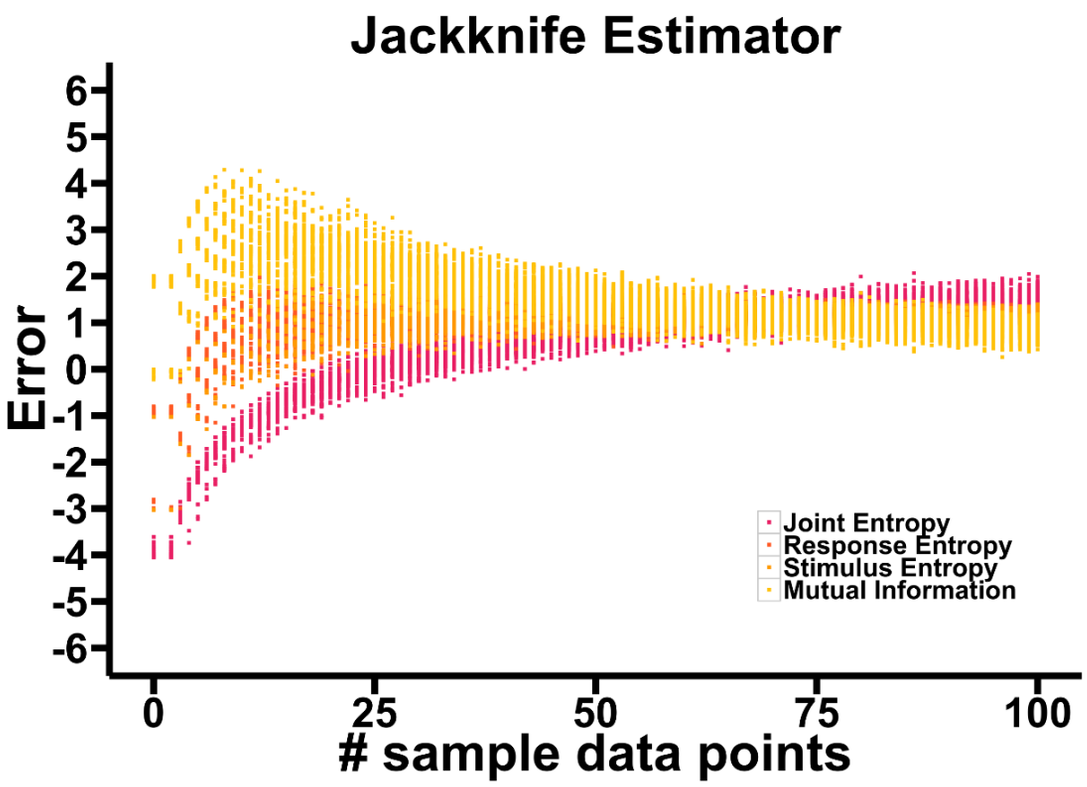


Figure 62) Error of the Jackknife entropy estimator described by (Paninski 2003) of stimulus, response and joint entropy, and mutual information for a range of different sample sizes from 1 to 100.

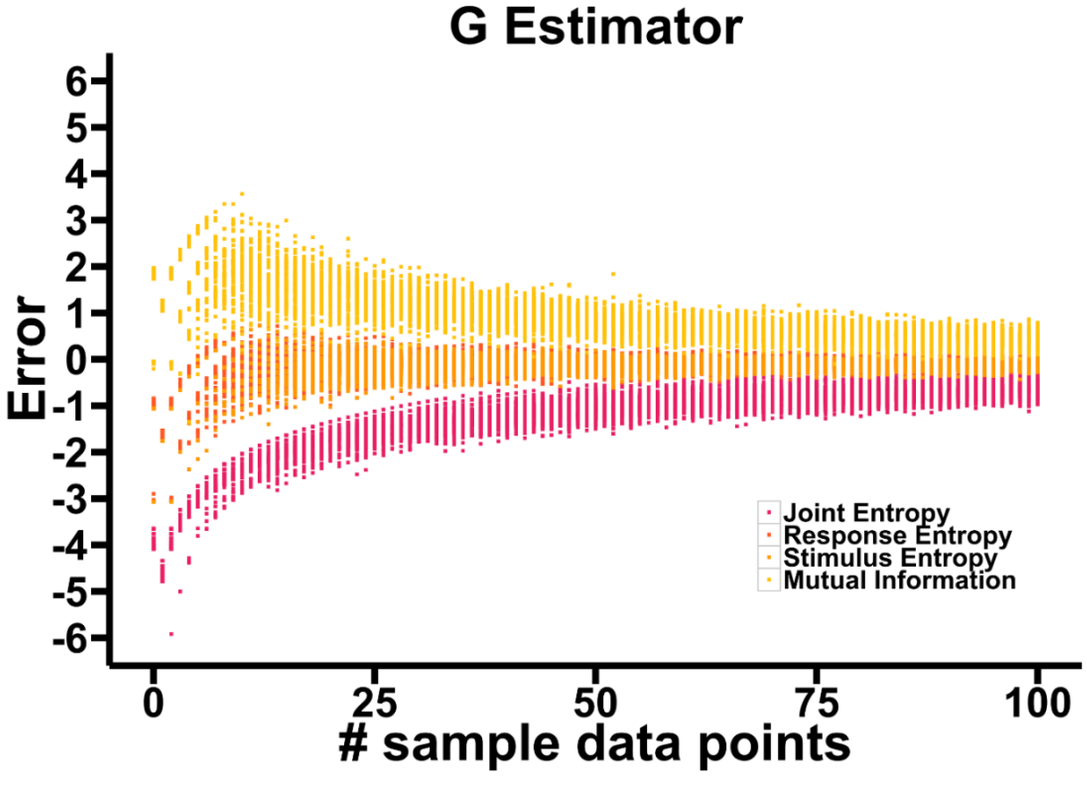


Figure 63) Error of the G estimator from (Grassberger 2003) of stimulus, response and join entropy, and mutual information for a range of different sample sizes from 1 to 100.

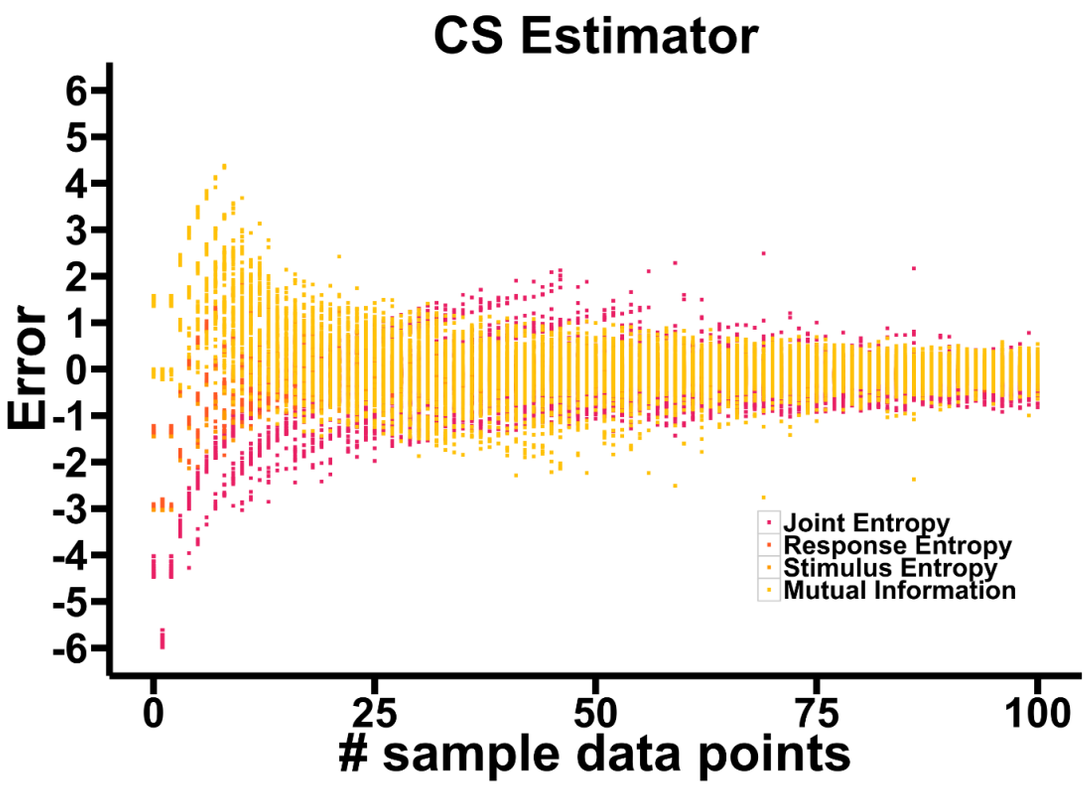


Figure 64) Error of the CS estimator from (Chao, Shen 2003) of stimulus, response and join entropy, and mutual information for a range of different sample sizes from 1 to 100.

Appendix E – Replication of results with an additional dataset

Further work has replicated many of the findings in this thesis with an additional dataset. We present these results here. Note that this section is largely separate from the main body of work and rather extensive, and as such has its own bibliography.

Fleeting Perceptual Experience and the Possibility of Recalling Without Seeing

William Jones[a,1], **Hannah Pincham**[b], **Ellis Luise Gootjes-Dreesbach**[d], **and Howard Bowman**[a, c]

[a]Centre for Cognitive Neuroscience and Cognitive Systems, University of Kent, Canterbury, UK
[b]South Eastern Sydney Local Health District, NSW, Australia
[c]Department of Psychology, University of Birmingham, Birmingham, UK
[d]Point Estimate Limited, Ellesmere Port, Cheshire, UK

We explore an intensely debated problem in neuroscience, psychology and philosophy: the degree to which the "phenomenological consciousness" of the experience of a stimulus is separable from the "access consciousness" of its reportability. Specifically, it has been proposed that these two measures are dissociated from one another in one, or both directions. However, even if it was agreed that reportability and experience were doubly dissociated, the limits of dissociation logic mean we would not be able to conclusively separate the cognitive processes underlying the two. We take advantage of computational modelling and recent advances in state-trace analysis to assess this dissociation in an attentional/experiential blink paradigm. These advances in state-trace analysis make use of Bayesian statistics to quantify the evidence for and against a dissociation. Further evidence is obtained by linking our finding to a prominent model of the attentional blink – the Simultaneous Type/Serial Token model. Our results show evidence for a dissociation between experience and reportability, whereby participants appear able to encode stimuli into working memory with little, if any, conscious experience of them. This raises the possibility of a phenomenon that might be called sight-blind recall, which we discuss in the context of the current experience/reportability debate.

Introduction

The ability to seperate functionally independent mental processes, and to be able to describe this seperation – or lack thereof – is critical to modern cognitive neuroscience. Of these problems of independence, the distinction between the subjective experience of the character of a stimulus (the "phenomenological awareness" of it) and the ability to objectively report on it (the "access consciousness" of it) has been one that has been particularly hotly contested. Block[1] is a notable proponent of a distinction between the two, arguing that it is possible to experience stimuli without being able to access them, and thus report on that experience. The believed locus of phenomenological awareness is iconic memory, initially, on the basis of the Sperling paradigm[2], with others supporting the concept of phenomenological awareness to varying degrees on the basis of experiments on Kanizsa triangles[3], other, modified versions of the Sperling paradigm[4], and short term memory experiments[5]. However, despite this large body of supporting literature, the theory is contested; for example, Dehaene and co-workers[6] have challenged this theory on the basis of change blindness, while others have pointed out that certain changes to the Sperling paradigm seem to compromise some key results[7].

A paradigm that is well placed to shed light on this topic, and has been used previously[8] to explore the all-or-none nature of subjective experience, is the attentional blink. The attentional blink is a phenomenon seen during RSVP (Rapid Serial Visual Presentation) in which participants frequently fail to detect a second target for a short time after the presentation of an encoded first target; see T2|T1 accuracy in figure 1[9,10]. Recently, Pincham et al[11] noted that the temporal pattern of T2 visibility (which they called the experiential blink) is dissimilar to that of report accuracy (i.e. the classical attentional blink) and raised the possibility that this finding represents two distinct processes. However, having the tools to elicit dissimilar patterns of behaviour is not the same as being able to determine whether the cognitive processes that underlie them are distinct. Tackling such problems is usually performed by looking for functional dissociations. These arise when we find variables that allow us to independently modify performance on two separate tasks, providing putative evidence that the cognitive processes embodied by the tasks are in some way separate. Such dissociation logic has been widely applied, and made an important contribution to the investigation of functional independence in the mind in such diverse sub-fields as short and long term memory[12], word comprehension[13] and consciousness[14].

In the context of our question, there are many who have claimed that the experience or awareness of a

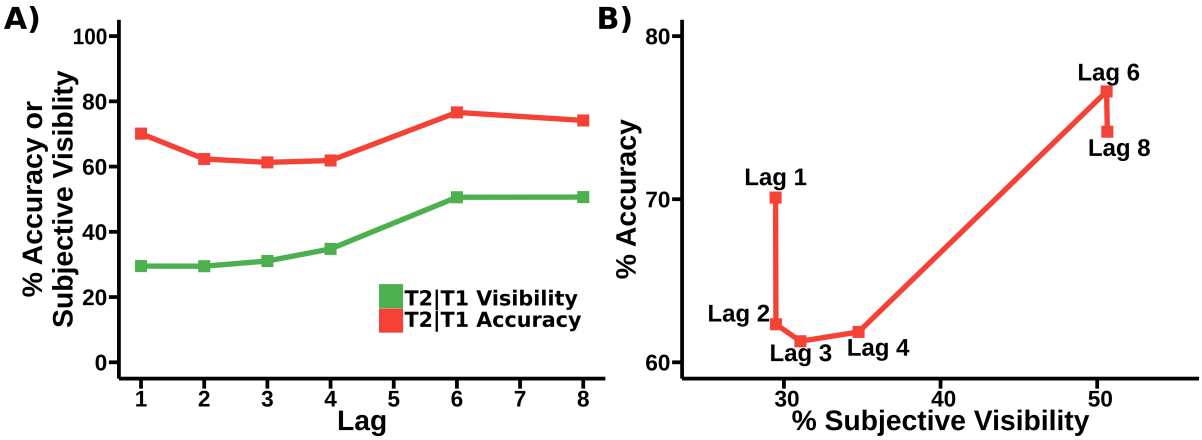


Figure 1. A) Results from[11], comparing accuracy and subjective visibility across lags in the attentional blink . The T2 visibility curve demonstrates what Pincham and Bowman term the Experiential blink of subjective report. B) State-trace plot comparing T2|T1 accuracy and T2 visibility from A). Note the apparent non-monotonicity of the relationship between accuracy and visibility. (Note, the T2|T1 blink curve here shows some very minor differences to that presented in[11]. This is because T2 accuracy in the original paper was mislabeled and in fact presented the accuracy of the conjunction of T2 and T1, whereas here we display the conditional probability of T2 given T1. None of the findings in[11] are impacted by this difference).

stimulus and its reportability are doubly dissociated. As previously discussed, in the direction of awareness without report, we have the "phenomenological consciousness" of Block. In the opposite direction, there exist several paradigms that seem to provide evidence for modulation of behaviour without awareness, for example continuous flash suppression[15], visual masking[16], blindsight[17], or episodic face recognition[18]. However, we would argue that these paradigms provide evidence for a weaker claim than reportability without awareness; that of *influence* without experience. In every case, the identity of the unexperienced stimulus is not directly reportable, it merely influences the report of, or response to, something else. In contrast, the criterion for a true demonstration of reportability without awareness would be of free recall of a stimulus identity in the absence of awareness, which, if definitively demonstrated, would be both striking and surprising.

Regardless, even if a double dissociation of the required kind between experience and reportability was widely agreed to exist, there has been a long standing debate about the use of double dissociations as a measure by which to assess functional differentiation[19–21]. In this work, we adopt an alternative method to traditional dissociation logic. This alternative suggests that a dissociation arises, given certain assumptions, when it is not possible to demonstrate a monotonic relationship between task performances. In the context of the attentional blink, there is evidence that such non-monotonicity exists between accuracy and subjective visibility report[11] (see figure 1), and one of the main contributions of this paper is to provide quantitative evidence for such an effect.

In order to provide statistical quantification, a method called state-trace analysis is typically employed. State trace analysis examines the monotonicity of data, across a state-trace plot in which our two task performances form the axes. In this work, we follow Prince, Brown and Heathcote[22] and Davis-Stober et al.[21] in advocating the use of a Bayesian approach to the analysis of these problems. The main reason for this is that we are solving a model comparison problem: comparing whether a non-monotonic or monotonic model best fits our data. Strictly speaking, a classical statistics approach would not enable us to find evidence for a non-monotonic outcome, since it would naturally take the role of the null. For a more detailed discussion on the various potential choices of statistical methods and their respective virtues, see[22].

While dissociations can tell us about specific effects, placing findings in larger theoretical context is pivotal to the forward progress of science, especially when the theory is encapsulated in a computational model. In particular, a theoretical interpretation of the data from[11] may be that items are encoded into working memory simultaneously, but only experienced serially. In combination with state-trace analysis, this allows us to explore not only the direction of the effect, but also some plausible mechanisms by which it may arise. In terms of specific models, the Simultaneous Type/Serial Token[10] model is well placed to explore this question: it models data in the relevant context (the attentional blink), and naturally deals with the difference between simultaneity and seriality.

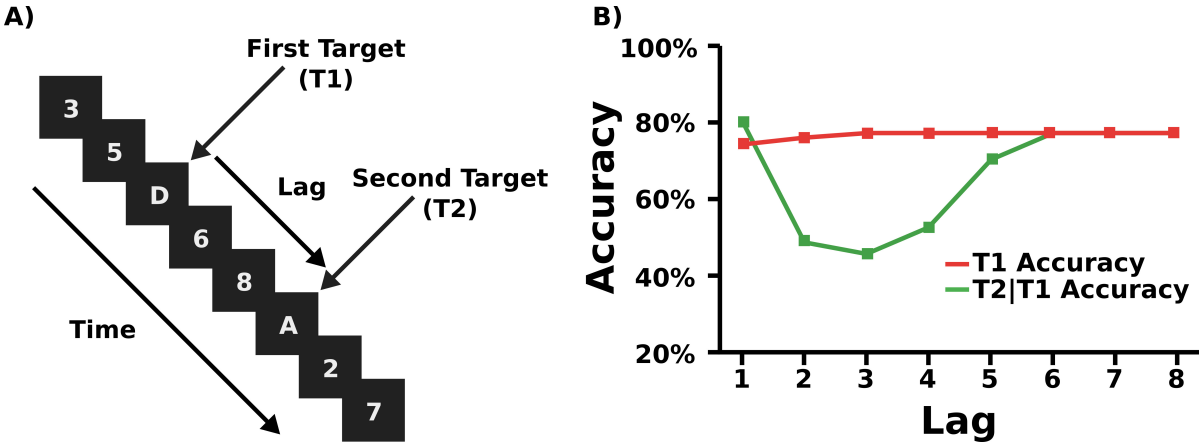


Figure 2. A) A typical attentional blink RSVP stream. Participants are instructed to report the two letters at the end of the stream. B) Example illustration of expected accuracy for T1 and T2|T1 at each lag during a typical attentional blink study with a Stimulus Onset Asynchrony (SOA), the amount of time between the onset of each stimulus, of 80-120ms.

In this paper, we make two original contributions. We first apply Bayesian state-trace analysis to the results of our attentional blink experiment in which we collected both report accuracy and subjective visibility (see figure 1), and compare the respective evidence for a monotonic and a non-monotonic relationship between the two measures. Secondly, we explore our results in the context of the Simultaneous Type/Serial Token (STST) model. Since the STST model does not natively deal with subjective experience, one of the contributions of this paper is development of a simple method by which this might be incorporated into the model. Given this method, we then compare the behavioural and EEG data that the model predicts to the human data from[11], and the results from our state-trace analysis.

The Attentional Blink Paradigm

Rapid serial visual presentation (RSVP) is a technique in which multiple stimuli are presented rapidly, one after the other in a fixed location. Typically, this stream of stimuli is composed of one or more targets to be detected or identified and a number of distractor stimuli to be ignored. The attentional blink (AB) is a deficit in performance on a second target when more than one target is to be identified[9,10]. It arises approximately 100-500ms after the presentation of the first target, when it is successfully encoded. Typically, the AB is elicited using alphanumeric stimuli, but images, letters, digits or words will all elicit the blink. For an example of a typical attentional blink RSVP stream, see figure 2.

The main parameter of the attentional blink is the relative serial positions at which the two targets are presented, known as lag, for example, at Lag 1 there are no intervening distractors between the targets, while at Lag 2, the two targets are separated by one intervening stimulus. The main attentional blink result is typically plotted as T2|T1 accuracy (second target accuracy, given the first target was correct) against lag. Excluding Lag 1, typically, when the two targets are close, accuracy is significantly reduced compared to recovery baseline (lags 7 and 8). A typical blink is shown in figure 2(B). Performance at Lag 1 is above the deepest point in the blink. This is known as Lag 1 sparing, and is itself a robust result of the attentional blink[23].

There has been extensive exploration of the attentional blink with respect to accuracy of report, but much less exploration of subjective visibility report in the attentional blink[8,11,24,25]. As we have discussed, the attentional limitations of the blink make it ideal for exploring dissociation between accuracy in reporting a stimulus and the strength of its conscious experience. Indeed,[11] mapped subjective report to lag, finding a blink of subjective experience, the so called Experiential Blink, akin to that of reportability, but without Lag 1 sparing. The results of this experiment are shown in figure 1.

Functional dissociations and reversed associations

As mentioned previously, the functional dissociation is a technique that has been widely implemented across the fields of psychology and neuroscience as a marker of the functional distinctness of mental processes. There are

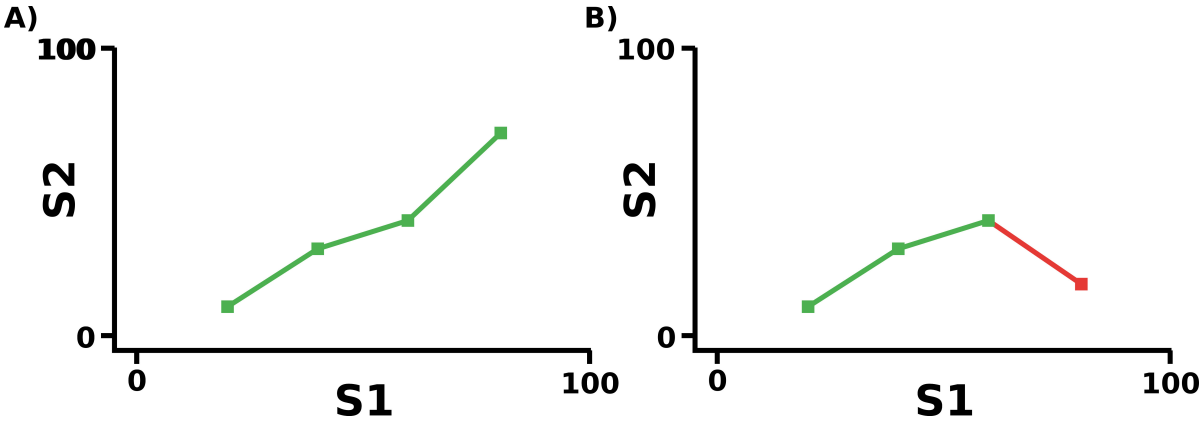


Figure 3. A) Example of a monotonic state-trace plot across 4 levels of a dimension factor D. It is possible to draw a monotonic (increasing) curve joining all points, therefore the relationship between the levels of the state factor is monotonic. B) Example of a non-monotonic state-trace plot across 4 levels of a dimension factor D. The point furthest to the right makes drawing either a monotonically increasing or monotonically decreasing curve impossible, therefore the relationship between the levels of the state factor is non-monotonic.

several types of functional dissociations, but all arise when one is able to independently modify performance on a set of one or more tasks without affecting performance on other tasks in the set. The ability to differentially affect behaviours on different tasks is seen as evidence that the mental processes underlying them are in some way functionally separate. However, despite their wide use in the literature, it has been argued that while dissociations are certainly indicative, they do not strictly provide either a necessary or sufficient basis for determining the separation of mental processes[19–21]. Broadly, it has been proposed that it is possible to construct cases in which dissociations exist but separate mental processes do not[19–21], and to create cases in which there are separate mental processes without dissociations. For an overview of these arguments, and a demonstration of how such behaviours can be constructed, see[26].

Regardless which side of this debate one stands, an alternative measure exists for which it is certain these issues will not arise: the reversed association proposed by[20]. The reversed association models the cognitive function that dissociations are trying to evaluate as a latent variable determining the relationship between a given task and task performance. It then assumes that, while the relationship between this latent cognitive function and task performance may not be proportional, it may at least be assumed to be monotonic in some direction[27]. Given this assumption of monotonicity between cognitive function and task performance, any tasks that share a single underlying cognitive process must then, by necessity, also share a monotonic relationship between their respective task performances. Therefore, under these assumptions, a non-monotonic relationship between task performances is sufficient to demonstrate a dissociation, this is our reversed association. Note that the opposite does not apply, a monotonic relationship is not sufficient to demonstrate that the cognitive functions underlying the two lack a dissociation. In order to undertake statistical inference for a reversed association, we turn to Bayesian statistics.

Quantifying the results – the Bayesian method

We describe state-trace analysis informally in terms of a state-trace plot, e.g. figure 3. We have a state factor consisting of our two tasks, with the performance on each task forming an axis on our graph. We then plot on this graph each level of our dimension factor, the variable that we are varying across our tasks. If we can draw a monotonically increasing (or decreasing) curve joining all the levels of our dimension factor, the relationship between our task performances across our variable is monotonic. In all other cases, it is non-monotonic. In the context of our attentional blink experiment, identity report and judging visibility are our two tasks so they give us our state factor, and the lags are the measure that we are varying across both tasks, so they give us our dimension factor. Plotting report accuracy on one axis and visibility on the other, we are trying to determine whether it is possible to draw a monotonic curve joining the data across each of our lags.

More formally, we have some state factor with two levels $S = \{S_1, S_2\}$, forming the state space over which we examine our question of interest, and some dimension space $D = \{D_1, \ldots, D_n\}$, a manipulation we are performing

across it. When concerned with monotonicity versus non-monotonicity, we wish to see if the ordering of the levels of our dimension factor are either the same or the reverse of one another across each of the two axes of our state factor. If this is possible, we diagnose monotonicity, and if it is not possible we do not. Often, we also introduce a trace $T = \{T_1, \ldots, T_n\}$ factor, but in our case, a trace factor is not required and we therefore exclude it from further discussion. Overall, we must consider each combination of $Q = D!$ orderings for each axis and Q^2 joint orderings. A visual example of both monotonic and non-monotonic state-trace plots can be found in figure 3.

At this point, the set of Q^2 joint orderings corresponds to the whole space of possible configurations of the state-trace graph, and currently it can be divided into two different partitions. These are the non-monotonic orderings and the monotonic orderings. With respect to our Bayesian statistics, we are attempting to choose between the monotonic model consisting of all monotonic orderings, and our non-monotonic model consisting of all other (non-monotonic) orderings. To do this, we calculate a Bayes factor expressing how much the data has changed our preference between our two models. This is the measure of the ratio of evidence for each model. Explicitly, denoting our data as y, the prior probabilities $P(x)$ where $x = M$ or NM as π_M and π_{NM} for the monotonic and non-monotonic models respectively, and the posterior probabilities $P(x|y)$ where $x = M$ or NM as $\pi_M^{(y)}$ and $\pi_{NM}^{(y)}$, we calculate the Bayes factor as:

$$BF_{M/NM} = \frac{\pi_M^{(y)}}{\pi_{NM}^{(y)}} / \frac{\pi_M}{\pi_{NM}}$$

We calculate our posterior using the library provided in[21]. We follow[21] in referring to this calculation as $BF_{M/NM}$, the bayes factor comparing the monotonic versus non-monotonic models.

Currently, we make use of a completely uniform prior, effectively assuming all possible orderings of the lags across the levels of the state factor are equally likely. In many data sets, including our own, this is clearly not true – we, for example, have strong prior expectations about the behaviour of the attentional blink. Previous work has approached this problem by using the prior to assert that certain constraints on the behaviour in the data are true. For example, in[21] the authors pre-suppose that dual task performance will always be worse than single task performance in their analysis of a data set from[28]. However, while we have expectations about the behaviour in the attentional blink, setting specific ordinal qualifications of behaviour across lags in a similar manner is non-trivial. While we wish to take advantage of as much prior knowledge as possible, the behaviour of the attentional blink is variable, and it is well established that setting a poor prior can compromise the integrity of results[29]. As well as setting a prior based on previous literature, we also therefore make use of an empirical prior method to derive a suitable prior. This method takes the set of constraints on the prior identified from the literature, and reduces the set to one that accurately fits the data, using a measure of the validity of constraints orthogonal to the contrast of interest. Details of this method can be found in supplementary material. We denote the validity of a prior calculated using this method as $BF_{D/N(D)}$, and similarly any Bayes factor calculated from a prior that accounts for information on our dimension axis (whether generated from our empirical priors method or not) as $BF_{(M/NM)|D}$.

We must also consider how to apply this type of analysis across a group of participants. Notably, state-trace analysis does not work well with approaches based on averaging. In particular, it is possible both to average multiple non-monotonic datasets into a monotonic dataset, and multiple monotonic datasets into a non-monotonic one. A simple alternative analysis is the grouped Bayes factor introduced by[22]. This method treats each of our participants (of which there are M) as independent from one another and calculates the group Bayes factor as the product of each individual Bayes factor:

$$GBF = \prod_{i=1}^{M} BF_i$$

As long as participants are independent samples and the results are reasonably homogeneous (not, for example, being driven by a single outlier), this grouped Bayes factor is a good summary of the group level effect. This will be the case in the data we analyse with one exception that will be discussed seperately.

STST model

In addition to the methods of state-trace analysis, we explore the potential dissociation of subjective experience and report accuracy through modelling. Specifically, we investigate the hypothesis that the differences in behaviour in the data from[11] that we analyse in this paper are the result of the systems of subjective experience

and working memory encoding being dissociated. We suggest that stimuli are experienced in a serial manner (reflecting the unitary nature of consciousness), but simultaneously encoded into working memory. The Simultaneous Type/Serial Token (STST) model[10] is in a uniquely strong position to explore this, though the model does not natively deal with subjective experience. In this section, we explore a simple set of additions to the STST model that allow it to read out a measure of subjective experience in addition to reporting accuracy. Before this however, we briefly summarise the workings of the Simultaneous Type/Serial Token model.

The STST model, see figure 4, is a two stage model that builds on a type/token distinction to simulate how items are bound into temporal contexts. In this definition, the type of a stimulus encompasses all of its instance invariant properties: the features that do not change between occurrences. Take the letter K for example; parts of its type are its semantic features (e.g. it's a letter, it's after J in the alphabet) and its visual features (e.g. its shape and colour). Conversely, a token represents a specific episodic occurrence of a type e.g. where it occurred in time relative to other items. In the STST model, types are processed in parallel, with many types simultaneously but fleetingly represented, and it is the act of sequentially binding a type to a token that creates a solidified representation in working memory.

The first stage of the model concerns the types and consists of four layers supporting different aspects of visual processing: the input layer, the masking layer, the item layer and the task-filtered layers. The second stage of the model governs the tokenisation process, and consists of the binding pool and the tokens. Items first arise in the input layer, and then pass through the masking layer, which implements masking, and would most naturally be associated with iconic memory[2]. From here, items enter the item layer, which creates a brief, self-sustained representation. Then, the final layer of the stage: the task filtered layer, provides a salience filter that excites task relevant nodes while inhibiting others. From the task filtered layer, sufficiently active items can activate tokens through the binding pool, and become bound to them through a tokenisation process. This tokenisation process takes several hundred milliseconds, though it is shorter for more active items. In order to reach sufficient activation to achieve this binding however, most stimuli will need to benefit from the blaster. When an item becomes sufficiently active in the task filtered layer, the blaster provides a brief, powerful enhancement to the entire task filtered and item layers that allows items to reach the threshold for tokenisation. During this process, a powerful inhibitory signal holds the blaster low to prevent it from re-firing and corrupting the tokenisation process: it is this inhibition of the blaster that generates the attentional blink. A walk through of how an individual item becomes encoded into working memory can be seen in figure 4.

Through these mechanisms, the Simultaneous Type/Serial Token model creates an account of working memory encoding in which types are processed simultaneously, but due to the way the blaster and the tokenisation process work, types can only be bound in serial. There exists a computational model of STST from which it is possible to generate both behavioural data, and also "virtual" ERP's[30,31] that closely mimic the results from human participants. It is an ideal choice for modelling the data which we are exploring, because it is specific to the paradigm we are using (the attentional blink), and it already deals naturally with the difference between simultaneity and seriality.

As discussed, the published STST model does not however, deal with subjective experience, and one of the contributions of this paper is to propose and implement a system by which this can be obtained. However, very many, and often any behaviours can be obtained from a model with sufficient modification and parameter adjustments[32]. In order to make the fairest possible assessment of the hypothesis in question, the dissociability of subjective experience and report accuracy during the attentional blink, we therefore limited ourselves in two ways in our modelling. Firstly, we would attempt to build on top of the existing model to provide a new "readout" without changing the existing model in any way. Secondly, this readout must be simple; ideally arising from one or two principles.

The result of these conditions is the following model to encapsulate serial experience: Subjective visibility is indexed by the strength of the P3 ERP component. When an item is above a given amplitude (the threshold of subjectivity), it is being "subjectively experienced" and when it is below, it is not. Additionally, this experience is serial. If the individual activation traces for two items are both above the threshold, then the second item cannot be experienced until the first one falls below the threshold. For an illustration of this, see figure 5. Specifically, the strength of an item's subjective experience is the duration for which its activation trace exceeds the threshold of subjectivity, subject to no other stimulus already being above the threshold. In this manner, a system allowing a subjective experience that is exclusively serial in manner is created, with only one addition on top of the existing model. We call this readout-enhanced STST model, the Simultaneous Encoding, Serial Experience model (SESE). In order to evaluate the success of this modified STST model, we will compare its behavioural output to that of human participants and the virtual ERPs it generates to human EEGs in the data from[11]. This

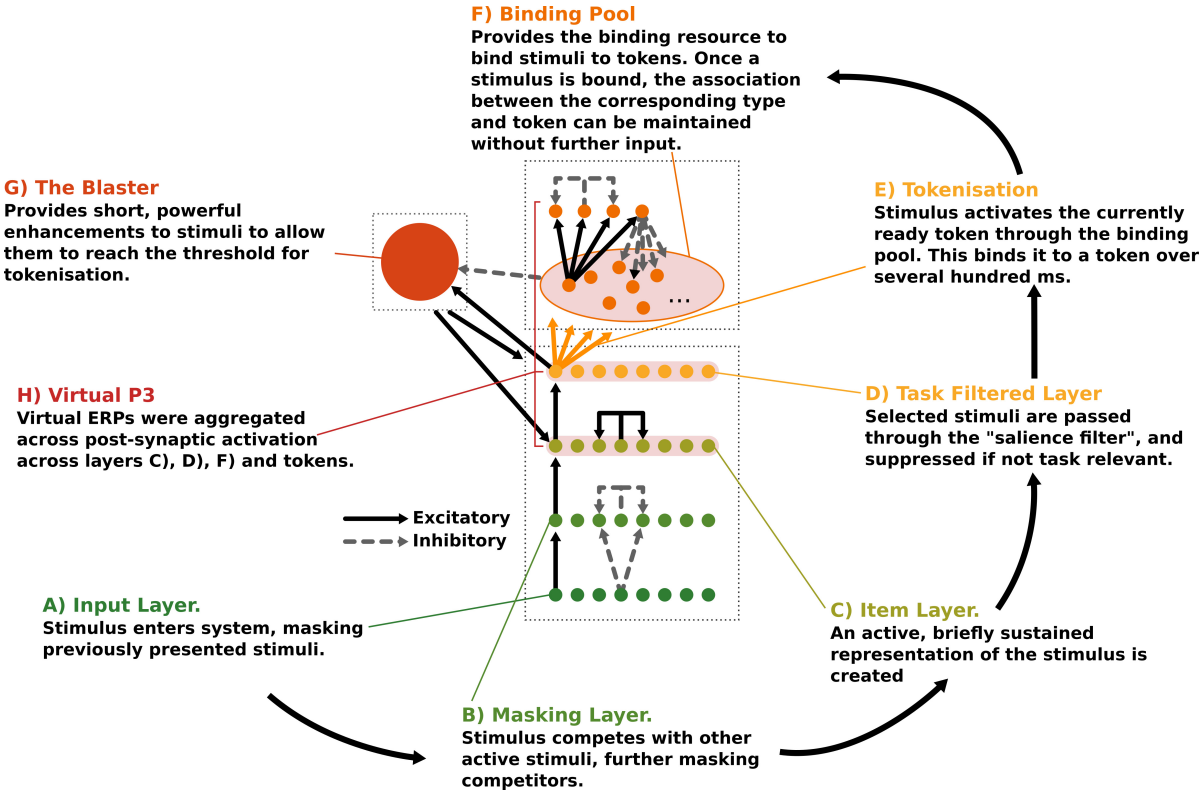


Figure 4. A) Input Layer. Stimuli enter the system through this layer. As well as providing input, this layer implements backward masking through inhibitory connections to all other stimuli in the masking layer. B) Masking Layer. Simulates further masking dynamically through lateral inhibitory connections to all other stimuli. These lateral inhibitory connections are weaker than the forward ones from the input layer, such that backward masking is stronger than forward masking. C) Item Layer. Creates a temporary representation of a stimulus through self-reinforcing connections. D) Task Filtered Layer. Implements a "salience filter" to filter out task irrelevant stimuli, by enhancing task relevant stimuli, and suppressing others. E) Tokenisation. When a stimulus has reached an appropriate level of activation, it excites the currently ready token through the binding pool. In a process that takes several hundred ms, the token is bound to the type. Once this binding has occurred, the type-token connection can be maintained without any further input. F) The Binding Pool. Contains the binding resources that enable stimuli to bind to tokens. G) The Blaster. Provides a short, powerful enhancement to items in the item and task filtered layers when there is sufficient activation in the task filtered layer to indicate the 'detection' of a target and warrant the onset of tokenisation. While the tokenisation process is ongoing, a powerful inhibitory signal from the binding pool prevents the blaster firing again. H) Virtual P3. A virtual P3 can be generated from the STST model from the excitatory post synaptic potentials of the item layer, the task filtered layer, and a subset of the tokens and binding pool (the token gates and the binder gates).

specification of subjective experience mandates a change to how we calculate the grand P3 ERPs from the model. The ERPs generated from the model in[31] are calculated by summing all components together. In this model, when a first target's activation trace crosses the threshold, it starts contributing to the P3, however, the activation traces of other targets do not contribute to the P3. A more detailed desciption of how virtual ERPs can be obtained from the model is available in supplementary material section D.

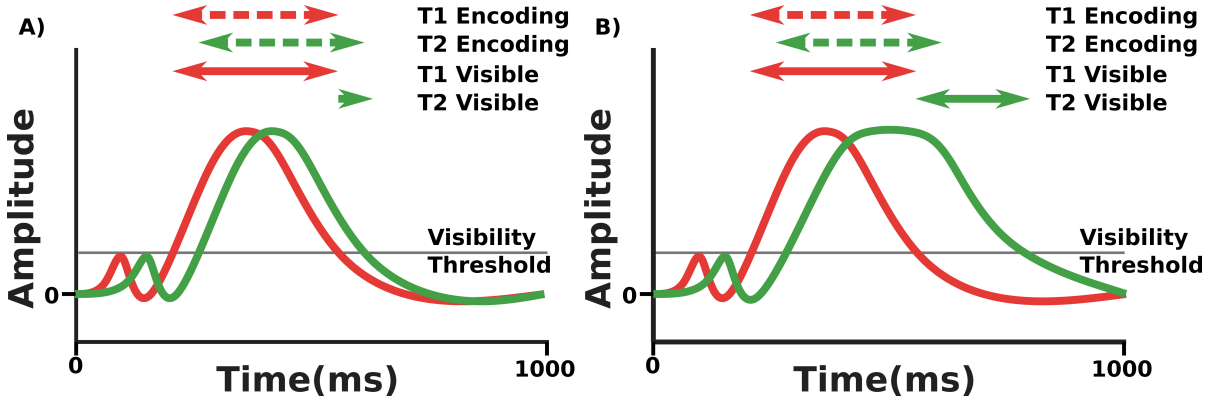


Figure 5. A) Seriality of experience in the SESE model. In A), though the amplitude of the response of both stimuli is the same, the duration of the experience of the second stimulus is greatly reduced because it cannot be experienced until the first stimulus falls below the threshold. Comparatively, in B), the response amplitude of both stimuli is the same, although the T2's activation trace is longer with a slightly delayed onset, consequently they are both experienced for similar durations.

Predictions and Validation

Our current model makes some strong predictions, some of which cannot be immediately validated through the analysis of our first, dataset which we distinguish by referring to it as the colour-marked task (since in the task, the T1 is colour marked, which is not the case in the letters-in-digits task that we introduce shortly). In this section, we discuss these analyses and propose several further analyses to support our hypothesis.

One critiscism of an analysis based on the colour-marked data we present in figure 1 is that the the very substantial differences in report accuracy and subjective visibility at Lag 1 may be due to the use of a colour-marked T1. Previous experiments that have examined subjective report in the attentional blink often find some degree of sparing of subjective visibility at lag 1 (see, for example[8,24]), which is not observed in the colour-marked T1 data. In light of this, we propose a replication without a colour marked T1, giving a pure letters-in-digits paradigm. Details of the experimental procedure will be given in our materials and methods section, but the behavioural results can be seen in figure 6, and interestingly, we do see sparing for subjective visibility at Lag 1, although we will still be able to show the dissociation between report accuracy and visibility at Lag 1 that is central to our argument.

We also need to buttress ourselves against the possibility that we are observing a dissociation between report accuracy and subjective experience for reasons that do not entail the sight-blind recall effect we are considering. This might occur if there is a different mechanism modulating visibility at Lag 1, than at other data points. This is a very pertinent concern, since the Lag 1 data-point is often argued to be unique in respect of attentional blink lags; it is, for example, by far the most vulnerable to order errors[23], or integration of both targets into one perceptual episode[25]. We take two routes to addressing this potential concern. Firstly, and most directly, we show that with the removal of the Lag 1 data-point in the replication (pure letters-in-digits) experiment just discussed, the effect still remains non-monotonic.

Secondly, contrary to a temporal integration explanation, a clear prediction of our proposal is that "if the individual P3s for two items are above the (conscious awareness) threshold, then the second item cannot be experienced until the P3 for the first one falls below threshold". As a result, the visibility (relative to accuracy) for T1 should remain intact at Lag 1 compared to other lags, since it will be experienced to completion, or, in other words, the co-active T2 cannot interrupt the ongoing experience of T1. According to a temporal integration account, visibility of T1 should be impaired at Lag 1, since integration fundamentally suggests a T1-T2 "composite" is constructed, which would surely imply an impact of T2 onto T1. In contrast, we predict

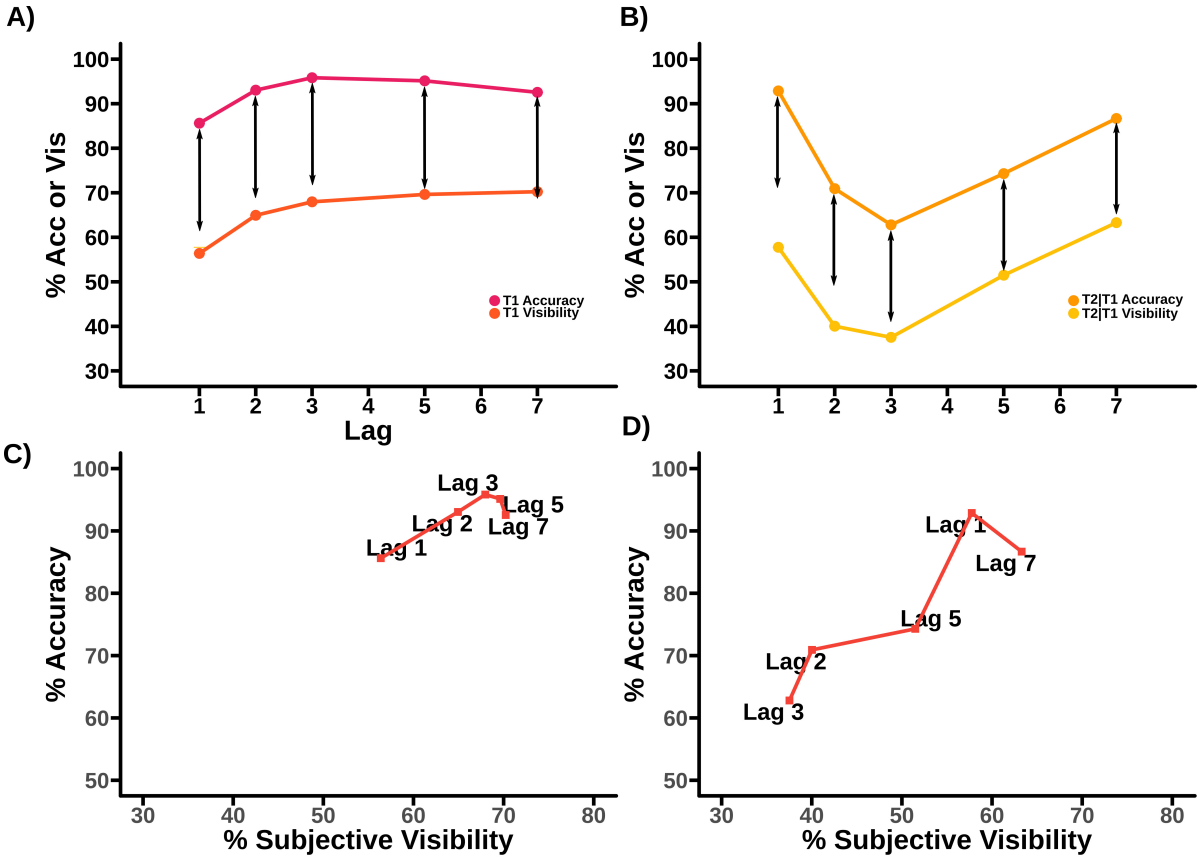


Figure 6. Behaviour of replication (pure letters in digits) data, comparing accuracy and subjective visibility across lags in the attentional blink. A) A comparison of report accuracy and visibility ratings for T1. B) A comparison of report accuracy and visibility ratings for T2. C) A state-trace plot comparing accuracy and visibility for T1. D) A state-trace plot comparing accuracy and visibility for T2. What we show as T2|T1 visibility is the visibility rating of T2 on all trials in which T1 was correctly reported. Note that compared to the analysis in[11], T2 visibility shows a level of Lag 1 sparing. This dataset also measures visibility of the first target, which was not collected in the (colour-marked AB) study of[11]. Importantly, however, the basic dissociation of report accuracy and subjective visibility at short lags that underlies our hypothesis is qualitatively present for T2; see panel B). For example, Lag 1 sparing is substantially higher for report accuracy than subjective visibility relative to other lags. This is illustrated by the black arrows, which indicate a constant distance for each graph. This can also be seen by noticing that, for T2 report accuracy, Lag 1 is considerably higher than Lag 7, while for subjective visibility it is marginally lower. Notice that the T1 curves do not seem to show the dissociation at early lags between report accuracy and subjective visibility that we see for T2. In particular, the differences in vertical distance across lag that are present in panel A) may just be a facet of the small dip in T1 accuracy at later lags, a feature that we have not observed previously and which may just reflect “sampling error”.

that T1 is isolated from the interference of a proximal T2. To address this concern, we propose a state-trace analysis of the T1 data of the replication (letters-in-digits) experiment. This has several advantages. First, it allows us to robustly examine whether visibility is changing differently with respect to accuracy across lags, when compared to our first (colour-marked) experiment. Second, a monotonic finding for T1 in the replication experiment would provide evidence directly against target integration.

One further analysis we perform is to examine report accuracy when participants indicate an absence of subjective visibility at Lags 1 and 3. This is a key analysis for the idea of sight-blind recall. That is, being able to show above chance report accuracy for T2, when participants select the bottom subjective visibility bin, i.e. nothing seen, suggests recall without experience. Showing that this phenomenon is larger at lag-1 than lag-3 further supports our position that co-activation (although not co-experience) of T1 and T2 particularly drives the dissociation of visibility from report accuracy. A preliminary version of this analysis was reported in the supplementary material of[11]. To maximise the available data for this analysis, we perform it on the second set of data from[11], which sampled fewer lags with more trials, compared to the first set of data from[11], which we have examined thus far in this paper. Focusing on this higher-powered data set enabled us to more robustly measure this effect.

Materials and methods

Original colour-marked RSVP Data

Ethics

All experiments were performed in accordance with the relevant guidelines and regulations. The study was approved by the Psychology Research Ethics Committee at the University of Cambridge, UK and participants provided informed, written consent.

Data

Our set of data is a behavioural attentional blink dataset previously presented in[11]. Full details of the experimental procedure is given in the original paper, we summarize this here for clarity. Data was collected for two experiments, a behavioural set that sampled a large number of lags over fewer trials per lag (Experiment 1), and an electrophysiological set that additionally collected EEG data, and sampled fewer lags (Experiment 2).

Targets were uppercase letters and distractors were single digits, each trial contained one or two targets - T1 occurred on every trial and was always presented in red, and T2 (if it occurred) was presented in white. Targets could be any one of 21 letters, with 5 letters excluded because of similarity to numbers. Each RSVP stream contained 15 items. T1 randomly appeared as the fourth, fifth or sixth item in the RSVP stream. Stimulus Onset Asynchrony (SOA), the amount of time between the onset of each stimulus, was 90ms. At the end of each RSVP stream, participants were asked to rate the subjective visibility of T2 using a 6 point self-report scale. The numbers 1 2 3 4 5 6 were presented in a horizontal line on the screen, with the description "not seen" presented beneath the number 1 and the description "maximal visibility" presented beneath the number 6. Participants then reported the identity of T1 and T2 (even if a second target did not occur). Participants were required to guess if they were unsure of the target identities. In Experiment 1, T2 appeared at lags 1, 2, 3, 4, 6, 8, or not at all with equal frequency. Results of this experiment for 18 participants were presented in figure 1. In Experiment 2, targets appeared at Lag 1 (40% of trials), Lag 3 (40% of trials), Lag 6 (10% of trials) and not at all (10% of trials). Experiment 1 deliberately sampled a large number of lags in order to examine the relationship between T2 accuracy and subjective visibility across the entire AB curve, while Experiment 2 sampled fewer in order to facilitate the creation of robust EEG data. Note that in contrast to the original study, for our state-trace analysis of second targets (T2s), we only include trials in which T1 is present and T1 and T2 are reported in the correct order in order to avoid order errors as a confound. This applies for both our accuracy and visibility ratings.

Implementation specifics

Setting the prior

We set the prior of our Bayesian analysis from prior literature, specifically based on the results from[24]. This paper presents both a classic attentional blink with lag 1 sparing of report accuracy, and a similar "experiential" blink of subjective report in which lag 1 is spared a great deal less. Due to the well-established evidence for the pattern of behaviour in the attentional blink, we encoded strong expectations of behaviour, including lag 1 sparing, of the report accuracy in our data. Comparatively, the evidence for the behaviour of subjective report during the blink is less well established, so we refrained from imposing such strong constraints about it, particularly at the important lag 1 data point. We also recognise some uncertainty about the deepest point in the attentional blink:

given the SOA of 90ms, we could reasonably expect either of lags 2 or 3 to be the deepest point in the blink. We therefore set our prior to be consistent with several potential deepest points. Finally, Lag 8 is a serial position outlier (A common finding in attentional blink experiments is that a last lag that is a serial position outlier, e.g. if there is no Lag 7 and most lags in the experiment are short, participants will come to learn this regularity and optimize the allocation of attentional resources to short lags, causing lag 8 performance to be relatively low across the experiment.) in our experiment and was therefore removed from our analysis. These considerations resulted in a uniform prior subject to the following constraints across our data: for report accuracy, Lags 1, 4 and 6 would be held to be larger than Lags 2 and 3, with Lag 1 additionally being held to also be larger than Lag 4. For subjective report, Lag 6 would be held to be higher than Lag 4, Lag 4 higher than Lag 3, and Lag 3 higher than Lag 2. The validity of these constraints, as determined by our empirical priors method discussed in the supplementary material section A was strong, but not completely homogenous. We therefore applied our method of empirical priors to reduce them to a set with a better fit. After application of our method, our prior was still uniform, subject to constraints as follows: For report accuracy, Lags 1 and 6 would be held to be larger than lags 2 and 3, and Lag 1 additionally would be held to be larger than Lag 4. The constraints for subjective report remained unchanged.

Distribution of data

The state-trace method we are applying, based on the work of[21,22], assumes a binomial distribution of the data. This is suitable for our accuracy data, which is a dichotomous variable, but not for our visibility scale that forms a multinomial distribution over 6 values. Consequently, we grouped our visibility results into two bins, a high visibility bin and a low visibility bin. To decide the fairest way of applying this split, we calculated the grouped bayes factor comparing the validity of the constraints for each possible method of splitting the data, for both the full and empirically determined prior. The results (see supplementary material section C) clearly show that the "best" split is that of assigning the top 50% of visibility ratings to the high visibility bin and the bottom 50% to the low visibility bin.

Replication pure letters-in-digits RSVP Data

Ethics

All experiments were performed in accordance with the relevant guidelines and regulations. The study was approved by the Faculty of Sciences Ethics Committee at the University of Kent, UK and participants provided informed, written consent.

Data

Our data is a set previously presented in[33], collected by Ellis Luise Gootjes-Dreesbach as part of her doctoral research at the University of Kent. 12 young adults took part in this study, aged 18-30 with a mean age of 21.83 years. Targets were upper case letter and distractors single digits. Targets could be any one of 21 letters, with 5 letters excluded because of similarity to numbers. Each trial contained two targets, with no colour marking for either target. Each RSVP steam contained 20 items. T1 randomly appeared as the 7th, 8th or 9th item in the steam. T2 was pseudorandomly presented at Lags 1, 2, 3, 5 or 7, ensuring an equal number of trials in each condition. Stimulus Onset Asynchrony (SOA), the amount of time between the onset of each stimulus, was 83ms. At the end of the stream, participants were asked to respond (via the keyboard) to four questions about the visibiltiy and identity of T1 and T2. The query for target visibility ('On a scale of 1-6, please indicate how well you saw the first [second] letter') was paired with an ASCII representation of a 6-point scale with the low end labelled as "not seen" and the high end labelled "maximal visibility". Target identity was queried by asking "What was the first [second] letter you saw? If you are not sure, give your best guess.". We analysed all trials whatever the report order. The whole experiment consisted of 4 blocks of 45 trials, each randomised with respect to lag and T1 position.

Implementation specifics

Setting the prior

This experiment sampled slightly different lags to the original colour-marked experiment, but we attempted to replicate the constraints used in the previous experiments as closely as possible for the analysis of T2. Specifically, we substituted all constraints in the previous experiment, with Lag 5 replacing Lag 4, and Lag 7 replacing Lag 6. For T1, lacking any precedent in the literature for the behaviour of T1 visibility, we placed no constraints on the possible orderings of our data. For this replication experiment, in order that constraints did not change from those in the original data set, we did not make use of our method of deriving empirical constriants.

Distribution of data

To provide the fairest comparison to our original (colour-marked) analysis, we maintained the previous split of visibility ratings into high and low bins.

Availability of data

All of the code used in this project has been open sourced on Github, subject to an MIT liscence. See https://github.com/william-r-jones/StateTrace for the modified state-trace code, and https://github.com/william-r-jones/SESE for the modified STST model. All of the data used in this paper is also available alongside this code where possible, though some datasets (notably the EEG data) are too large for this to be possible and have instead been made available using the Dataverse Project. See https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%

Results

Original Colour-Marked Data

State-Trace Results (T2)

Figure 7(A) shows validity for each participant for the original set of prior constraints derived from[24]. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (not log) $BF_{D/N(D)} = 1.22 \times 10^{9}$. However, we note that while the group validity is strong, four participants show the opposite pattern. Figure 7(B) shows the respective non-monotonicity for this set of constraints. Results are strongly and almost homogenously in favour of the non-monotonic model, with grouped (not log) $BF_{(M/NM)|D} = 2.25 \times 10^{-14}$.

Figure 7(C) shows validity for each participant for the set of prior constraints derived from the original using our empirical prior method. At the group level, the evidence is strongly in favour of the constraints fitting the data, with grouped (not log) $BF_{D/N(D)} = 1.07 \times 10^{13}$. However, we note that while the group validity is strong, there remains some variability across participants, though this situation has noticeably improved compared to 7(A). Figure 7(D) shows the respective non-monotonicity for this set of prior constraints. Results here are strongly and almost completely homogenously in favour of the non-monotonic model, with grouped (not log) $BF_{(M/NM)|D} = 1.17 \times 10^{-17}$.

Replication Letters-in-Digits Data

T2

Figure 8(A) shows validity for each participant for the prior adapted from the original colour-marked T1 data analysis. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (not log) $BF_{D/N(D)} = 1.46 \times 10^{11}$. Figure 8(B) shows the respective non-monotonicity for this set of constraints. Results are in favour of the non-monotonic model, with grouped (not log) $BF_{(M/NM)|D} = 1.14 \times 10^{-2}$.

T2 No Lag 1

Figure 9(A) shows validity for each participant for the prior adapted from the original colour-marked T1 data analysis, with Lag 1 removed. At the group level, the evidence is strongly in favour of the constraints fitting the data with grouped (not log) $BF_{D/N(D)} = 2.5 \times 10^{9}$. Figure 9(B) shows the respective non-monotonicity for this set of constraints. Results are in favour of the non-monotonic model, with grouped (not log) $BF_{(M/NM)|D} = 5.75 \times 10^{-4}$.

T1

Figure 10 shows the respective non-monotonicity test for T1. Results are in favour of the monotonic model, with grouped (not log) $BF_{(M/NM)|D} = 7.36 \times 10^{4}$.

Simultaneous Type/Serial Token Model Results

Our first comparison is the behavioural results of the STST model and those from[11]; see figure 11. Note the qualitative similarity in behaviour. Such a high similarity between empirical and model findings is rare without a fitting of model parameters to the data.

We also compared the human ERPs with the virtual ERPs generated by the STST model, see figure 12. For full details on how these are obtained, see the supplementary information. We present two sets of model ERPs, comparing each of them to the same human ERPs, i.e. Lag 1. Panel A) compares to model Lag 1 and B) to model Lag 2. It should be clear from this that there are features of both the models Lag 1 and Lag 2 that are similar to the human Lag 1. This is perhaps not surprising and suggests a fixed offset timing difference between model and human data. Additionally, there are further reasons why it is unrealistic to expect a more perfect fit between simulations and empirical findings. Firstly, the task modelled by STST does not have a colour marked T1, which

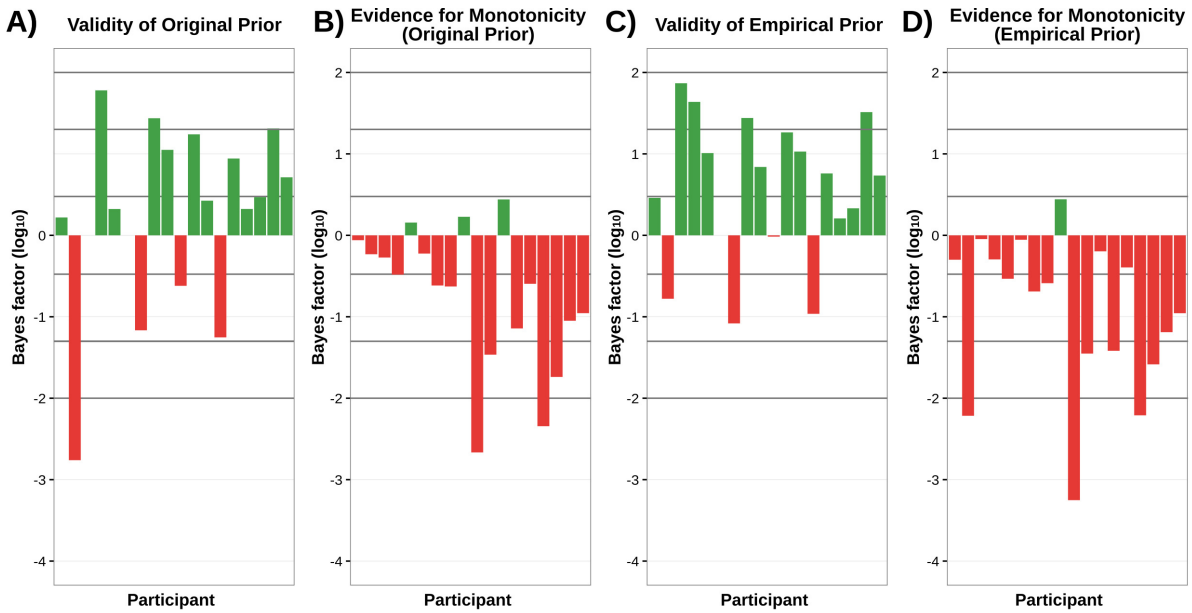


Figure 7. Log_{10} Bayes factors for each participant across 4 different tests, for T2 in the original (colour-marked T1) experiment. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{10000}$, $\frac{1}{1000}$, $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20, 100, and 1000 respectively. A) Evidence for validity of the prior by participant for the original prior based on[24]. B) Evidence for monotonicity (positive) vs non-monotonicity (negative) by participant for the original prior. C) Evidence for validity of the empirically derived prior. D) Evidence for monotonicity (positive) vs non-monotonicity (negative) by participant for empirically derived prior.

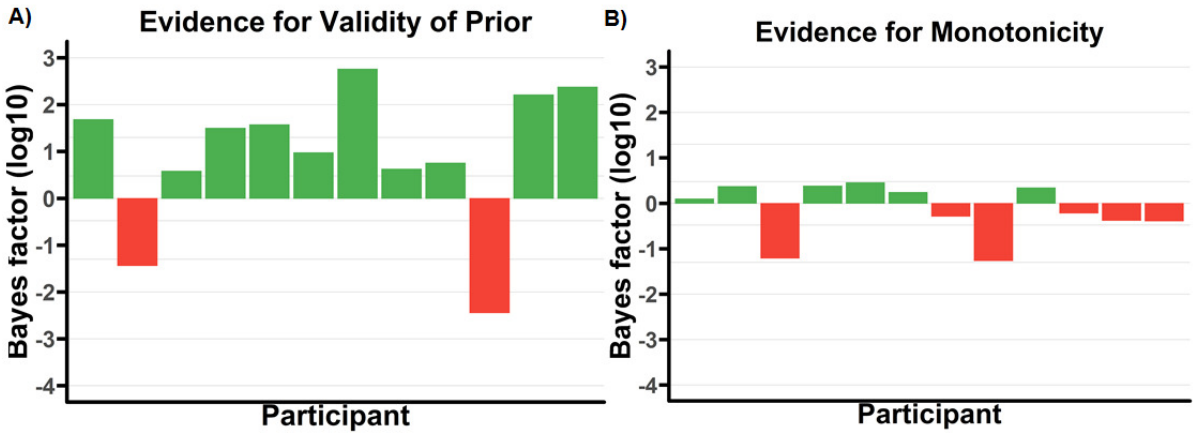


Figure 8. Log_{10} Bayes factors for each participant for monotonicity and validity of constraints for T2 in the replication (pure letters-in-digits) experiment. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{10000}$, $\frac{1}{1000}$, $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20, 100, and 1000 respectively. A) Evidence for validity of the prior adapted from the original (colour-marked T1) analysis. B) Evidence for monotonicity (positive) vs non-monotonicity (negative) by participant for this prior. Although the effect here is not as strong as it is for the original (colour-marked T1) experiment, the data does not exhibit the pattern in which the grouped Bayes Factor becomes a problematic measure, which arises, for example, if there is a single outlier subject driving the effect.

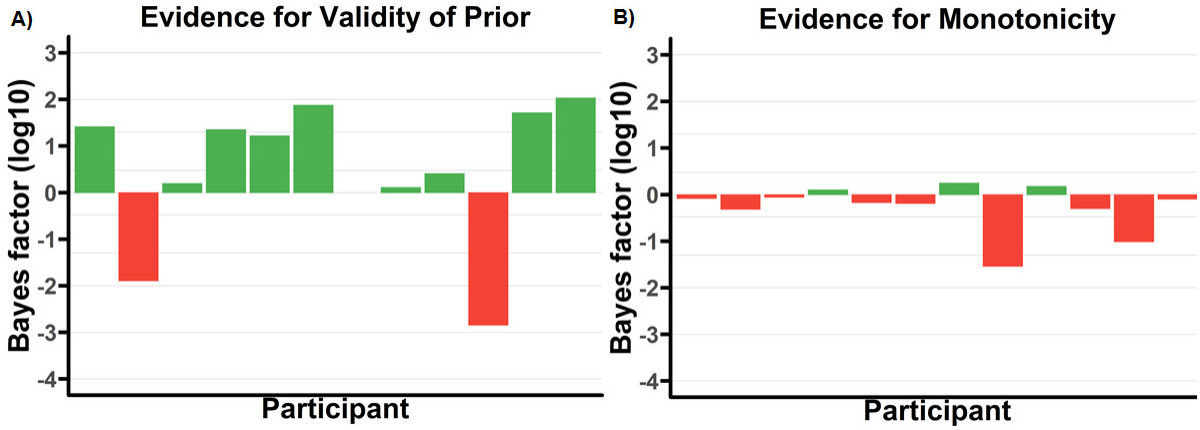


Figure 9. Log_{10} Bayes factors for each participant for monotonicity and validity of constraints for T2 in the replication (pure letters-in-digits) experiment with no Lag 1. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{10000}$, $\frac{1}{1000}$, $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20, 100, and 1000 respectively. A) Evidence for validity of the prior from the (colour-marked T1) analysis. B) Evidence for monotonicity (positive) vs non-monotonicity (negative) by participant for this prior.

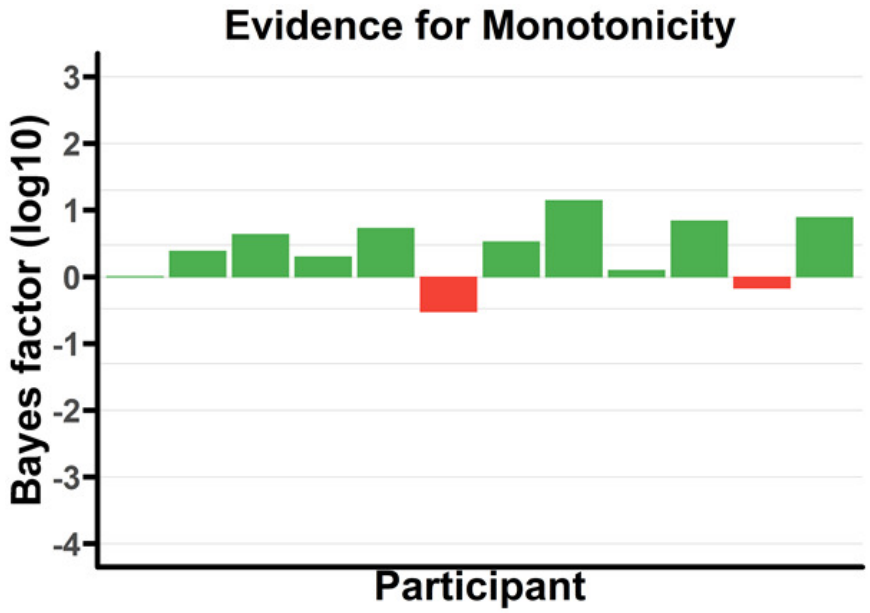


Figure 10. Log_{10} Bayes factors for each participant for monotonicity for T1 in the replication (pure letters-in-digits) experiment. Note that participants are in the same order in all graphs to facilitate comparison. Lines overlaying the figure correspond to bayes factors of $\frac{1}{10000}$, $\frac{1}{1000}$, $\frac{1}{100}$, $\frac{1}{20}$, $\frac{1}{3}$, 3, 20, 100, and 1000.

is likely to explain why the transient around 200ms in the human data is not replicated by STST. Secondly, we are comparing scalp EEG directly to model deflections, without recourse to a forward (lead field) model of how brain sources are projected into sensor space. Critically though, the key property that a clear conscious percept of T2 (i.e. the high visibility condition) coincides with a longer P3 is qualitatively present in both sets of virtual ERPs. This pattern resonates with the notion that conscious perception imposes a seriality constraint that is not required for encoding into working memory. Some further results are available in Supplementary Section E, where we compare human and virtual ERPs at later lags.

For illustrative purposes, we also present the activation traces for high and low visibility, for each of the T1, T2 and distractors seperately. We do this for each lag separately. This can be seen in figures 13(A) and 13(B) (Lag 1) and figures 13(C) and 13(D) (Lag 2). This clarifies how the Virtual ERPs in figure 12 emerge from the underlying STST activation traces. An STST virtual ERP, as presented in[31], is a summation of the traces in a panel of figure 13, including the low amplitude responses to distractors, which contribute to the "rougher" contours of the figure 12 model time series compared to the figure 13 target time series. Critically, the experience read-out mechanism we are proposing here means that the T1 and T2 traces are not simply summed when they are co-active. Rather, the T2 trace only starts contributing to the virtual P3 once the T1 trace has fallen below the visibility threshold, as shown in figure 5. Accordingly, only the the back-end of the T2 trace in figure 13(A) contributes, almost none of it in figure 13(B) and a much larger proportion in figure 13(C).

Report accuracy at minimal subjective visibility

To further justify the term sight-blind-recall, we directly investigated T2|T1 accuracy at the lowest level of subjective visibility. The question of interest is whether we can actually demonstrate that report accuracy is above chance when subjects report zero visibility of the T2. To this end, T2|T1 accuracy was calculated only on trials where participants selected a visibility rating of 1 (the lowest possible visibility rating, indicating 'not seen'). For each lag, T2|T1 accuracy was compared with the degree of accuracy expected due to chance (4.76%, one out of 21 letters presented), using one-sample t-tests. In other words, we investigated whether T2|T1 accuracy was greater than 4.76%, at relevant lags. As discussed, this analysis was conducted for lags 1 and 3 in the (colour-marked T1) second experiment from[11], as that is where the trial counts were sufficiently large to examine a specific subjective visibility (200 trials for each of those lags). As expected, accuracy was significantly greater than chance, despite participants indicating that the subjective visibility of the target was nil (lag 1: $\mu = 37.98\%$, $\sigma = 25.25\%$, $t(1,17)$, $p<.001$, $d=1.3156$), (lag 3:$\mu = 15.03\%$, $\sigma = 12.5\%$, $t(1,17)$, $p=.0014$, $d=0.8214$). We also examined the hypothesis that at minimum visibility report accuracy at lag 1 was greater than report accuracy at lag 3. We found evidence for this hypothesis, (lag 1 > lag 3, $t(1,17) = 5.2033$, $p<.001$, $d= 1.2264$).

Discussion

Monotonicity verseus Non-Monotonicity

Our state-trace analysis, comparing the measures of accuracy and subjective experience in the attentional blink, found strong evidence for a non-monotonic model of the relationship between these two measures at both the individual participant and group level. This was further supported by the methods developed as part of our own contributions to the current state-trace methodology. We would argue that our empirical priors approach identifies a more accurate set of results across the data, however it is encouraging that our results are similar both with and without our empirical priors.

Previous literature[21] has advocated the use of both the Grouped Bayes Factor (GBF) that we have calculated, as well as an Aggregated Bayes Factor (ABF) to confirm the homogeneity of the results, something we have not done. There seems little need to apply the ABF, since our data shows substantial homogeneity in both contrasts for which it is tested: for example, considering our main state-trace finding for our original colour-marked T1 data set, only three participants demonstrate even incidental evidence for a monotonic model (cf. figure 7B) with the original prior, and only one with the empirical prior (cf. figure 7D). Additionally, we note that the ABF cannot be used to confirm homogeneity, only identify heterogeneity.

There is one potential exception to this, figure 8B). In this instance, ignoring the absolute quantity of the effect, exactly half the participants show one Bayes factor direction, and half the other. This is heterogeneous in nature, which, as we have discussed, may be a problem case for the GBF. However, in this instance, we do not believe that we need to be overly concerned. The dangerous case of heterogeneous results in respect of the GBF is that it can potentially lead to a misleading summary of the overall effect. However, that is not the case in figure 8B). While it is true that we have a substantial number of participants supporting both monotonic and

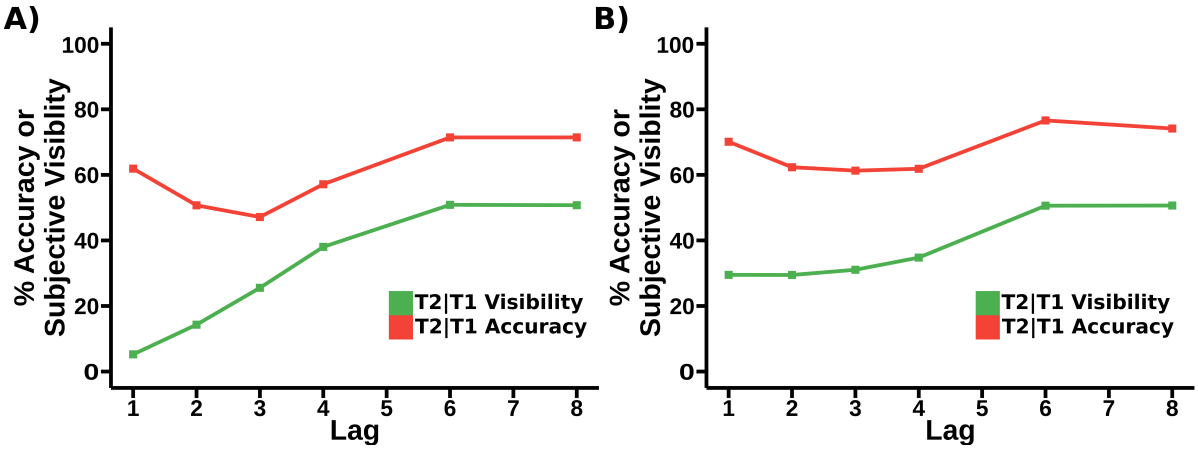


Figure 11. A) Accuracy and subjective visibility by lag for the STST model. B) T2|T1 Accuracy and T2 subjective visibility by lag for the data from[11], i.e. the original (colour-marked) task. Note that these results have appeared in a different figure (figure 1(A)) above, but we present them reformatted here to better facilitate a comparison. Importantly, as previously discussed, neither the function or the structure of the STST model, as given in[10] were changed when generating this fit.

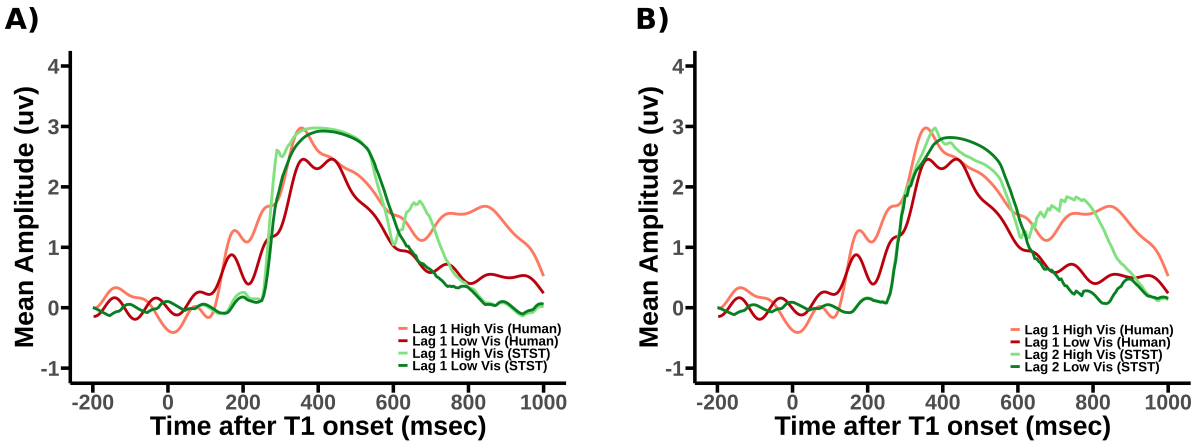


Figure 12. A comparison, for both high and low T2 visibility, given correctly reported T1, of the human ERPs from the original colour-marked T1 data analysis[11]. A) Lag 1 Human ERPs vs Lag 1 STST virtual ERPs. B) Lag 1 Human ERPs vs Lag 2 STST virtual ERPs. Importantly, as previously discussed, neither the function or the structure of the STST model, as given in[10] were changed when generating the virtual P3s. Note that the human ERPs presented are slightly different to those from[11], as ours exclude order errors to be consistent with previous sections.

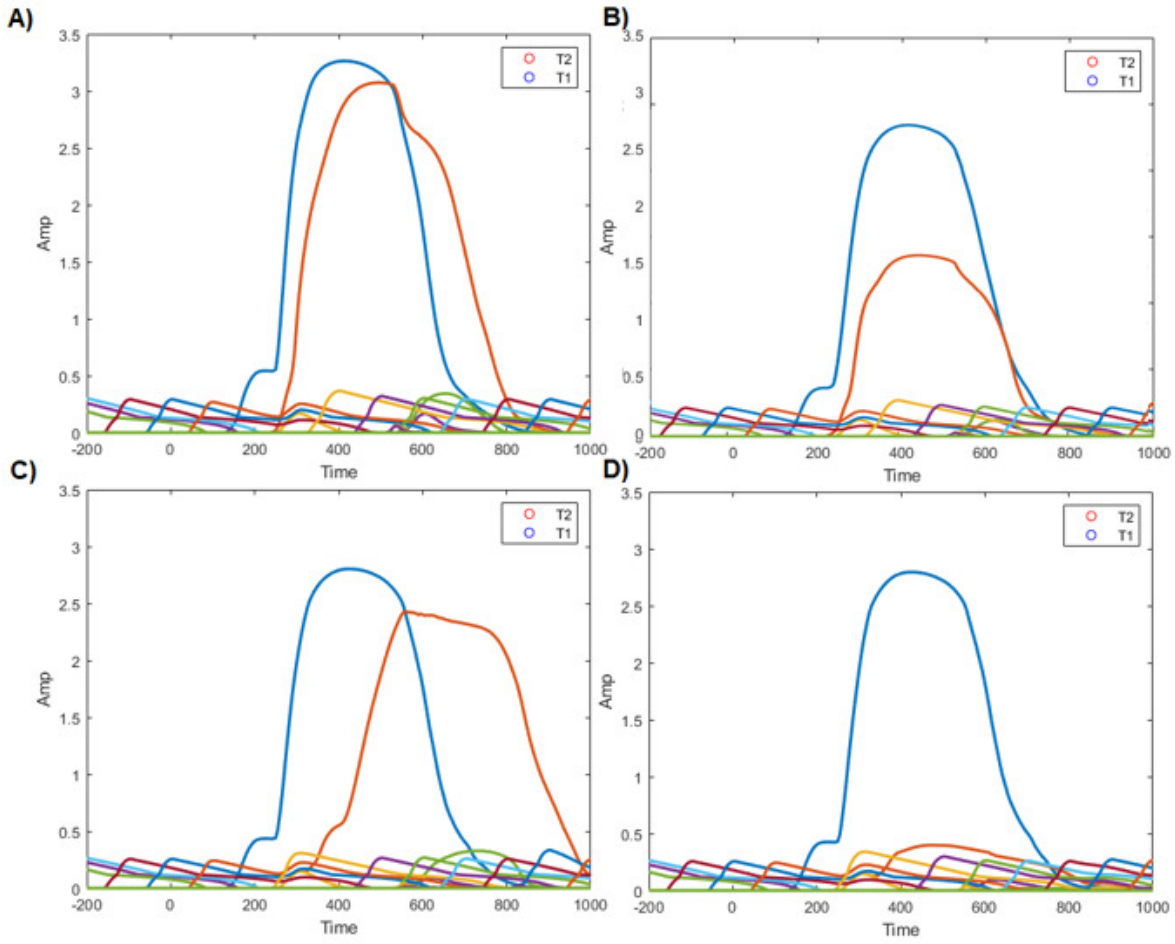


Figure 13. Activation traces by target for virtual data presented in figure 12, split up by visibility and lag. Unlabelled activation traces are from distractors. Each one of these activation traces corresponds to the sum of the excitatory post synaptic potential of the neurons on the 3rd, 4th, 6th and 8th layers of the neural-STST model, corresponding to the item layer, the task filtered layer, the binder gates and the token gates. This is illustrated in figure 4. The 'full' activation traces that are presented in figure 12 are generated from the sum of each of these individual traces at each timepoint, subject to the seriality of experience we have discussed previously; when one target is being experienced, the activation trace of the other target (or indeed, distractors) makes no contribution to the grand activation trace. A) individually depicted activation traces from the SESE model for each target, for high visibility targets at Lag 1. B) individually depicted activation traces from the SESE model for each target, for low visibility T2s at Lag 1. C) individually depicted activation traces from the SESE model for each target, for high visibility targets at Lag 2. D) individually depicted activation traces from the SESE model for each target, for low visibility T2s at Lag 2.

non-monotonic directions, the only non-incidental Bayes Factors we have provide evidence for non-monotonicity. In this case, the most natural interpretation of the data is non-monotonicity, which supports the calculated GBF.

One aspect of our analysis that is notable is the lack of a trace factor. However, the introduction of a trace factor is only required in the case in which there are only two levels of the dimension factor; in other cases, the introduction of a trace factor is a convenience designed to sweep out the behaviour of a system. In our case, we have 5 levels of our dimension factor, which is very close to, or exceeds the combined total trace $\times$ dimension factors in other state-trace experiments[18,28,34].

Working Memory encoding without Subjective Experience

Our results suggest some kind of dissociation between working memory encoding and subjective report. Despite this, we have only demonstrated that a dissociation exists and have not definitively characterised it: we would claim that our findings are indicative of a particular relationship of dependency between working memory encoding and conscious perception, but no more than that. However, our results do not exist in a vacuum. It is clear that the dissociation we observe is a phenomenon of very short lags. In particular, it is largest at Lag 1. For example, in the original (colour-marked T1) study, the series of interactions performed in[11], in which lags were systematically excluded, suggest a strong dissociation at Lag 1, with weakening dissociations from Lag 2 to Lag 3 and nothing at higher lags, additionally, the state-trace analysis performed here on that same data showed non-monotonicity when all lags were included, but the removal of Lag 1 from the state-trace analysis nullified that effect, see supplementary material section B for details of this analysis. Furthermore, the state-trace analysis we perform here on the replication (letters-in-digits) data set shows non-monotonic patterns with all lags in and when Lag 1 is excluded, but the effect is lost when further lags are excluded.

A dissociation restricted to just very early lags, and particularly Lag 1, raises the possibility, but no more than that, of working memory encoding being a necessary, but not sufficient, condition for conscious perception (although, the existence of phenomenological awareness would mean WM encoding was also not necessary for conscious perception). This is because it is at these lags that the activation of T1 and T2 is most strongly simultaneous. Thus, we can say that it is specifically when T1 and T2 are active together that T2 is encoded into WM, with a weakened, or absent, perceptual experience, suggesting a capacity to encode T1 and T2, while the T2 conscious percept is impaired. In addition, our finding in subsection "Report accuracy at minimal subjective visibility", that there is above chance report accuracy when participants report zero visibility, an effect that is substantially stronger at Lag 1 than Lag 3, provides probably the most direct evidence that on some trials encoding into WM can occur without visibility.

We also view the P3s we have observed in the original (colour-marked-T1) experiment as consistent with this interpretation although certainly not definitive verification of it. For example, in figure 12, it is clear that the Lag 1 High Vis (human) is considerably longer than the Lag 1 Low Vis (human). Additionally, in[11], figure 5 compares the ERPs for T2 correct with T2 high visibility (compare the green traces in panels A and F), again the high visibility T2 has a substantially extended P3. This seems to suggest that consciously seeing the T2 dramatically extends the P3, while the curtailed P3 when T2 is just correct, but not necessarily vividly seen, might be considered indicative of a T2 being encoded, with little, if any, conscious experience.

This profile of findings could suggest a phenomenon called "sight-blind recall", however, further empirical support from the RSVP domain and beyond is required to fully justify this interpretation. In particular, the critical demonstration would be that when T2 is correctly reported but given a zero visibility response, the lag 1 P3 is the same as that for a T1 alone. We do not though have sufficient trials in our ERP experiment to reliably construct this average. This, then, is a key test that needs to be performed.

Importantly, this purported sight-blind recall is different from more familiar notions of preconscious processing, such as subliminal priming, implicit perceptual learning as well as related findings demonstrated with continuous flash suppression[15] and phenomena such as blindsight[17], or episodic face recognition[18]. These experiments demonstrate only an indirect effect on a later test; in no case is the "invisible" stimulus that is not consciously perceived directly reportable. We would argue that these results are not strong enough to demonstrate the "sight-blind recall" that we have described, indicating instead influence without experience. In contrast to this, our results suggest the potential for free recall of a stimulus that has not been conscious perceived, a much stronger result that we would argue is far closer to constituting sufficient evidence for "sight-blind recall" and working memory encoding without conscious experience.

The decoupling of subjective visibility from report accuracy at early lags is particularly striking in our original (colour-marked-T1) data set, where there is no evidence of Lag 1 sparing for subjective visibility at all; see figure 11(B). However, it is important to realise that the decoupling effect we have identified is not dependent upon

the complete absence of sparing for subjective visibility, and this is important, since other studies that collected subjective visibility, e.g.[8] and[24], did see lag-1 sparing for subjective visibility. Importantly, the replication (letters-in-digits) data set, indeed, has sparing of subjective-visibility; see figure 6. However, critically, this kick-up at early lags is, in relative terms, considerably smaller for visibility than for report accuracy. Accordingly, we are still able to demonstrate the state-trace non-monotonicity that is central to the argument in this paper, and, in fact, the interaction that was central to[11] can also be demonstrated, see[33].

These findings though raise the question of why different lag-1 subjective visibility patterns have been observed, i.e. why is it that the original (colour-marked-T1) data did not show lag-1 sparing for subjective visibility, but[24] and our replication (letters-in-digits) data set did? Considering our data sets, one factor that surely impacts this is the T1 colour-mark in the original study. This, we believe, makes the T1 perceptually strong and, also, more easily distinguishable from the T2. Indeed, in this data set, T1 report accuracy is considerably higher than T2 report accuracy performance at all lags.

In contrast, the replication (letters-in-digits) study was a straight letters-in-digits task, with no colour marking. This may have caused the T2 to be more strongly perceived, since the T1 is not as strong as it is in the original (colour-marked-T1) study. It is less clear how to reconcile our findings with[24], since they did have a colour-marked T1. However, their colour-marking may not have been as salient as ours: cyan in theirs versus red in ours. This could potentially mean that there is also increased relative strength for T2s in their experiment, increasing its visibility. A definitive answer to these inconsistencies, though, awaits further empirical work.

Broadening out fron the attentional blink, there are several pieces of work that present findings consistent with our results. Firstly, evidence of working memory maintenance without conscious awareness[35] sits very nicely with our results, and this is even more the case for such a demonstration with the attentional blink[36]. If we have indeed found a case in which working memory representations can be formed, without awareness of their formation then we would have identified an explanation for how items could enter working memory without being experienced, which then could be maintained without experience. Our results may help explain how these pre-conscious working memory traces arise by giving them a mechanism through which they can be encoded without conscious experience.

[37] also present experimental conditions in which they are able to use metacontrast masking to vary the subjective report of consciousness, while stimulus discriminability is maintained. Further, the authors find that as SOA decreases (down to around 50ms, at which point the effect reverses) shorter SOAs result in lower subjective experience, consistent with our finding that subjective experience drops as T1 and T2 become closer.[37] is a landmark study; our results, though, move beyond their work by applying state-trace analysis rather than single dissociations, and by considering identification with free recall, rather than two alternative forced choice decisions. In this sense, our objective behaviour relies upon a significantly more complex cognitive process.

Taking our results along with those from[1,3,4] that indicate some degree of perception without reportability, it may be tempting to conclude that working memory encoding and perception are highly correlated but mutually dissociable processes. However, all of the studies above provide their evidence in the form of the single dissociations. Further state-trace analysis could provide additional evidence for the dual question to that studied in this paper.

From a theoretical point of view, it is interesting that perception is most taxed at Lag 1. As we have discussed,[11] note that this pattern of behaviour is consistent with a model of the attentional/experiential blink in which stimuli are consciously perceived in a serial manner, but encoded in a simultaneous manner. This is discussed in further detail below.

Integrated Percepts

One potential criticism of our results is that the low subjective experience at Lag 1 is caused by the rather unique nature of the Lag 1 data point. Lag 1 is the only data point without any intervening distractors, and is, notably, by far the most vulnerable point to order errors[23], or integration of both targets into one perceptual episode[25]. In this case, the poor report of subjective experience of T2 might be confounded by the presence of T1. Participants might report poor T2 visibility not because T2 was not vividly experienced, but because the experience of T1 in the same perceptual episode causes confusion. This issue was discussed at length in[11], but we return to the point, since it remains an important potential confound that is worth revisiting in the light of the new findings being presented in this paper.

We additionally note that there are an unusually small number of putative integrated percepts in the experiment of[11]. The colour marking of T1 in this experiment reduced the classical indicator of integrated percepts, order errors, from 30% in classic letters/digits tasks[38] to approximately 10% in the task from[11]. Further, we note that

the pattern of behaviour we see at Lag 1, with low subjective experience and high accuracy is also visible to a lesser extent at lags 2 and 3, in which there are intervening distractors.

Another important point that stands against an integrated percepts explanation is the evidence that the reduction in relative subjective visibility can also be observed at Lag 2, and perhaps also weakly at Lag 3. The interaction analysis in[11] showed this, and the state-trace analysis we performed in this paper, suggested a non-monotonic pattern was still found in the replication (letters-in-digits) task when Lag 1 was removed. The integration argument is though classically ascribed specifically to Lag 1 and not later lags, in which there are intervening distractors. A further reason for believing that perceptual integration is unlikely to explain our findings is that it seems T1 is immune to the decoupling of report accuracy and subjective visibility, a point we discuss next.

Target Specificity of Decoupling

Importantly, the replication (letters-in-digits) data set that we analyse in this paper strengthens the specificity of the argument we are able to make. This further data set has enabled us to, firstly, replicate the decoupling between report accuracy and subjective visibility for T2. This was done with the state-trace analysis of T2 reported in subsection "Replication (letters-in-digits) Data" of section "Results". In addition,[33] reports the classic T2 interaction between Report Measure (report accuracy vs subjective visibility) and Lag for the letters-in-digits data set, which we reported in[11] for the original (colour-marked T1) data set.

Secondly, and perhaps most significantly, while subjective visibility ratings for T1 were not collected in the original (colour-marked T1) data set, the replication data set has that data point. As a result, we have been able to investigate whether there is a dissociation of report accuracy and subjective visibility for T1; and, importantly, there does not seem to be one.[33] failed to find an interaction between Report Measure (report accuracy vs subjective visibility) and Lag, and, in this paper, we identified a monotonic state-trace pattern for T1 in the replication data set; see subsection "Replication (letters-in-digits) Data" and figure 10.

The immunity of T1 to the report accuracy – subjective visibility dissociation suggests that the relationship between working memory encoding and conscious perception is unchanged across lags, and, notably, that co-activation of T1 with T2 (as occurs at very short lags) does not impair the conscious experience of T1, in the way it does T2. This finding is wholly consistent with the serial experience interpretation we are arguing for in this paper. That is, at very short lags, particularly Lag 1, T1 typically starts being perceived before T2 does, conferring it occupancy of the exclusive "focus of conscious experience", and the, late coming, T2 is excluded. This manifests in a, relative (to report accuracy), loss of visibility for T2, but not for T1, which is what we observe. In other words, the T1 claims "the brain's experiencer" before T2 arrives, and holds it until T2 has decayed, but there is no such exclusivity to the encoding into working memory.

This T1 immunity to the report accuracy – visibility dissociation also stands against a perceptual/ event integration interpretation. This is because, at its very heart, event integration suggests a composite of T1 and T2 is experienced. But, if that were the case, one would surely expect any impairment in T2 visibility associated with that composite, to also impact T1, In other words, if one is going to argue that T2 subjective visibility being low at Lag 1 is due to a confused "joint" binding, why would that decoupling of subjective visibility and report accuracy not also impact T1?

Simultaneous Type/Serial Token Model

There is no certainty with regard to an explanation of data such as we are presenting in this paper, but a computational account is as good a demonstration as one can have that a group of theoretical positions are consistent with each other, since a computational model has to run and generate this range of phenomena. Thus, we would argue that the STST computational account and the extension of it in the current paper is the demonstration that the theoretical positions we are taking are reconcilable. In particular, this shows that the subjective visibility findings we have named the Experiential Blink are reconcilable with the STST computation model, in particular, additions to the simultaneous type/serial token (STST) model of temporal attention allow it to index subjective experience as well as report accuracy, with the goal of providing a model that can explore the dissociations we discuss in this paper. In order to verify this model, we compared its predictions with the human data from[11]. The first comparison we made is between the behavioural results, specifically, we compare the respective report accuracies and subjective visibilities predicted by the SESE model to those from the human data. The results from this can be seen in figures 11(A) and 11(B). Overall, there is a strong similarity between the two. One notable difference is that the SESE model is simulating a slightly more difficult task than the human data – report accuracy lower by around 10%. Perhaps because of this, the SESE model also demonstrates a more marked downturn in subjective report at earlier lags than the human data.

We also compared the virtual ERPs generated by the SESE model with the human ERP data. The most significant difference between the two is the respective late dynamics of SESE compared to the human data, with the SESE data ERPs showing differences to the human data from approximately 600ms onward. Despite this, there is still a strong qualitative fit between the SESE data and the human data. It is important to note that we have taken the STST model exactly as it was formulated over 10 years ago, i.e. in[10]. Most notably, we have not refitted the parameters of the model in order to improve the match to the experimental data presented in this paper. This surely means that the match between model and experimental data is not going to be quantitatively perfect. In this respect, it is perhaps only reasonable to just expect a qualitative match between model and experimental results. In this context, the quality of match to the empirical data is, we would argue, impressive. Most importantly, the simulations we have run with SESE have provided a proof of principle that the explanation presented in figure 5 for why report accuracy and subjective visibility diverge is tenable. This explanation rests on the concept that encoding into working memory can proceed in parallel, but conscious perception cannot, a concept which we have noted suggests a theory called simultaneous encoding/serial experience. The natural electrophysiological correlate of this is a time-extended P3 when both T1 and T2 are consciously perceived, as opposed to just T1. This is what we observe in our data, and simulations in figure 12.

It is also important to observe that without a full investigation of the range of input strengths and parameter values within the STST family of models, the full range of patterns of data that can be embraced by the SESE model is not certain. For example, in its current configuration, the model generates very low visibility at lag-1 (see figure 11), which seems inconsistent with the observation that subjective visibility can exhibit sparing at lag-1, just substantially less than observed for report accuracy; see figure 6B). However, within the STST family of models, there may be a region of parameter settings that enable weak sparing for visibility at lag-1. In particular, the model is on something of a "knife-edge" at lag-1 and small changes in input strength and parameter settings can greatly change the model's behaviour.

One possible way in which sparing could be obtained for visibility would be if the T1 activation trace were high amplitude but short in duration, only excluding perception of T2 for a short period and thereby enabling it to be seen relatively vividly. If this were accompanied by very weak activation traces for T2 during the blink, weak lag-1 sparing of visibility may be obtainable. In this respect, aspects of the eSTST model[23] could be relevant, since they enable a more marked difference in dynamics between sparing and the blink. These aspects ensure that it is hard to reactivate the blaster (STST's attentional enhancement) once a blink has been initiated, naturally leading to weak T2 activation traces at lags 2 and 3. This said, modelling sparing of visibility at lag-1 is likely, at the least, to require retuning of STST's parameters, a step we have avoided to date.

A potentially far-reaching claim of the SESE model is that the generation of P3s is more involved than previously proposed (see[31]) for STST. We are not in a position to completely define this approach with full neural detail; that has to await further work. However, the new interpretation is required in order to be consistent with the results we present here and particularly in[11]. Specifically,[11] suggests that the P3 indexes conscious perception, not working memory encoding, so if we are proposing seriality of conscious perception, we have to propose seriality of the P3. Although a definitive mechanistic explanation awaits further modelling work, the intuition is that the activation traces currently generated by STST (which aggregate across a number of layers of the model) are precursors to the actual P3 and are earlier in the processing pathway. These activation traces feed into our "readout" mechanism, which is serial, excluding the second target from contributing to the P3 until the first has completed being experienced, i.e. has dropped below threshold.

Thus, we are imagining that the activation traces for T1 and T2 that the original STST model generate remain unchanged and can unfold in parallel, as they currently do at lag-1. Working memory encoding is still driven from these traces, but conscious experience is driven by the traces read-out, a notion that could be related to ideas of self-observation prominent in theories of conscious experience[39–41]. This readout enhancement can be considered speculative at this point. However, we include the idea here, since one purpose of theory is to provide strong claims that empirical work can attempt to disprove. This is a classic example of a scientific prediction that would be considered unlikely unless one subscribes to the theoretical position associated with the SESE theory. These are exactly the predictions that can carry the most evidence if experimentally investigated.

Indeed, it is central to scientific progress that testable predictions are made from models, in order that formalised theories can be disproved, the key to scientific progress from a Popperian perspective. In this spirit, the SESE model that we have presented in this paper makes two particularaly strong claims. The first being that that the P3 at lag-1 does not have the form of a double-amplitude single-target P3. Note, the vanilla STST, without readout-enhancement, does generate a double-amplitude P3 at lag-1, see figure 7 of[31]. Critically, it is important to rule out the possibility that the observed lag-1 P3 is reduced in amplitude because it is at ceiling.

That is, the specific prediction is that the lag-1 P3 is a similar amplitude to a single-target P3 and the distribution of P3s observed is not skewed according to a ceiling effect. The second key prediction that the SESE P3 readout mechanism predicts is that the steady state visual evoked potential (SSVEP) weakens or even de-synchromises during the P3. This is because if one asserts that an ongoing P3 for a target excludes the activation trace for another target, it should also exclude or dampen the activation traces of distractors (which drive the steady state response). Clearly, the SSVEP is at least partially from generators substantially earlier in the processing pathway than those that might directly drive the P3. Nonetheless, some sort of reduction in the power of the SSVEP may be observable. Disproving the first of these predictions would be a major problem for the readout-enhanced STST theory. Finding evidence for the second would provide converging evidence for the theory.

Seriality and STST

It is important to clarify the STST theory in the light of the findings and the serial experience ideas presented here. The following are key points to consider.

1. The original STST theory already makes a seriality assertion[10]. This, though, is a seriality over a longer time-frame than we are considering in this paper. That is, it proposes that the attentional blink has the role of delaying the start of a second episode, in order that all the bindings associated with a first episode can be completed before the next one starts. Thus, the seriality it focusses on is "across" the attentional blink, e.g. between a T1 and a T2 at, say, lag 5. As currently framed, it is focussed on working memory encoding, and does not explicitly speak to conscious experience.

2. The seriality considered in the current paper, is focussed on what happens when targets are very close together in time, e.g. at Lag 1. The original STST theory presented in[10] incorporated the notion of a "joint encoding" at Lag 1, whereby both T1 and T2 can be encoded into WM, but with a loss of episodic information, e.g. order and conjunction properties. The Experiential blink and the experience read-out theory presented in this paper extends the "joint encoding" notion from the original STST model, by arguing that there can be "joint encodings", but for T2 to be experienced, it has to be sufficiently strong that it can outlive the experiencing of T1. This is a new idea to the STST framework. The serialising considered here is specifically about conscious experience (the serialising of point 1. above is about working memory encoding), and it specifically occurs within a single episode, not across them.

Conclusion

We have examined the evidence for a dissociation between working memory encoding and subjective report in the attentional blink, and developed our own additions to current state-trace methodology. Our data stands clearly for a dissociation between working memory encoding and subjective report, and examining the data shows that this is the result of an increase in accuracy and a decrease of subjective visibility at lags 1, 2 and 3. Overall, we may have found evidence for a case in which it is possible to encode a stimulus into working memory without consciously perceiving it, a phenomenon we call sight-blind recall; however, a good deal more evidence needs to be acquired before this claim can be made with confidence. The SESE model is consistent with findings from human participants, and the results of the state trace analysis of this current work. However, more work will be required to determine the further predictions that the SESE model makes, and the sparseness of literature with respect to the experiential blink will require further experimentation to validate the predictions presented in this paper and those that will emerge. In particular, although there are a number of competing explanations of the decoupling of report accuracy and subjective visibility we observe (see[11] for a detailed consideration of many of these), evidence for the capacity to encode in parallel and experience in sequence is accumulating.

References

1. Block, N. How many concepts of consciousness? *Behavioral and brain sciences* **18**, 272–287 (1995).
2. Sperling, G. The information available in brief visual presentations. *Psychological monographs: General and applied* **74**, 1 (1960).
3. Vandenbroucke, A. R., Fahrenfort, J. J., Sligte, I. G. & Lamme, V. A. Seeing without knowing: neural signatures of perceptual inference in the absence of report. *Journal of cognitive neuroscience* **26**, 955–969 (2014).

4. Bronfman, Z. Z., Brezis, N., Jacobson, H. & Usher, M. We see more than we can report "cost free" color phenomenality outside focal attention. *Psychological science* **25**, 1394–1403 (2014).

5. Sligte, I. G., Scholte, H. S. & Lamme, V. A. Are there multiple visual short-term memory stores? *PLOS one* **3**, e1699 (2008).

6. Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in cognitive sciences* **10**, 204–211 (2006).

7. De Gardelle, V., Sackur, J. & Kouider, S. Perceptual illusions in brief visual presentations. *Consciousness and cognition* **18**, 569–577 (2009).

8. Sergent, C. & Dehaene, S. Is consciousness a gradual phenomenon? evidence for an all-or-none bifurcation during the attentional blink. *Psychological science* **15**, 720–728 (2004).

9. Raymond, J. E., Shapiro, K. L. & Arnell, K. M. Temporary suppression of visual processing in an rsvp task: An attentional blink? *Journal of experimental psychology: Human perception and performance* **18**, 849 (1992).

10. Bowman, H. & Wyble, B. The simultaneous type, serial token model of temporal attention and working memory. *Psychological review* **114**, 38–70 (2007). LR: 20070117; CI: ((c) 2007; JID: 0376476; ppublish.

11. Pincham, H. L., Bowman, H. & Szucs, D. The experiential blink: Mapping the cost of working memory encoding onto conscious perception in the attentional blink. *Cortex* **81**, 35–49 (2016).

12. Warrington, E. K. The double dissociation of short-and long-term memory. *Human Memory and Amnesia (PLE: Memory)* **4**, 61 (2014).

13. Cousins, K. A., York, C., Bauer, L. & Grossman, M. Cognitive and anatomic double dissociation in the representation of concrete and abstract words in semantic variant and behavioral variant frontotemporal degeneration. *Neuropsychologia* **84**, 244–251 (2016).

14. Cohen, M. A., Cavanagh, P., Chun, M. M. & Nakayama, K. The attentional requirements of consciousness. *Trends in cognitive sciences* **16**, 411–417 (2012).

15. Hsieh, P.-J., Colas, J. T. & Kanwisher, N. Pop-out without awareness unseen feature singletons capture attention only when top-down attention is available. *Psychological science* (2011). Pmid:21852451.

16. den Bussche, E. V., Hughes, G., Humbeeck, N. V. & Reynvoet, B. The relation between consciousness and attention: An empirical study using the priming paradigm. *Consciousness and cognition* **19**, 86–97 (2010).

17. Marshall, J. C. & Halligan, P. W. Blindsight and insight in visuo-spatial neglect (1988).

18. Heathcote, A., Freeman, E., Etherington, J., Tonkin, J. & Bora, B. A dissociation between similarity effects in episodic face recognition. *Psychonomic bulletin & review* **16**, 824–831 (2009).

19. Bogartz, R. S. On the meaning of statistical interactions. *Journal of experimental child psychology* **22**, 178–183 (1976).

20. Dunn, J. C. & Kirsner, K. Discovering functionally independent mental processes: The principle of reversed association. *Psychological review* **95**, 91 (1988).

21. Davis-Stober, C. P., Morey, R. D., Gretton, M. & Heathcote, A. Bayes factors for state-trace analysis. *Journal of mathematical psychology* **72**, 116–129 (2016).

22. Prince, M., Brown, S. & Heathcote, A. The design and analysis of state-trace experiments. *Psychological methods* **17**, 78 (2012).

23. Wyble, B., Bowman, H. & Nieuwenstein, M. The attentional blink provides episodic distinctiveness: sparing at a cost. *Journal of Experimental Psychology: Human Perception and Performance* **35**, 787 (2009).

24. Nieuwenhuis, S. & de Kleijn, R. Consciousness of targets during the attentional blink: a gradual or all-or-none dimension? *Attention, Perception, & Psychophysics* **73**, 364–373 (2011).

25. Simione, L., Akyrek, E. G., Vastola, V., Raffone, A. & Bowman, H. Illusions of integration are subjectively impenetrable: Phenomenological experience of lag 1 percepts during dual-target rsvp. *Consciousness and cognition* **51**, 181–192 (2017).

26. Newell, B. R. & Dunn, J. C. Dimensions in data: Testing psychological models using state-trace analysis. *Trends in cognitive sciences* **12**, 285–290 (2008).

27. Loftus, G. R., Oberg, M. A. & Dillon, A. M. Linear theory, dimensional theory, and the face-inversion effect. *Psychological review* **111**, 835 (2004).

28. Sense, F., Morey, C. C., Prince, M., Heathcote, A. & Morey, R. D. Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior research methods* **49**, 853–862 (2017).

29. Lindley, D. V. A statistical paradox. *Biometrika* **44**, 187–192 (1957).

30. Wyble, B. & Bowman, H. Computational and experimental evaluation of the attentional blink: Testing the simultaneous type serial token model. In *CogSci*, 2371–2376 (2005).

31. Craston, P., Wyble, B., Chennu, S. & Bowman, H. The attentional blink reveals serial working memory encoding: Evidence from virtual and human event-related potentials. *Journal of cognitive neuroscience* **21**, 550–566 (2009).

32. Roberts, S. & Pashler, H. How persuasive is a good fit? a comment on theory testing. *Psychological review* **107**, 358 (2000).

33. Gootjes-Dreesbach, E. L. *Awareness & Perception in Rapid Serial Visual Presentation*. Ph.D. thesis, University of Kent, (2015).

34. Tulving, E. Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior* **20**, 479–496 (1981).

35. Soto, D. & Silvanto, J. Reappraising the relationship between working memory and conscious awareness. *Trends in cognitive sciences* **18**, 520–525 (2014).

36. Bergström, F. & Eriksson, J. Maintenance of non-consciously presented information engages the prefrontal cortex. *Frontiers in human neuroscience* **8**, 938 (2014).

37. Lau, H. C. & Passingham, R. E. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences* **103**, 18763–18768 (2006). Pmid:17124173.

38. Chun, M. M. & Potter, M. C. A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental psychology: Human perception and performance* **21**, 109 (1995).

39. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Frontiers in human neuroscience* **8**, 443 (2014).

40. Lau, H. & Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences* **15**, 365–373 (2011).

41. Cleeremans, A. Connecting conscious and unconscious processing. *Cognitive science* **38**, 1286–1315 (2014).

42. Morris, C. N. Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association* **78**, 47–55 (1983).

Acknowledgements

We thank Zoltan Dienes for pointing us in the direction of reversed associations, and Andrew Heathcote, Brad Wyble, and Marius Usher for their valuable insights into our work. We also thank Denes Szucs for his assistance and supervision in the collection of the data used in this paper. Finally, we thank two anonymous referees, who gave extensive and insightful suggestions on this paper, which have enabled us to greatly improve it.

Author contributions statement

H.B and W.J designed research, H.P and E.L.G.D collected data, W.J performed research, analysed data and wrote paper. H.B. worked on drafts of this paper.

Additional information

The author(s) declare no competing interests.

Supplementary Information

Section A - Extensions to the state-trace method

Previous state-trace analysis has generally been in a position to make strong statements about the ordinal relationships of the variables for which the measures of interest (e.g., accuracy and visibility) are calculated, allowing them to make strong statements with their priors. For example, in their experiment on short term memory[21] are able to a-priori assume in their data that accuracy in a change detection task is higher when participants have the opportunity to verbalise the first target than when they did not. In comparison, while we have strong expectations about some behaviours of the attentional blink such as lag 1 sparing in letters-in-digits tasks[10,23,38], the variability in, for example, depth of the blink between experiments, means we are not in a position to make such strong ordinal statements as these previous works. We therefore propose a data driven method that makes use of an orthogonal measure to the monotonicity contrast. This method takes two sets of a-priori "constraints" on the data, restrictions on potential orderings in the prior entered into Bayesian inference. These are an "irrevocable" set containing those constraints that no theorist would believe violable, and for which any evidence against can only be considered a measurement error - for example, we would expect lag 1 accuracy to be larger than lag 2 accuracy at the participant level in the letters-in-digits attentional blink - and a "free" set encoding those behaviours that we might expect to change between experiments – for example, the lowest point in the blink. Orderings of the dimension (or trace) factor that do not fit the constraints are considered a-priori to have a prior probability of 0, with all other orderings equally likely. Our method then removes constraints from the free set that do not fit the data on the basis of our orthogonal measure of validity. The result is a theoretically grounded, empirically derived set of constraints on the data.

This orthogonal measure is a dimension vs non-dimension factor, analogous to and intersecting with, the trace vs non-trace factor used in[21]. In the same way as this trace vs non-trace factor, this gives us a measure of how accurately the data conforms to a given set of ordering constraints across both the dimension and trace factors. We call this measure $BF_{(D\&T)/N(D\&T)}$, or when no trace factor is present such as in the main body of this paper, as $BF_{D/N(D)}$ in order to prevent confusion about the trace factor that does not exist in our analysis. In the case in which the trace factor has only one level (such as in our data), this measure is also equivalent to how well the data conforms to exclusively the dimension axis versus how well it does not. This measure specifically quantifies the ratio of evidence for the intersection of both the trace and dimension constraints versus all other points, thereby providing a measure of validity that the overall set of constraints we select fit our data.

In order to make use of this measure to derive a prior, we first pick a set of order constraints on the state and dimension axes from prior data, $C = \{c_1, \ldots, c_n\}$. This set of constraints should be the fullest set that can be reasonably expected to fit the data, but should not contain constraints that contradict one another. We then divide this set C into two subsets, those constraints in C for which violation can only constitute a measurement error (the irrevocable set), and those about which we might expect variation between experiments (the free set). We label these $E = \{e_1, \ldots, e_l\}$ and $F = \{f_1, \ldots, f_q\}$ respectively. Next, we introduce the concept of group validity for a given set of constraints, denoted GE. This is the product of $BF_{(D\&T)/N(D\&T)}$ across all our M participants for the set of constraints C, specifically:

$$GE(C) = \prod_{i=1}^{M} BF_{(D\&T)/N(D\&T)_i}$$

For each item in F, we denote the "leave one out" subset of constraints ($\overline{F}_j$) as:

$$\overline{F}_j = E \cup \{f_1, \ldots, f_{j-1}, f_{j+1}, \ldots, f_q\}$$

We then calculate $GE(\overline{F}_j)$ for all $j \in q$. For the largest evaluated $GE(\overline{F}_j)$ with $GE(\overline{F}_j) > GE(E \cup F)$, we then remove f_j from F. This procedure is repeated on the new F with f_j removed until there does not exist a set such that $GE(\overline{F}_j) > GE(E \cup F)$, or until $F = \{\}$. The resulting $E \cup F$ is the "empirical prior". We note that this method is very similar in its essence to the parametric empirical Bayes (PEB) method[42], however, we note that the specifics of our application allow us to solve the problem in a greatly simplified manner.

Our method is justified as follows. Firstly, it is clear that setting our empirical prior based on $BF_{(D\&T)/N(D\&T)}$ will, on its own, converge to a prior set of constraints that best fit the data. Secondly, since we are starting from the fullest (strictest) set of constraints that are theoretically grounded and pruning from this set, it is impossible for us to introduce spurious constraints that fit the data by chance, but are incompatible with our theoretical understanding. Equally, because we hold some constraints "irrevocable" we are protected from removing constraints that are highly likely a-priori, based on measurement errors. Finally, $BF_{(D\&T)/N(D\&T)}$ is an orthogonal measure to the $BF_{(M/MN)|/(D\&T)}$. Since $M|(D\&T) \cup NM|(D\&T) \subseteq D\&T$ (the union of the

monotonic and non-monotonic orderings given some set of constraints is contained inside the set of all possible orderings given those constraints) the changes in the balance of probabilities between $M|(D\&T)$ and $NM|(D\&T)$ (calculated as $BF_{(M/MN)|/(D\&T)}$ have no effect on the respective probabilities of a given set of constraints $D\&T$ versus their complement $N(D\&T)$.

Section B - Lag 1 as a cause of non-monotonicity in the original colour-marked T1 task

In the main body of the paper, we find evidence for a strongly non-monotonic relationship between accuracy and subjective report in the original colour-marked T1 task. As noted in[11], this appears to be driven by differences in the behaviour at early lags, particularly Lag 1. Here, we attempt to quantify this effect by removing Lag 1 from the state-trace analysis, and examining how it changes. As well as removing the lag from the dataset, we must also adjust our constraints. The strongest performance was on the empirically derived constraints, so for this analysis we use these, minus any constraints on the lag 1 datapoint that are now no longer applicable. We find that, despite the fact that grouped (not log) evidence is almost completely unchanged ($BF_{D/N(D)} = 1.07 \times 10^{13}$ with lag 1, $BF_{D/N(D)} = 1.01 \times 10^{13}$ without), our grouped (not log) bayes factor changes from extremely strong evidence for non-monotonicity at $BF_{(M/NM)|D} = 1.17 \times 10^{-17}$, to no strong evidence either way . The results for each subject individually can be seen in figures 14 and 15. From this we conclude that Lag 1 is a strong driver of the effect of non-monotonicity that we see in our state-trace analysis of the original colour-marked T1 task. However, the situation changes for the replication letters-in-digits experiment.

Section C - Binning Method for high vs low visibility trials

In order to determine which binning method was appropriate for separating the data from[11] into high and low visibility trials, we evaluated the grouped validity for each potential binning method. This showed quite clearly (see figure 16) that the split with the strongest validity was an even split with the 3 lowest visibility ratings forming the low bin, and the 3 highest visibility ratings forming the high bin.

Section D - Subjective experience in the Simultaneous Type/Serial Token model

In this section, we detail how the STST model is used to simulate ERPs, the setup of the STST model used to extract a visibility rating, and how the visibility rating was calculated. Our virtual ERPs are calculated from a computational implementation of the STST model, neural-STST[10,31]. As in the STST model described in the STST model section, the neural STST model is organised as layers of nodes, connected via weighted connections. These connections are the analogue of synaptic projections in the brain, and in order to calculate the P3, we therefore introduce the concept of excitatory post synaptic potential to these virtual nodes. This is calculated as the activation value of the node multiplied by the weight value of its connections to the subsequent layer. The virtual P3 is then calculated as the sum of these excitatory post synaptic potentials across a subset of the layers. We follow previous work in using the 3rd, 4th, 6th and 8th layers of the neural-STST model, corresponding to the item layer, the task filtered layer, the binder gates and the token gates. As in previous work[31], we also implement a retinal delay of a model equivalent of 70ms. Compared to previous works using virtual ERPs from the STST model, we selected a slightly different stimulus range over which to calculate this virtual P3. Specifically, we sample a range of stimulus strengths with greater variability (-0.078 to +0.078 -> -0.1625 to +0.1625), at a slightly higher average stimulus and distractor strength (0.520 -> 0.570). This approach is consistent with previous simulations with the STST model, where we allow input strength ranges to vary reflecting the fact that different experiments being modelled might have quite different stimulus types and sensitivities. Compared to previous iterations of virtual P3 generation, we do not directly sum the components of each item in the stream to create the P3. We instead only consider the contribution to the P3 of a target to the extent that it does not conflict with the P3 of another, active target.

In order to calculate subjective report from these virtual P3's we, as described in the main body of the paper, calculate the number of time steps that a stimulus spends above a given threshold. For the results given in this paper, this threshold is 0.05. Additionally, although this method gives us a continuous subjective report, for the purposes of comparison with the human data from[11], it is necessary to be able to divide these subjective reports into the discrete cases of high/low visibility. Since we are unable to be sure that each lag contains the full range of possible subjective reports, we do this by lag. Since we also do not know how the visibilities are distributed across each lag, but wish to make a simple, even split as far as possible, we use the average as the splitting point for high/low visibility. It is also necessary to normalise these time steps counts into visibility ratings that can be compared to the human data. In the spirit of the simplicity that has driven the creation of the model so far, we simply normalise the timesteps by a linear factor. To keep the range plausible and remain data driven, the

value we selected was the most visible stimulus in the entire experiment, and divided each visibility rating by this in order to give a "percentage visibility". In this way, we provide a very simple index of both continuous and binned subjective report that requires no changes to the original model.

Section E - Further SESE model ERPs

In figure 17 we provide some further results comparing human and SESE generated ERPs. This compares Human lag 3 with SESE lag 3, and human lag 3 with SESE lag 4.

Section F - Justifying Non-Monotonic Pattern in Figure 6

It is interesting to note that in the replication (pure letters-in-digits) data set, non-monotonicity goes up when lag-1 is removed: compare figures 8B and 9B. There are a number of points that can be made about this.

1. The identification of a non-monotonic pattern when lag-1 is excluded is not inconsistent with the attentional and experiential blink curves we observe for this data set – see figure 6B, where the distance between T2 report accuracy and T2 subjective visibility are further apart at lag-2 than at higher lags.

2. Non-monotonicity with lag-1 excluded is not so obvious from figure 6D, although, there is a definite kink for lag-2 relative to lags 3 and 5. Furthermore, small fluctuations in the lag-2 data point, which there certainly are across participants, could create a non-monotonic pattern driven by lag-2.

3. The importance of Lag 1 in the averaged data is not necessarily accurately reflecting each individual. Accordingly, removing the lag-1 point does not consistently effect each individual participant. Although the overall trend is for more evidence for non-monotonicity, 4 of the 12 participants, for example, gain evidence for monotonicity with the removal of the lag 1 data-point, compare figure 8B and figure 9B.

4. Finally, and perhaps most importantly, it is well attested that averaged state-trace curves can fail to be representative of the across participant pattern. Indeed, it could be that the lag-2 point is only at the position shown in figure 6D for the average and not for any of the participants.

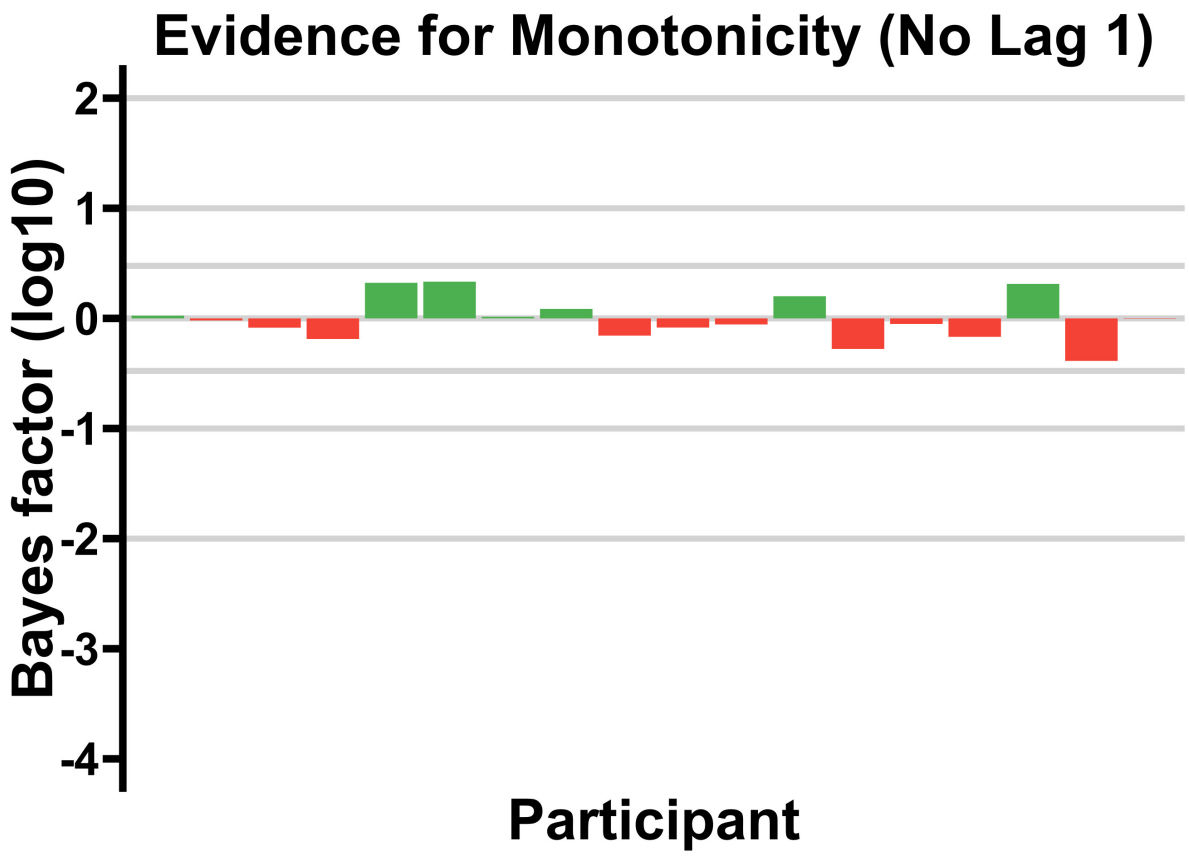


Figure 14. Respective monotonicity vs non-monotonicity for the original (colour-marked T1) dataset excluding the lag 1 datapoint. Results are weak and strongly hetrogenous, with grouped (not log) $BF_{(M/NM)|D} = 6.69 \times 10^{-1}$. This essentially provides no evidence either way for monotonicity, a strong constrast to the analysis with the Lag 1 datapoint included, which finds a strongly non-monotonic effect with grouped (not log) $BF_{(M/NM)|D} = 1.17 \times 10^{-17}$

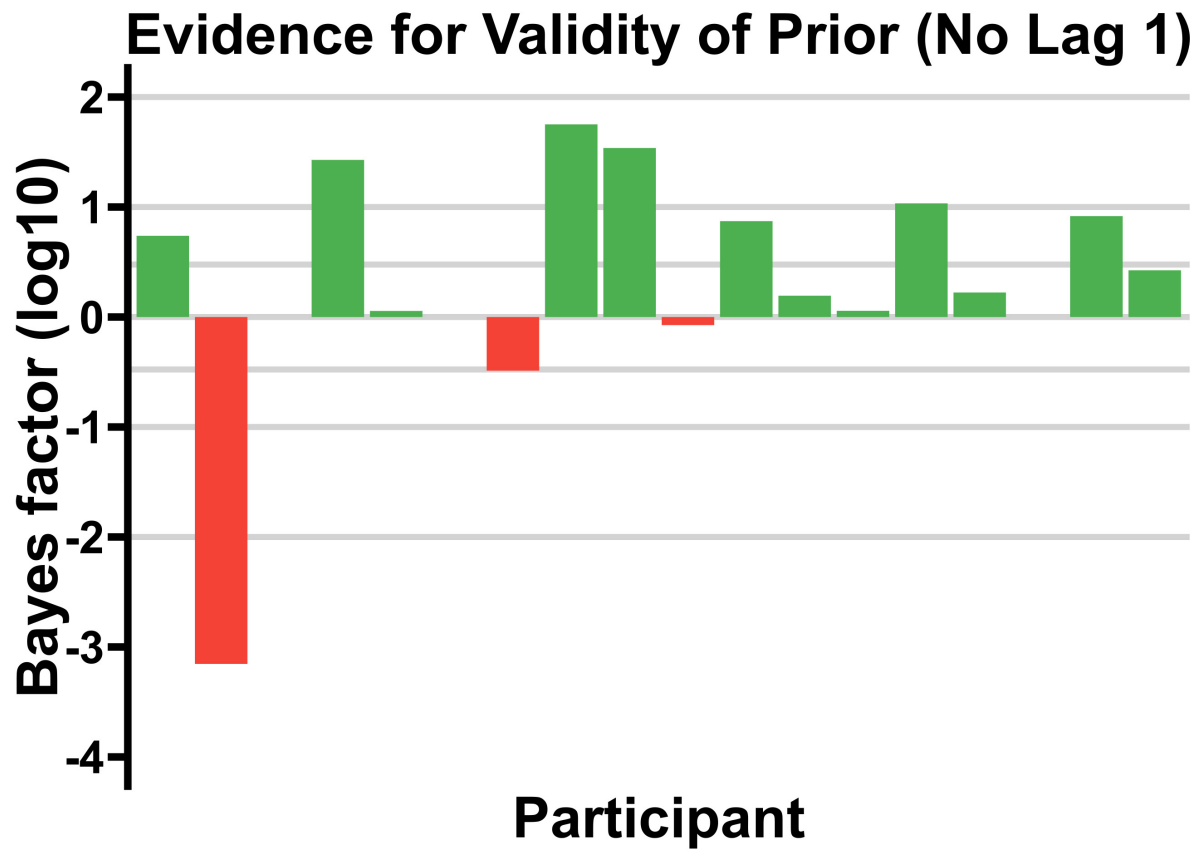


Figure 15. Validity for each participant for the set of prior constraints derived from the original using our empirical prior method, for the original (colour-marked T1) task dataset excluding the Lag 1 datapoint. Any constraints no longer valid without Lag 1 have been removed. At the group level, the evidence is strongly in favour of the constraints fitting the data, with grouped (not log) $BF_{D/N(D)} = 1.01 \times 10^{13}$, extremely close to the grouped validity with Lag 1 included with grouped (not log) ($BF_{D/N(D)} = 1.07 \times 10^{13}$.

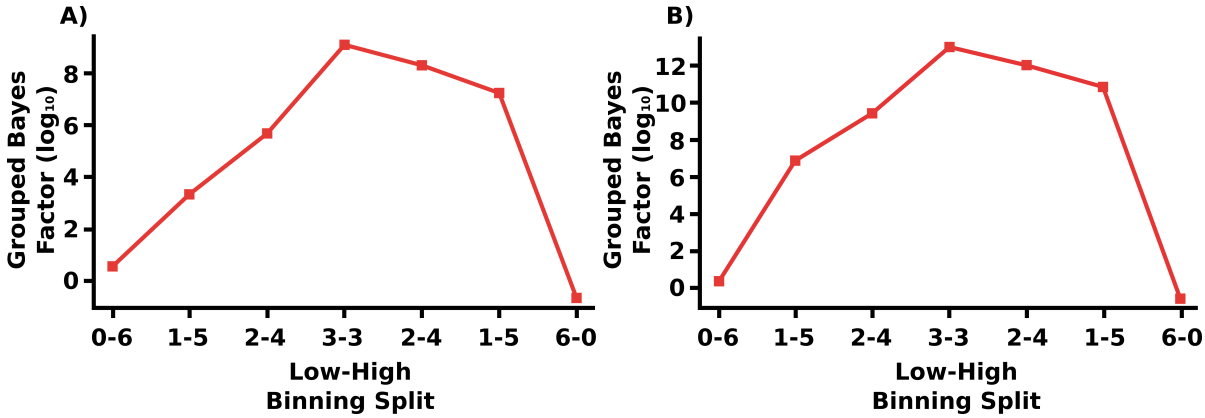


Figure 16. A) Grouped Bayes factor for validity of the original (colour-marked T1) dataset across each potential binning method for high and low visibility using the original set of constraints based on the data from (Nieuwenhuis, de Kleijn 2011). B) Grouped Bayes factor for validity across each potential binning method for the empirical prior constraints.

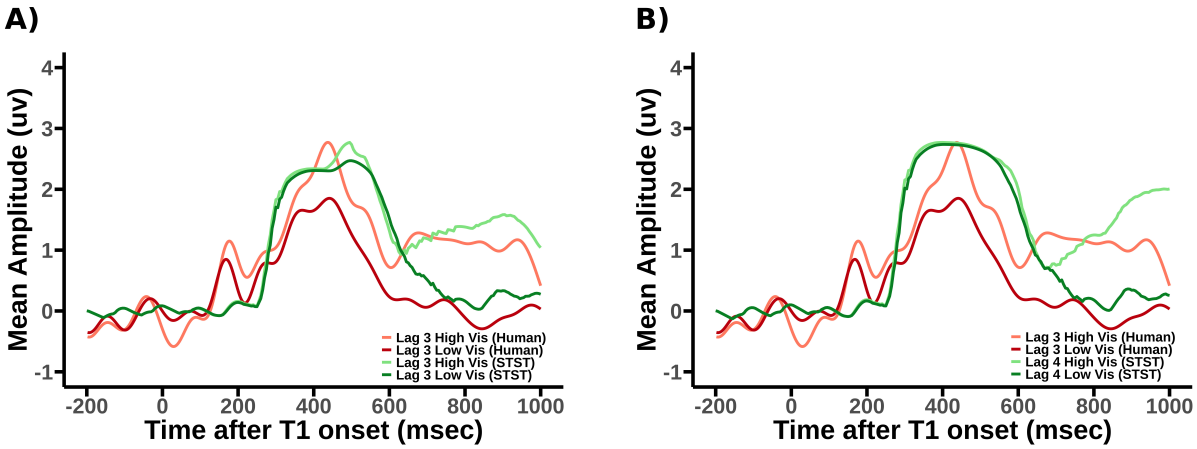


Figure 17. A comparison, for both high and low T2 visibility, given correctly reported T1, of the human ERPs from the original colour-marked T1 data analysis[11]. A) Lag 3 Human ERPs vs Lag 3 STST virtual ERPs. B) Lag 3 Human ERPs vs Lag 4 STST virtual ERPs. Importantly, as previously discussed, neither the function or the structure of the STST model, as given in[10] were changed when generating the virtual P3s. Note that the human ERPs presented are slightly different to those from[11], as ours exclude order errors to be consistent with previous sections.

References

ANTONY, M.V., 2001. Is 'consciousness' ambiguous? *Journal of Consciousness Studies,* **8**(2), pp. 19-44.

ANTOS, A. and KONTOYIANNIS, I., 2001. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms,* **19**(3-4), pp. 163-193.

AREND, I., JOHNSTON, S. and SHAPIRO, K., 2006. Task-irrelevant visual motion and flicker attenuate the attentional blink. *Psychonomic bulletin & review,* **13**(4), pp. 600-607.

ASHBY, F.G. and SOTO, F.A., 2015. Multidimensional signal detection theory. *Oxford handbook of computational and mathematical psychology,* , pp. 13-34.

BAARS, B.J., 1997. In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies,* **4**(4), pp. 292-309.

BAMBER, D., 1979. State-trace analysis: A method of testing simple theories of causation. *Journal of mathematical psychology,* **19**(2), pp. 137-181.

BARRETT, A.B., DIENES, Z. and SETH, A.K., 2013. Measures of metacognition on signal-detection theoretic models. *Psychological methods,* **18**(4), pp. 535.

BATES, D., SARKAR, D., BATES, M.D. and MATRIX, L., 2007. The lme4 package. *R package version,* **2**(1), pp. 74.

BEIRLANT, J., DUDEWICZ, E.J., GYÖRFI, L. and VAN DER MEULEN, EDWARD C, 1997. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences,* **6**(1), pp. 17-39.

BERGSTRM, F. and ERIKSSON, J., 2014. Maintenance of non-consciously presented information engages the prefrontal cortex. *Frontiers in human neuroscience,* **8**, pp. 938.

BLOCK, N., 2011. Perceptual consciousness overflows cognitive access. *Trends in cognitive sciences,* **15**(12), pp. 567-575.

BLOCK, N., 2007. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and brain sciences,* **30**(5-6), pp. 481-499.

BLOCK, N., 1995. How many concepts of consciousness? *Behavioral and brain sciences,* **18**(02), pp. 272-287.

BOGARTZ, R.S., 1976. On the meaning of statistical interactions. *Journal of experimental child psychology,* **22**(1), pp. 178-183.

BOGEN, J.E., 1997. An example of access-consciousness without phenomenal consciousness? *Behavioral and Brain Sciences,* **20**(1), pp. 144.

BOWMAN, H. and WYBLE, B., 2007. The simultaneous type, serial token model of temporal attention and working memory. *Psychological review,* **114**(1), pp. 38-70.

BOWMAN, H., WYBLE, B., CHENNU, S. and CRASTON, P., 2008. A reciprocal relationship between bottom-up trace strength and the attentional blink bottleneck: Relating the LC–NE and ST2 models. *Brain research,* **1202**, pp. 25-42.

BRAUN, J., 1998. Vision and attention: the role of training. *Nature,* **393**(6684), pp. 424.

BRONFMAN, Z.Z., BREZIS, N., JACOBSON, H. and USHER, M., 2014. We See More Than We Can Report "Cost Free" Color Phenomenality Outside Focal Attention. *Psychological science,* **25**(7), pp. 1394-1403.

CARRUTHERS, G. and SCHIER, E., 2017. Why are we still being hornswoggled? Dissolving the hard problem of consciousness. *Topoi,* **36**(1), pp. 67-79.

CARRUTHERS, P., 2006. *The architecture of the mind.* Clarendon Press.

CHALMERS, D.J., 1995. Facing up to the problem of consciousness. *Journal of consciousness studies,* **2**(3), pp. 200-219.

CHAO, A. and SHEN, T., 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics,* **10**(4), pp. 429-443.

CHUN, M.M. and POTTER, M.C., 1995. A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental psychology: Human perception and performance,* **21**(1), pp. 109.

COHEN, M.A., CAVANAGH, P., CHUN, M.M. and NAKAYAMA, K., 2012. The attentional requirements of consciousness. *Trends in cognitive sciences,* **16**(8), pp. 411-417.

CONRAD, R., 1964. Acoustic confusions in immediate memory. *British journal of Psychology,* **55**(1), pp. 75-84.

COUSINS, K.A., YORK, C., BAUER, L. and GROSSMAN, M., 2016. Cognitive and anatomic double dissociation in the representation of concrete and abstract words in semantic variant and behavioral variant frontotemporal degeneration. *Neuropsychologia,* **84**, pp. 244-251.

CRASTON, P., WYBLE, B., CHENNU, S. and BOWMAN, H., 2009. The attentional blink reveals serial working memory encoding: Evidence from virtual and human event-related potentials. *Journal of cognitive neuroscience,* **21**(3), pp. 550-566.

DAVIS-STOBER, C.P., MOREY, R.D., GRETTON, M. and HEATHCOTE, A., 2016. Bayes factors for state-trace analysis. *Journal of mathematical psychology,* **72**, pp. 116-129.

DAVIS-STOBER, C.P., MOREY, R.D., GRETTON, M. and HEATHCOTE, A., 2015. Bayes factors for state-trace analysis. *Journal of mathematical psychology,* .

DE GARDELLE, V., SACKUR, J.'.^. and KOUIDER, S., 2009. Perceptual illusions in brief visual presentations. *Consciousness and cognition,* **18**(3), pp. 569-577.

DEHAENE, S., 2014. *Consciousness and the brain: Deciphering how the brain codes our thoughts.* Penguin.

DEHAENE, S. and CHANGEUX, J., 2011. Experimental and theoretical approaches to conscious processing. *Neuron,* **70**(2), pp. 200-227.

DEHAENE, S. and CHANGEUX, J., 2005. Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *PLoS biology,* **3**(5), pp. e141.

DEHAENE, S., CHANGEUX, J., NACCACHE, L., SACKUR, J.'.^. and SERGENT, C., 2006. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in cognitive sciences,* **10**(5), pp. 204-211.

DEHAENE, S., KERSZBERG, M. and CHANGEUX, J., 1998. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences,* **95**(24), pp. 14529-14534.

DEHAENE, S., SERGENT, C. and CHANGEUX, J., 2003. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences,* **100**(14), pp. 8520-8525.

DENNETT, D., 2000. Facing backwards on the problem of consciousness. *Explaining Consciousness–The "Hard Problem,* , pp. 33-36.

DIENES, Z., 2015. Behavioural methods in consciousness research.

DIENES, Z. and SETH, A., 2010. Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and cognition,* **19**(2), pp. 674-681.

DUNN, J.C. and KIRSNER, K., 1988. Discovering functionally independent mental processes: The principle of reversed association. *Psychological review,* **95**(1), pp. 91.

FEINSTEIN, D.L., KALININ, S. and BRAUN, D., 2016. Causes, consequences, and cures for neuroinflammation mediated via the locus coeruleus: noradrenergic signaling system. *Journal of neurochemistry,* **139**, pp. 154-178.

FERLAZZO, F., LUCIDO, S., DI NOCERA, F., FAGIOLI, S. and SDOIA, S., 2007. Switching between goals mediates the attentional blink effect. *Experimental Psychology,* **54**(2), pp. 89-98.

FLEMING, S.M. and LAU, H.C., 2014. How to measure metacognition. *Frontiers in human neuroscience,* **8**, pp. 443.

FODOR, J.A., 1983. *The modularity of mind: An essay on faculty psychology.* MIT press.

FRANKLIN, S. and GRAESSER, A., 1999. A software agent model of consciousness. *Consciousness and cognition,* **8**(3), pp. 285-301.

FRISTON, K., 2012. Ten ironic rules for non-statistical reviewers. *NeuroImage,* **61**(4), pp. 1300-1310.

GALVIN, S.J., PODD, J.V., DRGA, V. and WHITMORE, J., 2003. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic bulletin & review,* **10**(4), pp. 843-876.

GARNER, W.R., HAKE, H.W. and ERIKSEN, C.W., 1956. Operationism and the concept of perception. *Psychological review,* **63**(3), pp. 149.

GILMORE, G.C., HERSH, H., CARAMAZZA, A. and GRIFFIN, J., 1979. Multidimensional letter similarity derived from recognition errors. *Perception & psychophysics,* **25**(5), pp. 425-431.

GRASSBERGER, P., 2003. Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138,* .

HACKER, P., 2010. Hacker's challenge. *The Philosophers' Magazine,* (51), pp. 23-32.

HAN, Y., JIAO, J. and WEISSMAN, T., 2015Adaptive estimation of Shannon entropy, *2015 IEEE International Symposium on Information Theory (ISIT)* 2015, IEEE, pp. 1372-1376.

HARRISON, S.A. and TONG, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. *Nature,* **458**(7238), pp. 632.

HAUSSER, J. and STRIMMER, K., 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research,* **10**(Jul), pp. 1469-1484.

HEATHCOTE, A., FREEMAN, E., ETHERINGTON, J., TONKIN, J. and BORA, B., 2009. A dissociation between similarity effects in episodic face recognition. *Psychonomic bulletin & review,* **16**(5), pp. 824-831.

HENSON, R., 2006. Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in cognitive sciences,* **10**(2), pp. 64-69.

HSIEH, P., COLAS, J.T. and KANWISHER, N., 2011. Pop-out without awareness unseen feature singletons capture attention only when top-down attention is available. *Psychological science,* .

HUMPHREY, N.K., 1974. Vision in a monkey without striate cortex: a case study. *Perception,* **3**(3), pp. 241-255.

JAMES, W., 1898. *Philosophical conceptions and practical results.* The University Press.

JIAO, J., VENKAT, K., HAN, Y. and WEISSMAN, T., 2017. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory,* **63**(10), pp. 6774-6798.

KALAT, J.W., 2014. No title. *Consciousness and the Brain: Deciphering How the Brain Codes our Thoughts,* .

KORNELL, N., SON, L.K. and TERRACE, H.S., 2007. Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science,* **18**(1), pp. 64-71.

KOUIDER, S., DE GARDELLE, V., SACKUR, J. and DUPOUX, E., 2010. How rich is consciousness? The partial awareness hypothesis. *Trends in cognitive sciences,* **14**(7), pp. 301-307.

KRANCZIOCH, C., DEBENER, S., MAYE, A. and ENGEL, A.K., 2007. Temporal dynamics of access to consciousness in the attentional blink. *NeuroImage,* **37**(3), pp. 947-955.

KRANTZ, D.H. and TVERSKY, A., 1971. Conjoint-measurement analysis of composition rules in psychology. *Psychological review,* **78**(2), pp. 151.

KRUEGER, C. and TIAN, L., 2004. A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological research for nursing,* **6**(2), pp. 151-157.

LAMME, V.A., 2001. Blindsight: the role of feedforward and feedback corticocortical connections. *Acta Psychologica,* **107**(1-3), pp. 209-228.

LAMONT, C.H. and WIGGINS, P.A., 2016. The Lindley paradox: The loss of resolution in Bayesian inference. *arXiv preprint arXiv:1610.09433,* .

LAMY, D., SALTI, M. and BAR-HAIM, Y., 2009. Neural correlates of subjective awareness and unconscious processing: an ERP study. *Journal of cognitive neuroscience,* **21**(7), pp. 1435-1446.

LAU, H.C. and PASSINGHAM, R.E., 2006. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences,* **103**(49), pp. 18763-18768.

LINDLEY, D.V., 1957. A statistical paradox. *Biometrika,* **44**(1/2), pp. 187-192.

LIU, C.C. and AITKIN, M., 2008. Bayes factors: Prior sensitivity and model generalizability. *Journal of mathematical psychology,* **52**(6), pp. 362-375.

LOFTUS, G.R., 2002. Analysis, interpretation, and visual presentation of experimental data. *Stevens' handbook of experimental psychology,* .

LOFTUS, G.R., 1978. On interpretation of interactions. *Memory & cognition,* **6**(3), pp. 312-319.

LOFTUS, G.R., OBERG, M.A. and DILLON, A.M., 2004. Linear theory, dimensional theory, and the face-inversion effect. *Psychological review,* **111**(4), pp. 835.

MANISCALCO, B. and LAU, H., 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition,* **21**(1), pp. 422-430.

MARSHALL, J.C. and HALLIGAN, P.W., 1988. Blindsight and insight in visuo-spatial neglect.

MARTENS, S., MUNNEKE, J., SMID, H. and JOHNSON, A., 2006. Quick minds don't blink: Electrophysiological correlates of individual differences in attentional selection. *Journal of cognitive neuroscience,* **18**(9), pp. 1423-1438.

MARTENS, S. and WYBLE, B., 2010. The attentional blink: Past, present, and future of a blind spot in perceptual awareness. *Neuroscience & Biobehavioral Reviews,* **34**(6), pp. 947-957.

MASSON, M.E. and ROTELLO, C.M., 2009. Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **35**(2), pp. 509.

MILLER, G., 1955. Note on the bias of information estimates. *Information theory in psychology: Problems and methods,* .

MISRA, N., SINGH, H. and DEMCHUK, E., 2005. Estimation of the entropy of a multivariate normal distribution. *Journal of multivariate analysis,* **92**(2), pp. 324-342.

MOREY, R.D., PRATTE, M.S. and ROUDER, J.N., 2008. Problematic effects of aggregation in z ROC analysis and a hierarchical modeling solution. *Journal of mathematical psychology,* **52**(6), pp. 376-388.

NELSON, T.O., 1984. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological bulletin,* **95**(1), pp. 109.

NEMENMAN, I., 2011. Coincidences and estimation of entropies of random variables with large cardinalities. *Entropy,* **13**(12), pp. 2013-2023.

NEMENMAN, I., SHAFEE, F. and BIALEK, W., 2002Entropy and inference, revisited, *Advances in neural information processing systems* 2002, pp. 471-478.

NEWELL, B.R. and DUNN, J.C., 2008. Dimensions in data: Testing psychological models using state-trace analysis. *Trends in cognitive sciences,* **12**(8), pp. 285-290.

NIEUWENHUIS, S. and DE KLEIJN, R., 2011. Consciousness of targets during the attentional blink: a gradual or all-or-none dimension? *Attention, Perception, & Psychophysics,* **73**(2), pp. 364-373.

NIEUWENHUIS, S., GILZENRAT, M.S., HOLMES, B.D. and COHEN, J.D., 2005. The role of the locus coeruleus in mediating the attentional blink: a neurocomputational theory. *Journal of Experimental Psychology: General,* **134**(3), pp. 291.

NIEUWENHUIS, S., VAN NIEUWPOORT, I.C., VELTMAN, D.J. and DRENT, M.L., 2007. Effects of the noradrenergic agonist clonidine on temporal and spatial attention. *Psychopharmacology,* **193**(2), pp. 261-269.

NIEUWENSTEIN, M.R., POTTER, M.C. and THEEUWES, J., 2009. Unmasking the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance,* **35**(1), pp. 159.

NISBETT, R.E. and WILSON, T.D., 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review,* **84**(3), pp. 231.

OLIVERS, C.N. and MEETER, M., 2008. A boost and bounce theory of temporal attention. *Psychological review,* **115**(4), pp. 836.

OLIVERS, C.N., VAN DER STIGCHEL, S. and HULLEMAN, J., 2007. Spreading the sparing: Against a limited-capacity account of the attentional blink. *Psychological research,* **71**(2), pp. 126-139.

OVERGAARD, M., ROTE, J., MOURIDSEN, K. and RAMSY, T.Z., 2006. Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Consciousness and cognition,* **15**(4), pp. 700-708.

OVERGAARD, M. and SORENSEN, T.A., 2004. Introspection distinct from first-order experiences. *Journal of Consciousness studies,* **11**(7-8), pp. 77-95.

PANINSKI, L., 2003. Estimation of entropy and mutual information. *Neural computation,* **15**(6), pp. 1191-1253.

PERSAUD, N., MCLEOD, P. and COWEY, A., 2007. Post-decision wagering objectively measures awareness. *Nature neuroscience,* **10**(2), pp. 257-261.

PINCHAM, H.L., BOWMAN, H. and SZUCS, D., 2016. The experiential blink: Mapping the cost of working memory encoding onto conscious perception in the attentional blink. *Cortex,* **81**, pp. 35-49.

PITTS, M.A., PADWAL, J., FENNELLY, D., MARTÍNEZ, A. and HILLYARD, S.A., 2014. Gamma band activity and the P3 reflect post-perceptual processes, not visual awareness. *NeuroImage,* **101**, pp. 337-350.

PRATTE, M.S., ROUDER, J.N. and MOREY, R.D., 2010. Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **36**(1), pp. 224.

PRINCE, M., BROWN, S. and HEATHCOTE, A., 2012. The design and analysis of state-trace experiments. *Psychological methods,* **17**(1), pp. 78.

PRINZ, J., 2006. Is the mind really modular. *Contemporary debates in cognitive science, ed.RJ Stainton,* , pp. 22-36.

RAYMOND, J.E., SHAPIRO, K.L. and ARNELL, K.M., 1992. Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of experimental psychology: Human perception and performance,* **18**(3), pp. 849.

ROBERTS, S. and PASHLER, H., 2000. How persuasive is a good fit? A comment on theory testing. *Psychological review,* **107**(2), pp. 358.

ROUDER, J.N. and LU, J., 2005. An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic bulletin & review,* **12**(4), pp. 573-604.

SALTI, M., BAR-HAIM, Y. and LAMY, D., 2012. The P3 component of the ERP reflects conscious perception, not confidence. *Consciousness and cognition,* **21**(2), pp. 961-968.

SANDBERG, K., TIMMERMANS, B., OVERGAARD, M. and CLEEREMANS, A., 2010. Measuring consciousness: is one measure better than the other? *Consciousness and cognition,* **19**(4), pp. 1069-1078.

SENSE, F., MOREY, C.C., PRINCE, M., HEATHCOTE, A. and MOREY, R.D., 2017. Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior research methods,* **49**(3), pp. 853-862.

SENSE, F., MOREY, C.C., PRINCE, M., HEATHCOTE, A. and MOREY, R.D., 2016. Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior research methods,* , pp. 1-10.

SERGENT, C. and DEHAENE, S., 2004. Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological science,* **15**(11), pp. 720-728.

SHAFTO, J.P. and PITTS, M.A., 2015. Neural signatures of conscious face perception in an inattentional blindness paradigm. *Journal of Neuroscience,* **35**(31), pp. 10940-10948.

SHALLICE, T., 1988. *From neuropsychology to mental structure.* Cambridge University Press.

SHANAHAN, M., 2006. A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and cognition,* **15**(2), pp. 433-449.

SIMIONE, L., AKYREK, E.G., VASTOLA, V., RAFFONE, A. and BOWMAN, H., 2017. Illusions of integration are subjectively impenetrable: Phenomenological experience of Lag 1 percepts during dual-target RSVP. *Consciousness and cognition,* **51**, pp. 181-192.

SIMONS, D.J. and RENSINK, R.A., 2005. Change blindness: Past, present, and future. *Trends in cognitive sciences,* **9**(1), pp. 16-20.

SLIGTE, I.G., SCHOLTE, H.S. and LAMME, V.A., 2008. Are there multiple visual short-term memory stores? *PLOS one,* **3**(2), pp. e1699.

SOTO, D. and SILVANTO, J., 2014. Reappraising the relationship between working memory and conscious awareness. *Trends in cognitive sciences,* **18**(10), pp. 520-525.

SPERLING, G., 1960. The information available in brief visual presentations. *Psychological monographs: General and applied,* **74**(11), pp. 1.

SQUIRES, K.C., HILLYARD, S.A. and LINDSAY, P.H., 1973. Cortical potentials evoked by confirming and disconfirming feedback following an auditory discrimination. *Perception & psychophysics,* **13**(1), pp. 25-31.

STIGLER, S.M., 1986. Laplace's 1774 memoir on inverse probability. *Statistical Science,* , pp. 359-363.

TAATGEN, N.A., JUVINA, I., SCHIPPER, M., BORST, J.P. and MARTENS, S., 2009. Too much control can hurt: A threaded cognition model of the attentional blink. *Cognitive psychology,* **59**(1), pp. 1-29.

TEPLAN, M., 2002. Fundamentals of EEG measurement. *Measurement science review,* **2**(2), pp. 1-11.

TEUBER, H., 1955. Physiological psychology. *Annual Review of Psychology,* **6**(1), pp. 267-296.

THOMAS, R.D., 1999. Assessing sensitivity in a multidimensional space: Some problems and a definition of a generald'. *Psychonomic bulletin & review,* **6**(2), pp. 224-238.

TIMMERMANS, B. and CLEEREMANS, A., 2015. How can we measure awareness? An overview of current methods. *Behavioural methods in consciousness research,* , pp. 21-46.

TRÜBUTSCHEK, D., MARTI, S., OJEDA, A., KING, J., MI, Y., TSODYKS, M. and DEHAENE, S., 2017. A theory of working memory without consciousness or sustained activity. *Elife,* **6**, pp. e23871.

TULVING, E., 1981. Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior,* **20**(5), pp. 479-496.

VALIANT, G. and VALIANT, P., 2011Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs, *Proceedings of the forty-third annual ACM symposium on Theory of computing* 2011, ACM, pp. 685-694.

VAN DEN BUSSCHE, E., HUGHES, G., VAN HUMBEECK, N. and REYNVOET, B., 2010. The relation between consciousness and attention: An empirical study using the priming paradigm. *Consciousness and cognition,* **19**(1), pp. 86-97.

VANDENBROUCKE, A.R., FAHRENFORT, J.J., SLIGTE, I.G. and LAMME, V.A., 2014. Seeing without knowing: neural signatures of perceptual inference in the absence of report. *Journal of cognitive neuroscience,* **26**(5), pp. 955-969.

VELICHKOVSKY, B.B., 2017. Consciousness and working memory: Current trends and research perspectives. *Consciousness and cognition,* **55**, pp. 35-45.

VOGEL, E.K., LUCK, S.J. and SHAPIRO, K.L., 1998. Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance,* **24**(6), pp. 1656.

WARRINGTON, E.K., 2014. The double dissociation of short-and long-term memory. *Human Memory and Amnesia (PLE: Memory),* **4**, pp. 61.

WYBLE, B., BOWMAN, H. and NIEUWENSTEIN, M., 2009. The attentional blink provides episodic distinctiveness: sparing at a cost. *Journal of Experimental Psychology: Human Perception and Performance,* **35**(3), pp. 787.

WYBLE, B. and BOWMAN, H., 2005Computational and experimental evaluation of the attentional blink: Testing the simultaneous type serial token model, *CogSci* 2005, pp. 2371-2376.

ZALTA, E.N., NODELMAN, U. and ALLEN, C., 2005. Stanford encyclopedia of philosophy. *Palo Alto CA: Stanford University,* .

ZEMAN, A., 2005. What in the world is consciousness? *Progress in brain research,* **150**, pp. 1-10.