High-level cognitive and neural contributions

to conscious experience and metacognition in visual perception

Brian Maniscalco

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

High-level cognitive and neural contributions to conscious experience and metacognition in visual perception

Brian Maniscalco

Visual processing in humans has both objective and subjective aspects. Objective aspects of visual processing consist in an observer's ability to accurately discern objective properties of visual stimuli. Subjective aspects of visual processing consist in an observer's visual experience of the stimuli and the observer's metacognitive evaluation of the reliability of objective visual processing. What is the nature of the relationship between objective and subjective visual processing? A wide range of views exists in the literature today, but a broad distinction can be drawn between (1) views holding that objective and subjective visual processing are intimately interrelated, such that changes in subjective processing should be associated with changes in objective processing; and (2) views holding that subjective visual processing is a separate, higher-order process, such that it is possible to change subjective processing without changing objective processing. Here we perform a series of psychophysical experiments to arbitrate between these views. To make the data analysis more powerful, we created a novel extension of signal detection theory for analyzing the informational content of subjective ratings of perceptual clarity and confidence (Appendix A).

We constructed a wide array of signal detection theoretic models capturing different hypotheses on the relationship between objective and subjective visual processing and performed a formal model comparison analysis in order to discern which model structures best accounted for a data set in which objective stimulus discrimination performance was dissociated from subjective ratings of visual clarity (Chapter 1). Results from this analysis favor a higher-order view of subjective visual processing. If the higher-order view is correct, it should be possible to disrupt the informational content

carried by subjective ratings of perceptual clarity and decision confidence without affecting an observer's objective ability to visually discriminate stimuli. We found two lines of novel empirical evidence for such dissociations. We show that when subjects perform a working memory task in which the contents of working memory require extensive manipulation, ratings of confidence in a concurrent perceptual task carry less information about perceptual task performance, even taking the influence of task performance into account (Chapter 2). Similarly, we show that transcranial magnetic stimulation to dorsolateral prefrontal cortex selectively impairs the metacognitive sensitivity of visual clarity ratings without affecting perceptual task performance (Chapter 3). Finally, we show that perceptual and metacognitive performance can dissociate over time as an observer performs a continuous block of trials in a visual discrimination task, contrary to views holding that perceptual discrimination and metacognition are closely intertwined processes (Chapter 4). We show that this dissociation can be partly attributed to individual variability in gray matter volume of regions of anterior prefrontal cortex previously linked to visual metacognition. We interpret these results as suggesting that limited prefrontal resources can be dynamically allocated to support the performance of either perceptual or metacognitive processes.

Taken together, these results provide converging evidence supporting a higher-order view of subjective visual processing. Functionally, objective and subjective processing are organized hierarchically, such that downstream subjective processes reflect the properties of objective processing but can be independently manipulated. Anatomically, these high-level subjective processes are linked to regions of prefrontal cortex rather than posterior perceptual areas.

Table of Contents

Figures and Tables

Acknowledgements

I would like to thank Hakwan Lau for being a great advisor and friend over the years. I have been with Hakwan's lab at Columbia from the beginning and I am sad to have to go. Hakwan managed to strike the right balance between providing me with mentorship and support but also providing me with freedom and encouraging me to become an independent thinker. I have always admired his ability to be both extremely critical and extremely creative, and from what I can tell, both of these skills are necessary to be a great scientist. I hope to develop these skills in equal measure in myself.

I have been fortunate to work with many intelligent and kind graduate students and postdocs over the years, including Dobromir Rahnev, Matthew Davidson, Chrystele Ody, Guillermo Solovey, Ai Koizumi, Wei Guo, Jorge Morales, and Travis Riddle. I would like to give special thanks to Matthew for introducing me to meditation, which has significantly enriched my life, and to Ai for helping me out so much in the past year.

My graduate work is indebted to all of the assistance and hard work of our past and present lab managers and research assistants, including Li Yan McCurdy, Guy Graney, Yoshiaki Ko, Tashina Graves, Sara Steele, Cynthia Shi, Linda Bahdo, Audrey Rakovich Seville, Elliott Huang, David Zhou, Laila Iravani, Joo Won Bang, Franc Camps-Febrer, Liz LaMarca, Nishtha Lamba, Michael Watson, Dylan Kotliar, Richard Barzaga, Alan Wu, and Kevin Hui.

I must give special acknowledgement to Li Yan McCurdy for collecting the voxel-based morphometry data described in Chapter 4 and performing the initial analysis to extract the regions of interest, and to Elisabeth Rounis for collecting the transcranial magnetic stimulation data described in Chapter 3.

Finally, I would also like to thank Jennifer Mangels for giving me my first opportunity to do psychology research many years ago. Jennifer was my first mentor and without her influence I would probably not be where I am today. I must also thank Jonah Hassenfeld, who was also in that lab, for our

marathon philosophical discussions which have come to help shape my worldview. I will always remember the years I worked in Jennifer’s lab with special fondness.

Dedication

For Mom, Dad, Chris, Benedetta, and Viggy.

Preface

Most of the work presented in this dissertation has been taken from published papers and manuscripts in preparation:

- Chapter 1 and Appendix B are based on Maniscalco and Lau (in review)
- Chapter 2 is based on Maniscalco and Lau (in preparation)
- Chapter 3 is based on Rounis, Maniscalco, Rothwell, Passingham, and Lau (2010) Cognitive Neuroscience
- Chapter 4 is based on Maniscalco, McCurdy, and Lau (in review)
- Appendix A is based on Maniscalco and Lau (in press) in Fleming and Frith (Eds) The Cognitive Neuroscience of Metacognition

In order to acknowledge the contributions of my collaborators in all phases of research from experimental design and data collection through analysis, interpretation, and writing, I will refer to our collective efforts using the pronoun "we."

General Introduction

What does it mean for an observer to "see" something? In everyday usage, the concept of "seeing" seems simple and straightforward. To see a stimulus is simply to be visually aware of it. On reflection, however, we can discern conceptual distinctions between the component processes and properties of seeing. For instance, at the most basic level, seeing a stimulus involves using visual information to register the presence of the stimulus, to identify what kind of stimulus it is, and to make appropriate behavioral responses to it. Aspects of vision such as these involve the observer's ability to accurately assess and interact with the external world. We may refer to these as the *objective aspects of visual processing*, or *objective vision* in brief.

However, for humans, seeing is not limited to the process of interpreting and interacting with the external world. Seeing also involves the process of *knowing what one sees* and *knowing how well one is seeing*. For instance, as you read this sentence, you are not merely engaged in the process of identifying the letters, words, word meanings, and so on. Concurrently with such objective visual processing, you have an explicit *visual experience* of colors and forms in a visual field, and a *felt sense* of how the visual stimuli cohere into wholes that have meaning. In this way, you are not only objectively seeing words, but you are also *aware* of yourself *as seeing* these words; that is, in addition to the bare process of objective seeing, there is another aspect of visual processing that represents the fact that an act of seeing is occurring. In addition to merely *knowing* that you are seeing, you may also have a sense of *how well* you are able to see. You probably experience the words in this sentence as being clearly legible, but in very poor viewing conditions you might have a sense that the words are difficult to see, or might feel uncertain that you can accurately identify a certain letter. Aspects of vision such as these involve an observer's ability to register and evaluate the nature and quality of their own objective visual processing. We may refer to processes such as these as the *subjective aspects of visual processing*, or *subjective vision* in brief.

A simple thought experiment helps to draw out the contrast between objective and subjective vision. We might imagine that a simple robot equipped with a video camera and an artificial limb could make basic visual discriminations and generate appropriate goal-directed behavior on the basis of this visual processing. Such a robot could be said to have a basic kind of objective ability to see, insofar as it is able to use visual information to accurately characterize the state of the world and interact appropriately with the world on that basis. However, our intuitions would suggest that the robot's objective ability to see is not necessarily accompanied by a first-person phenomenological experience of subjective visual qualities such as form and color; at any rate, there seems to be no logical contradiction in assuming that the robot would lack such visual experience. From a computational perspective, unless we were to endow this robot with additional capacities, the robot would not have a representation of itself *as* a system that is seeing something, nor any way to evaluate *how dependable* its visual discriminations are. In the phrasing of Karmiloff-Smith (1992), although there may be visual knowledge *in* the robot, this is not explicitly represented as visual knowledge *for* the robot. This example serves to illustrate that subjective vision is distinct from, and not necessarily entailed by, objective vision.

Given that objective and subjective vision are distinct processes, what is the relationship between them as implemented in the human observer? To what extent are they interlocked or dissociable? What computational architecture best describes the stream of processing in the human mind that allows us to objectively and subjectively see? The current work will address questions such as these. Broadly speaking, our main concern will be to determine whether subjective vision is best characterized as a low-level sensory process, or as a higher-level process somewhat removed from the lower levels of basic perceptual processing. We address this question by collecting data from human observers in visual psychophysics experiments and using formal modeling techniques to determine which view is best supported by the data. To anticipate, we find converging lines of evidence that the subjective aspects of visual processing are best characterized by higher-level, rather than low-level, perceptual processes.

Aspects and measures of subjective vision

We have already distinguished between two aspects of subjective vision: (1) a representation or experience of a stimulus *as being seen*, and (2) a representation or experience of a stimulus as being seen *well* or *poorly*. The first aspect we might call *visual awareness*, as it concerns whether a visual stimulus is explicitly registered in consciousness or whether it goes unnoticed and unexperienced, receiving only unconscious processing. The second aspect we might call *visual metacognition*, as it concerns to what extent objective visual processing is experienced or judged to be clear, accurate, reliable, etc. as opposed to poor, uncertain, degraded, etc.

Visual awareness

The study of perceptual awareness (and its complement, unconscious perceptual processing) has historically been beset by controversies regarding how to measure whether a stimulus is perceived consciously or not. Early approaches relied on taking subjective reports at face value—if an observer reports not being aware of a stimulus, then he is not aware of it (e.g. Peirce & Jastrow, 1884; Sidis, 1898; Stroh et al, 1908; Adams, 1957). However, theoretical advances in psychophysical research came to view perceptual reports as being a flexible, cognitive *decision process* that is sensitive to factors including, but not limited to, the nature of perceptual processing (e.g. Tanner & Swets, 1954; Green & Swets, 1966). Essentially, subjective reports might be contaminated by various response biases. For instance, consistent with reports that observers tend to be underconfident in sensory discrimination tasks (Björkman, Juslin, & Winman, 1993), observers might be reluctant to characterize extant but faint, ambiguous, or degraded experiences as instances of conscious perception. Additionally, the observer's criteria for mapping perceptual experience onto behavioral reports might shift as a function of the experimental context (Tanner & Swets, 1954).

In response to such methodological concerns, Eriksen (1960) proposed that awareness be measured by the observer's objective ability to discriminate the stimulus. Under this proposal, awareness can only be said to be absent if the observer's objective performance in identifying the stimulus is at chance levels. However, this method does not seem to respect the already mentioned conceptual distinctions between objective and subjective processing, and *a priori* rules out the possibility of unconscious perceptual processing, a phenomenon that has subsequently received strong empirical support from case studies on blindsight patients (Weiskrantz, 1986). Subsequent approaches to measuring perceptual awareness have proposed to corroborate subjective reports with more objective considerations, such as by demonstrating qualitative differences between the processing of sub- and supra-threshold stimuli (Cheesman & Merikle, 1986; Debner & Jacoby, 1994), or by measuring the extent to which subjective reports of confidence or post-decision wagering predict task performance (Dienes, Kwan, & Goode, 1995; Kunimoto, Miller, & Pashler, 2001; Persaud, McLeod, & Cowey, 2007).

However, for such proposals the conceptual question of what exactly we are measuring still looms large. For instance, if an observer consistently reports "clear" and "very clear" visual experiences of a stimulus, we would have strong reason to believe he experiences *something*, even if his distinction between "clear" and "very clear" experience does not carry predictive value regarding task performance (Dienes, 2004; Maniscalco & Lau, 2012). Considerations like these are suggestive that we should always take the direct contents of an observer's subjective reports into consideration in some form or another when assessing perceptual awareness (Maniscalco & Lau, 2012). For all their potential methodological flaws, subjective reports nonetheless seem to provide a crucially important window into the nature of an observer's subjective experience, since experience as such is a private, first-person phenomenon that is inaccessible to direct objective measurement—alas, there is no such thing as a "consciousness meter" (Chalmers, 1996).

Visual metacognition

There are some senses in which the measurement of visual metacognition is not as methodologically problematic as the measurement of visual awareness. Whereas assessing visual awareness involves making an *absolute* distinction between the presence or absence of a private visual experience, visual metacognition can be cast as making *relative* distinctions between visual processing that is experienced with more or less phenomenological clarity, or engenders more or less confidence. Thus, for instance, even if we cannot be sure precisely what an observer might mean to communicate by reporting that a visual experience is clear or cause for high confidence, it is less problematic to interpret *differences* in reports of clarity or confidence across trials or experimental conditions, particularly for within-subject comparisons.

Additionally, the construct of visual metacognition avails itself to a more straightforward methodological treatment than does visual awareness. Because metacognitive reports can be seen as evaluations of the efficacy of objective perceptual processing, and because the efficacy of objective perceptual processing can be directly measured, it is relatively straightforward to operationalize metacognition as the degree to which metacognitive reports predict objective performance (see Appendix A for our signal detection theory approach to doing so). The analogous procedure is not available in the case of measuring visual awareness, since we have no direct access to the subjective experiences that subjective reports of awareness bear upon.

Thus far we have equivocated between reports on the *phenomenological clarity* of a visual experience, and the degree of *confidence in the efficacy* of visual processing, treating both of these as forms of visual metacognition. Conceptually, the degree of clarity with which a visual content is experienced could be seen as a form of metacognitive appraisal intrinsically woven into the phenomenological character of the experience; clearer, more vivid, less ambiguous experiences are more likely to be associated with effective objective performance (and higher reports of confidence). Indeed, according to some proposals, sensory representations only enter awareness if they are assessed to be sufficiently statistically reliable, implying a close connection between perceptual awareness, perceptual clarity, and perceptual confidence (Lau, 2008a). A similar construct to perceptual clarity is the construct of *perceptual fluency*, the sense of ease with which a stimulus is perceived; perceptual fluency has similarly been taken to serve a kind of experientially grounded metacognitive function (Oppenheimer, 2008).

Nonetheless, confidence is a more general concept, in that it characterizes the efficacy of perceptual performance without explicitly specifying the qualitative clarity of perceptual experience as the source of such characterizations. Thus, in principle, we might expect that judgments of confidence admit of more sources of influence than judgments of visual clarity. For instance, an observer's introspective rating of confidence in his perceptual decision may be influenced not only by how clearly he perceived the stimulus phenomenologically, but also by more abstract, non-sensory 'fringe' experiences such as a 'gut feeling' of rightness (James, 1890; Mangan, 2001; cf type 2 blindsight in Weiskrantz, 1997), as well as other non-perceptual considerations such as the observer's previous experience with similar stimuli, the observer's access to performance feedback, etc.

This conceptual observation that judgments of confidence may draw upon more sources of information than do judgments of perceptual clarity is supported by empirical observations. Patients with a neurological condition termed 'blindsight' have damage to areas of the primary visual cortex, V1, which entails the loss of all visual experience in the corresponding part of the patient's visual field. Yet, such patients can make objective forced-choice discriminations about stimuli presented in the subjectively 'blind' portions of the visual field at above-chance levels (Weiskrantz, 1986). Interestingly, in experiments where blindsight patient GY has been asked to place a wager on every trial regarding the accuracy of his perceptual decisions, his wagers can predict accuracy for stimuli presented to his blind field at above-chance levels of performance (Persaud et al., 2007; Persaud et al., 2011). Similarly, healthy observers are able to rate confidence in such a way as to predict objective performance in visual tasks at above-chance levels for trials on which they deny having detected the visual target (Kanai, Walsh, & Tseng, 2010).Thus, to the extent that judgments of confidence can carry effective information about perceptual processing even in the absence of reports of visual awareness, such judgments must have access to information about perceptual processing beyond what is captured by experiences of visual clarity.

However, although not identical in their content, ratings of confidence and visual clarity are conceptually and empirically alike. Conceptually, they both carry metacognitive appraisals about the efficacy of objective perceptual processing. In an open-ended phenomenological study, Ramsøy and Overgaard (2004) had subjects perform a visual discrimination task and asked the subjects to report the "degree of clearness of experience" on each trial. Subjects were not presented with an *a priori* reporting scheme, but rather were invited to construct their own rating scales, assigning meaning to each category of the scale in such a way as to capture the range of clarity in visual experience elicited by the experimental stimuli. All five subjects in the experiment wound up converging on essentially the same 4-category classification scheme, with the semantic content of the categories summarized by the authors as "no experience whatsoever," "brief glimpse," "almost clear experience," and "clear experience." Thus, subjects spontaneously reported the phenomenology of visual clarity on a graded scale resembling graded ratings of confidence in their ability to discriminate the target. Indeed, the subjects indicated in post-experiment interviews that for stimuli that elicited "no experience," the forced choice visual discrimination response was experienced as a pure guess, whereas "almost clear" and "clear" stimuli were associated with feelings of being almost certain and certain, respectively.

Empirically, although dissociable, ratings of clarity and confidence tend to be well correlated. An indirect empirical suggestion that such correlation might exist comes from studies showing that fluency in memory retrieval tasks is positively related with ratings of confidence (Kelley & Lindsay, 1993; Koriat, 1993). More direct evidence on the similarity between ratings of clarity and confidence comes from Sandberg, Timmermans, Overgaard, & Cleeremans (2010). Subjects performed a visual discrimination task and provided a metacognitive report on every trial after the forced choice visual discrimination. In one condition, subjects used the 4-point “perceptual awareness scale” of visual clarity previously described by Ramsøy and Overgaard (2004). In another condition, subjects used a 4-point confidence rating scale whose categories were “not confident at all,” “slightly confident,” “quite confident,” and “very confident.” In a third condition, subjects placed one of 4 possible wagers on their task performance with imaginary monetary bets. Distributions of rating scale responses as a function of stimulus strength, and the relationship of task performance to stimulus strength as a function of rating scale response, were qualitatively similar for ratings of clarity and confidence, whereas both scales were relatively less similar to the wagering scale.

Unpublished data from our lab further corroborates the relationship between ratings of visual clarity and confidence. Three subjects performed a metacontrast masking task similar to that described in Chapter 1. On each trial, a square or diamond was presented, followed by a metacontrast mask. The stimulus onset asynchrony (SOA) between the target and mask could take on one of eight possible values. In one experimental condition, after providing a forced choice discrimination regarded the identity of the visual target, subjects rated confidence in the discrimination on a scale of 1 to 4. In another condition, subjects reported the clarity with which the target was perceived on a scale of 1 to 4. All three subjects exhibited strong across-SOA correlations for average ratings of clarity and confidence (*r*s > .7, *p*s < .05).

For the purposes of the present work, then, we treat perceptual clarity and perceptual confidence as similar constructs. The distinction between these processes is not important for the broader questions we address in this manuscript, but nor does our methodological approach and interpretation of the data crucially rely upon not making a sharp distinction between them. We will sometimes use the phrase "perceptual metacognition" to refer to both types of judgments interchangeably.

Candidate computational architectures relating objective and subjective vision

Given that objective and subjective vision are distinct processes, how are they implemented in the physical structure and functional architecture of the human brain? There has been an explosion of interest in this question in the last two decades, and a wide range of views have been put forth in the literature (Tong, 2003; Tononi & Koch, 2008; Dehaene & Changeux, 2011; Lau & Rosenthal, 2011). Consideration of the wide range of views intended to account for perceptual awareness can be helpfully simplified and organized by adopting the taxonomy described in Lau and Rosenthal (2011). These authors distinguished between first-order theories, information integration theory, neuronal global workspace theory, and higher-order theories.

First-order theories (Pins & Ffytche, 2003; Tong, 2003; Zeki, 2003; Tse, Martinez-Conde, Schlegel, & Macknik, 2005; Lamme, 2006) hold that perceptual awareness arises as a function of early sensory processing regions in the brain. For instance, in the case of vision, visual awareness should be associated with processing in primary visual cortex (Tong, 2003) and/or extrastriate visual cortex (Zeki, 2003; Tse et al., 2005; Lamme, 2006).

Information integration theory (Tononi, 2008) holds that awareness arises as a function of the computational complexity of a physical system. Specifically, systems that feature higher degrees of computational integration have higher degrees of awareness, where "integrated information captures the information generated by causal interactions in the whole, over and above the information generated by the parts" (Tononi, 2008, p. 221).

Neuronal global workspace theory (Dehaene, Sergent, & Changeux, 2003; Dehaene & Changeux, 2011) builds on the cognitive global workspace theory first put forth by Baars (1989). According to this view, the contents of consciousness correspond to activations in a 'neuronal global workspace' consisting of a dynamic, coherent network of interacting neurons based crucially upon long range cortico-cortical connections in prefrontal and parietal cortex. If processing in early sensory areas is sufficiently strong and receives top-down amplification via attentional selection, then this low-level sensory activation may cross a threshold and gain access to the global neuronal workspace, and thus enter perceptual awareness (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006). Access to the global workspace entails extended processing that makes the first-order sensory processing more robust and accessible to a diverse range of higher cognitive functions such as introspective report and cognitive control (Dehaene & Changeux, 2011).

Higher-order theories (Lau, 2008a; Lau, 2011; Dienes, 2008; Pasquali, Timmermans, & Cleeremans, 2010) hold that perceptual awareness occurs as the result of higher-order cognitive and neural processing that is 'about' first-order sensory processing. For instance, a sensory representation may enter perceptual awareness if judged by higher-order mechanisms to have a sufficiently high degree of statistical reliability (Lau, 2008a; Cleeremans, 2008). Thus, in a sense, there is a relatively clean division of labor whereby lower-order processes are responsible for evaluating the state of the world (objective perceptual processing) and higher-order processes are responsible for evaluating the state of lower-order processes (subjective perceptual processing).

Thus, not only are there many views on the manner in which perceptual awareness is implemented anatomically and functionally, but these views span a wide gamut, ranging from views that locate perceptual awareness in the very earliest stages of cortical perceptual processing up through views that locate perceptual awareness in the very latest stages. This situation is a reflection of the empirical and methodological difficulties and ambiguities in studying perceptual awareness.

For the purposes of the current work, we will be particularly concerned to assess the *relationship* between objective and subjective perceptual processing. Because first-order theories locate perceptual awareness in sensory cortex, they posit a tight relationship between objective and subjective perception; in particular, changes in an observer's subjective reports about perceptual awareness and clarity should tend to be associated with changes in objective perceptual performance (Lau & Rosenthal, 2011). Although neuronal global workspace theory posits that perceptual awareness is associated with higher-level activations in prefrontal and parietal cortex, rather than being restricted to earlier sensory regions, it nonetheless is similar to first-order theories to the extent that it posits a direct relationship between objective and subjective perception (Lau & Rosenthal, 2011). Changes in subjective reports reflect changes in a sensory representation's access to the global workspace, which according to global workspace theory should entail changes in objective processing, e.g. changes in the robustness of the sensory representation and its degree of access to higher cognitive evaluation. By contrast, higher-order theories are unique in positing that changes in subjective perceptual processing can occur in the absence of a concurrent change in objective processing. (At its current stage of development, it is not yet entirely clear how the formalisms of information integration theory might map onto specific predictions about the relationship between objective and subjective perception.)

A similarly broad contrast between low-level and high-level views has occurred independently in the narrower literature focused specifically on perceptual confidence. Some labs contend that evaluations of confidence in perceptual processing bear simple and straightforward relationships to objective perceptual processing, such that confidence is essentially a direct measure of perceptual signal strength and the two therefore co-vary in reliable ways (Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009; Kepecs & Mainen, 2012). Other labs hold that confidence is constructed by higher-order evaluations of first-order perceptual processing, such that the two are partly dissociable (Fleming, Weil, Nagy, Dolan, & Rees, 2010; Pleskac & Busemeyer, 2010; McCurdy et al., 2013).

Thus, our fundamental concern in this work will be to help arbitrate between views positing that perceptual clarity and confidence are intimately related to objective perceptual processing and perhaps even based upon the same underlying information, and views positing that perceptual clarity and confidence are somewhat functionally removed from, and thus partially dissociable from, objective perceptual processing. In order to place our approach in the proper context, we will first discuss the conceptual and methodological importance of dissociating objective and subjective perception, prior empirical demonstrations of such dissociations, and the utility of analyzing objective and subjective perception in the computational framework of signal detection theory.

Methodological importance of dissociations

In order to isolate the neural and functional processes underpinning subjective perception, it is crucial to rule out the potential influence of experimental confounds. In the context of neuroimaging experiments seeking to discover the neural correlates of perceptual awareness, great emphasis has been placed upon controlling for stimulus confounds—i.e.in making perceptual stimuli as similar as possible across experimental conditions, in order to ensure that comparisons of neural activity across conditions captures differences in perceptual awareness in and of itself, and not low-level differences in sensory processing due to differences in the stimuli (e.g. Dehaene et al., 2001; Blake & Logothetis, 2002).

However, a less appreciated principle is that it is also necessary to control for performance confounds (Lau, 2008b; Weiskrantz, Barbur, & Sahraie, 1995), since experimental manipulations that induce changes in reports of perceptual awareness also tend to induce changes in objective perceptual performance. Indeed, in psychophysical tasks, objective performance and subjective measures of awareness and confidence typically exhibit robust correlations (see e.g. a representative example of how objective performance and subjective ratings exhibit similar changes as a function of stimulus strength in Figure B-1, reprinted from Figure S3 of Del Cul, Dehaene, Reyes, Bravo, & Slachevsky, 2009). Thus, the study of perceptual awareness and metacognition must take great care to disentangle the interrelated but distinct strands of objective and subjective perceptual processing. The most straightforward and effective way to accomplish this is to discover and leverage empirical dissociations between objective perceptual performance and subjective reports.

Dissociating objective and subjective perception takes on an even more important role in our endeavor to compare different theories that make different predictions about how objective and subjective processing are related. In this context, such dissociations are not just serving the role of helping to rule out confounds in the interpretation of experimental outcomes, but rather are central to the project of arbitrating between the different theories.

Dissociations between objective and subjective perception

Probably the most ubiquitous kind of objective-subjective dissociation in psychological science is the demonstration that some degree of stimulus processing can occur even in the absence of awareness of the stimulus (e.g. Kouider & Dehaene, 2007; Bargh & Morsella, 2008), even in spite of the previously discussed methodological difficulties involved in conclusively demonstrating the complete absence of stimulus awareness (Eriksen, 1960; Holender, 1986; Holender & Duscherer, 2004). However, many such studies involve demonstrating *indirect* effects of unconscious stimuli on behavior, such as effects of an unconscious prime on reaction times for a subsequently presented suprathreshold stimulus. A stronger and more informative kind of dissociation for the present purposes would involve differences in objective and subjective processing related to the very same stimulus. The aforementioned neuropsychological deficit of blindsight (Weiskrantz, 1986) qualifies as such a direct dissociation, insofar as it demonstrates that objective performance in forced-choice visual discriminations of a stimulus can occur at above-chance levels even though the blindsight patient profusely denies having any awareness of the stimuli.

However, even more germane to the current agenda than demonstrations of objective processing without awareness would be investigations on the extent to which the overall *relationship* between objective and subjective processing is malleable. This is because different theories of perceptual awareness and metacognition differ on how they characterize the relationship between objective and subjective processing, as already noted. Fortuitously, potential demonstrations of dissociation in the relationship between objective and subjective perception also allow us to side-step the methodologically thorny issues of demonstrating the absence of perceptual awareness. Such "objective-subjective relationship" dissociations are also conceptually stronger than "unconscious perception" dissociations, in the sense that weaker assumptions are needed for such dissociations to support inferences to underlying processes (Schmidt & Vorberg, 2006).

One such kind of dissociation consists in finding two experimental conditions that yield identical levels of objective perceptual performance and yet differ in average levels of reported perceptual clarity or confidence. For instance, Lau and Passingham (2006) had subjects perform a simple shape discrimination task and indicate whether the target was seen or unseen for stimuli that were backward-masked with a metacontrast mask. They systematically altered the stimulus onset asynchrony (SOA) between target and mask, ranging from -50 ms to 133 ms. It is well known that when percentage of correct responses is plotted as a function of SOA, a U-shaped function results (Breitmeyer, 1984), such that task performance is best for short and long SOAs and worse for intermediate SOAs, and this pattern was replicated by Lau and Passingham. When they plotted the percentage of "seen" responses as a function of SOA, they found that this curve was also U-shaped, but was asymmetric with respect to the percent correct curve, such that they could find pairs of SOAs exhibiting the same level of percent correct but differing levels of percent seen. They termed this finding a “relative blindsight” effect. Similarly, Rahnev et al. (2011) had subjects perform a tilt discrimination task for grating stimuli that could be either attended (cued) or unattended (uncued). They adjusted the contrast of the grating stimuli so that task performance was equivalent in the attended and unattended conditions, and found that subjects rated the average visibility of unattended stimuli to be higher than that of attended stimuli in spite of having identical levels of task performance. Notably, such dissociations significantly facilitate the study of subjective perception by providing a means of circumventing the problem of performance confounds discussed previously (Lau, 2008b).

A related kind of dissociation focuses on the relationship between objective perceptual performance and perceptual metacognitive sensitivity, where the latter term refers to the efficacy with which ratings of clarity or confidence discriminate between an observer’s own correct and incorrect perceptual decisions. Modest demonstrations that perceptual and metacognitive sensitivity can dissociate have come from analyses of individual differences. Maniscalco and Lau (2012) found that there is across-subject variability in the relationship between perceptual and metacognitive sensitivity, and Fleming et al. (2010) found that there is across-subject variability in metacognitive sensitivity even when stimuli are adjusted so as to yield the same level of objective task performance for all subjects. In the current work we will demonstrate several stronger, within-subject dissociations between perceptual and metacognitive sensitivity, and such dissociations will play a key role in abdicating between lower- and higher-order views of the subjective aspects of visual processing.

Dissociations between perceptual and metacognitive sensitivity have conceptual advantages over dissociations between perceptual sensitivity and average levels of reports of awareness, clarity, or confidence. Metacognitive sensitivity measures the informational content of subjective ratings, rather than just the mean reported level of such ratings. As a consequence, (1) the objective measure (perceptual sensitivity) and the subjective measure (metacognitive sensitivity) are more directly comparable, both being measures of sensitivity, and (2) the subjective measure is less prone to the potential influence of response biases. As we will see in the next section, comparisons of perceptual and metacognitive sensitivity provides a third advantage when analyzed in the context of signal detection theory (SDT), since this theoretical framework provides theoretical, quantitative predictions on how these measures *should* be related according to the assumptions of SDT. The comparison of such predictions to observed outcomes can provide useful context and insight in the interpretation of the data.

Signal detection theory analysis of objective and subjective perception

Signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005) provides a simple yet powerful computational framework for characterizing performance in perceptual tasks. An observer's response bias in a perceptual task is fluid and can change as a function of the experimental context (Tanner & Swets, 1954). However, even as an observer's criteria for how to translate perceptual processing into behavioral responses change, presumably the observer's underlying ability to objectively process the stimuli—i.e. the observer's perceptual sensitivity—remains unchanged. SDT's primary empirical virtue is that its measure of perceptual sensitivity, *d'*, remains constant even as an observer's response bias changes (Swets, 1986b). Thus, SDT provides a model that can distinguish and independently characterize an observer's perceptual sensitivity and response bias. This aspect of SDT makes it an ideal framework for the current purposes, since we are concerned with characterizing the relationship between objective perceptual processing—perceptual sensitivity—and subjective reports.

Recent theoretical developments extending SDT to the domain of metacognitive performance (Clarke, Birdsall, & Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003; Maniscalco & Lau, 2012; Appendix A) have made SDT even more useful for the current purposes. These developments show that SDT makes strong theoretical predictions regarding the relationship between perceptual and metacognitive performance. Given that an observer exhibits a given level of objective perceptual processing—i.e. a certain value of *d′* and response bias *c*—SDT makes a quantitative prediction regarding how informative the observer's confidence ratings should be about his perceptual performance. This prediction can be quantified with the measure meta-*d′*, which is defined such that an SDT-ideal observer has *d′* = meta-*d′* (Maniscalco & Lau, 2012; Appendix A). This prediction provides a theoretical reference point against which we can evaluate an observer's actual metacognitive performance. For instance, if the observer has *d′* = meta-*d′*, then the observer exhibits metacognitive performance consistent with the SDT-ideal observer, whereas if *d′* > meta-*d′*, then the observer is metacognitively suboptimal.

In addition to providing a benchmark for interpreting an observer's metacognitive performance in a given experimental condition, meta-*d′* also facilitates the comparison of objective and subjective perceptual processing across conditions. For instance, we have previously argued that analyses of subjective perception should be careful to control for performance confounds, and that one means of doing so is to compare experimental conditions where subjective perception differs and yet objective performance is the same. This empirical method of controlling for performance confounds is effective, but is limited by the (to date) relatively small number of experimental procedures that can produce the performance-matching dissociation. However, the framework of SDT provides a *theoretical*, rather than empirical, way of assessing subjective perception while controlling for performance confounds. Numerical comparisons between *d′* and meta-*d′* can quantify metacognitive sensitivity in such a way as to take into account the influence of the objective *d′* measure on the subjective meta-*d′* measure. For instance, if in condition A an observer has *d′* = 1 and meta-*d′* = 1, whereas in condition B he has *d′* = 2 and meta-*d′* = 1.5, then we can infer that, after taking into account his level of task performance, the observer's metacognition is worse in condition B, since e.g. meta-*d′* only achieves 75% of the value of *d′* as compared to the 100% value achieved in condition A. In this way, the SDT framework lends considerable flexibility and power in the study of the relationship between objective, perceptual sensitivity and subjective, metacognitive sensitivity.

For a fuller treatment on SDT approaches to measuring metacognition, see Appendix A and Maniscalco & Lau (2012).

Summary of theoretical questions and methodological approach

A brief summary of the above could be stated as follows. Although there are many theories of perceptual awareness and metacognition, a common denominator is that some predict that objective and subjective perceptual processing are intimately related, whereas others predict that the two are more loosely related and can dissociate in interesting ways. In this work, we will make use of experimental paradigms that yield various kinds of dissociations between objective perceptual performance and subjective ratings of perceptual clarity and confidence. We will make extensive use of the computational framework of SDT in order to demonstrate the existence of these dissociations and model the possible underlying mechanisms. We will use the SDT analyses of these empirical data in order to argue that higher-order theories of perceptual awareness and metacognition are best suited to account for the full range of the data.

Outline of the present manuscript

The current manuscript is composed of four chapters and two appendices.

In Chapter 1, we replicate the relative blindsight finding from Lau and Passingham (2006). We create multiple SDT models intended to capture the core computational principles of three classes of views on perceptual awareness—single channel models (akin to first-order theories), dual channel models (single channel models augmented with an additional, “unconscious” processing stream), and hierarchical models (akin to higher-order theories). Formal model comparison analysis suggests that the hierarchical model structure is best able to account for the relative blindsight dissociation.

In Chapter 2, we probe the effects of working memory demands on objective and subjective vision in two experiments. Concurrent with performing a basic perceptual task, subjects perform a demanding working memory task. The results suggest that when subjects are required to manipulate the contents of working memory under conditions of high load, metacognitive sensitivity is selectively reduced. The nature of the observed dissociation is consistent with higher-order, but not first-order, views. In conjunction with prior empirical findings, the results suggest a role of dorsolateral prefrontal cortex (DLPFC) in supporting perceptual metacognition.

In Chapter 3, we probe the effects of transcranial magnetic stimulation (TMS) to DLPFC on objective and subjective vision. We find that TMS selectively inhibits metacognitive sensitivity but not perceptual sensitivity. The nature of the observed dissociation is consistent with higher-order, but not first-order, views.The results provide further evidence for the role of DLPFC in supporting perceptual metacognition.

In Chapter 4, we perform four experiments to probe the dynamics of perceptual and metacognitive sensitivity over time in blocks of experimental trials. We find that changes in the two are weakly or negatively correlated, rather than the strong positive correlation that would be predicted by first-order views. We find that between-subject variability in this pattern can be explained by between-subject variability in gray matter volume of anterior prefrontal cortex. We construct a cognitive resource account of the neural findings and corroborate a prediction of this account in two further behavioral experiments.

In Appendix A, we provide an extensive formal treatment of the SDT model of objective perceptual performance and our application of the SDT model to the measurement of metacognitive sensitivity.

In Appendix B, we provide a supplement to the model analysis of Chapter 1, demonstrating the similarity of our SDT models to alternative models described elsewhere in the literature.

Chapter 1

The signal processing architecture underlying subjective reports of perceptual clarity

Introduction

What are the mechanisms that drive subjective and objective visual judgments in humans, and how are they related? As discussed in the General Introduction, several prominent classes of theories characterizing the relationship between objective and subjective vision are currently in circulation (Tong, 2003; Tononi & Koch, 2008; Dehaene & Changeux, 2011; Lau & Rosenthal, 2011). Here we consider the general forms of different signal processing architectures that map onto the various theories.

The most parsimonious kind of account holds that subjective and objective judgments, though distinct, are generated from the same underlying process (Single Channel models, Figure 1-1 left panel). For instance, on a common signal detection theory (SDT) account, perceptual decisions result from a binary comparison between an internal signal and a criterion (Green & Swets, 1966; Macmillan & Creelman, 2005), whereas subjective judgments of the quality of evidence are made by evaluating some transformation of the signal, such as its distance from the criterion (Clarke et al., 1959; Galvin et al., 2003). According to this kind of model, subjective and objective judgments are just different ways of evaluating the same underlying evidence (Figure 1-3; Appendix A).

Alternatively, even if subjective and objective judgments are based on the same evidence, the *quality* of evidence available for each kind of judgment might differ. For instance, a Hierarchical model (Figure 1-1 right panel) might suppose that evidence is first used to generate objective perceptual decisions, and subsequently undergoes further processing in order to make subjective judgments (Cleermans et al., 2007; Lau, 2008a; Fleming et al., 2010). On such an account, the evidence might

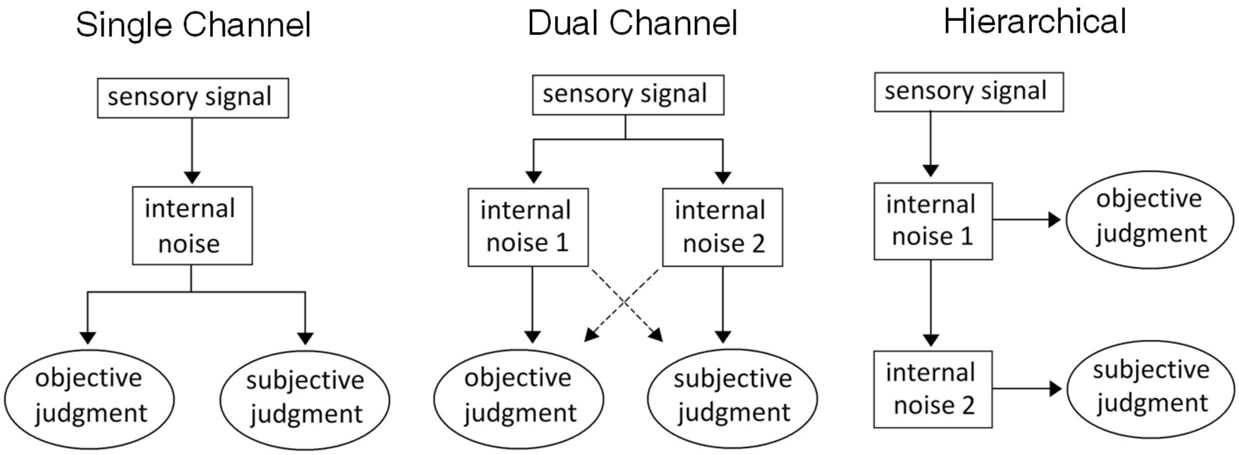

Figure 1-1. Schematic diagram for the three categories of models. (Left) According to a Single Channel model, the same process gives rise to both objective judgments (e.g. perceptual decisions about the identity of a stimulus) and subjective judgments (e.g. confidence ratings or visibility ratings). The model can still support some independence between task performance and subjective reports by supposing that the sensory evidence is a continuous variable that can be evaluated by setting various decision criteria (Figure 1-3; Appendix A; Green & Swets, 1966; Macmillan and Creelman, 2005). (Middle) An alternative model is that objective and subjective judgments are driven by two parallel processes, each influenced by independent sources of noise. Differential contribution of the two channels to objective and subjective judgments can lead to dissociations between the two kinds of responses. Note that the model can allow that each channel can contribute both kinds of judgment to some extent. In particular, one would expect that the channel which primarily influences one’s subjective ratings would also heavily influence one’s objective task response. For instance, when an observer subjectively reports clearly and vividly seeing squares, this should strongly correlate with objective judgments that the stimuli on the current trial are squares. (Right) Another alternative is that objective and subjective judgments are driven by different processes that are organized in a serial hierarchy, such that an early stage of processing generates the objective judgment and a later stage of processing generates the subjective judgment, as if the latter evaluates the quality of the former. Note that on this model, the second stage inherits the noise of the first stage, and thus the two are not entirely independent. However, the influence is one sided; the “subjective” stage does not influence the “objective” stage of processing.

become degraded by the time it is processed by subjective judgment mechanisms, due to a decaying signal and/or the accrual of noise (Pleskac & Busemeyer, 2010).

A third possibility is a Dual Channel model (Figure 1-1 middle panel) in which subjective and objective judgments are based on separate cognitive or neurophysiological processes (Jacoby, 1991; Jolij & Lamme, 2005; Del Cul et al., 2009; Morewedge & Kahneman, 2010). For instance, perhaps there are two independent visual processing routes, one of which supports conscious vision and another whose visual processing is entirely unconscious. On such an account, subjective and objective judgments access different sources of information (and noise).

In this chapter, we capitalize on a psychophysical paradigm that dissociates changes in objective perceptual decision performance from changes in subjective visibility ratings (Lau & Passingham, 2006) in order to evaluate SDT implementations of the model categories described above.

Methods

In the metacontrast masking procedure, stimulus identification performance varies with stimulus-mask onset asynchrony (SOA) in a U-shaped fashion (Figure 1-2). Visibility judgments follow a similar U-shape that is asymmetrical with respect to the objective performance curve, thus yielding similar levels of performance associated with different levels of subjective stimulus visibility. We compared the ability of various implementations of the Single Channel, Dual Channel, and Hierarchical models to capture the relative dissociation between subjective and objective judgments found in this data set.

Participants

59 students from the Columbia University undergraduate population participated in the experiment and were paid $10 for approximately one hour of participation. All participants were naïve

regarding the purpose of the experiment, had normal or corrected-to-normal vision, and signed an informed-consent statement. The research was approved by the Columbia University's Committee for the Protection of Human Subjects.

Experimental procedure

Subjects were seated in a dim room, 60 cm away from the computer monitor. Stimuli were generated using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) in MATLAB® (MathWorks, Natick, MA) and were shown on an iMac monitor (19 inch monitor size, 1680 x 1050 pixel resolution, 60 Hz refresh rate).

On each trial, a ring of eight shapes with a 4° radius was presented around a central fixation point (Figure 1-2). (A ring of stimuli was used with potential extension to fMRI in mind; to facilitate efficient retinotopic delineation of visual areas it is useful to present stimuli outside of the fovea. However, behavioral results similar to those reported here were also found with foveal presentation of single stimuli in Lau & Passingham, 2006.) Within each trial, each of the eight shapes was identical. The shapes could be either squares or diamonds with sides measuring 1.5° of visual angle. The shapes were presented for 33 ms on a gray background. Shapes were darker than the background, with the precise darkness determined separately for each subject by a thresholding procedure. A set of metacontrast masks designed to trace the outline of the square and diamond stimuli without physically overlapping with them (line width .025°) was subsequently displayed for 50 ms. Stimulus onset asynchrony (SOA) between stimulus and mask was determined randomly on each trial and counterbalanced among 8 possible durations, ranging from 0 ms to 116.7 ms in increments of 16.7 ms.

Following each stimulus presentation, subjects provided two responses. First, they made a forced choice objective judgment about the shapes of the stimuli (squares or diamonds). Next, they rated how subjectively visible the shape of the stimulus appeared using a 4 point scale. Specifically,

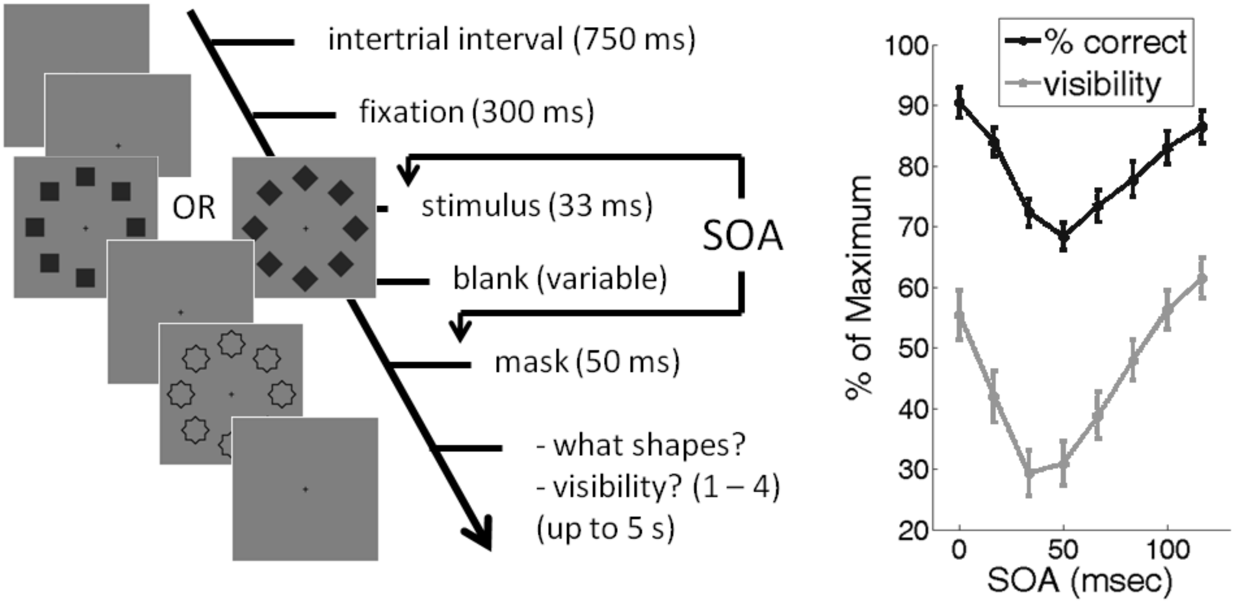

Figure 1-2. Experimental design and basic behavioral results. (Left) We used a paradigm based on metacontrast masking, similar to the one used in a previous study (Lau and Passingham, 2006). In every trial the subject was presented with a set of squares or diamonds (i.e. tilted squares). After a varying stimulus onset asynchrony (SOA; the temporal gap between the two sets of stimuli), a mask was presented. The mask did not overlap spatially with the targets, but nevertheless impaired their visibility. In each trial subjects first decided whether the targets were squares or diamonds, and then gave subjective visibility ratings (4 levels) to indicate how clearly they saw the identity of the targets. (Right) Replicating previous findings (Lau and Passingham, 2006), this masking procedure gives rise to a U-shaped masking function when stimulus identification performance is plotted against SOA. The average level of subjective visibility ratings across SOAs, however, did not take the same shape, and reflected a bias towards giving lower ratings at lower SOAs. Shown here were data from a selected group subjects (n=20) who particularly demonstrated this pattern of dissociation in a relatively pronounced fashion.

subjects were asked to rate how clearly they had perceived the stimuli. Subjects were encouraged to use the entire rating scale while still accurately characterizing what they had visually experienced. Stimulus presentation for the next trial commenced 1050 ms after subjects entered the visibility rating. However, if subjects failed to enter both the stimulus identity judgment and the visibility rating within 5 seconds of stimulus offset, the current trial was aborted and the next trial commenced automatically.

After receiving task instructions, subjects completed two blocks of 28 practice trials. Following practice, subjects completed a block of 120 trials in order to determine the Weber contrast of the stimuli at which threshold performance across all SOAs could be obtained. Because performance in this task is close to maximal with an SOA of 0 ms (Lau & Passingham, 2006), all trials in the thresholding procedure had the minimum stimulus-mask SOA of 0 ms. We reasoned that if near maximal performance at 0 ms could be controlled to be at threshold levels, performance at other SOA values would also be near threshold. Stimuli were initially set to a Weber contrast of -.15 and were subsequently adjusted online using a QUEST procedure (Watson & Pelli, 1983). Three separate QUEST tracks were recorded (40 trials each). Each QUEST track provided an independent estimate of the stimulus contrast needed to produce threshold performance (84% correct) at the minimum SOA. Trials for each track were interleaved randomly. Among the 3 resulting QUEST estimates, the median stimulus contrast was selected as the contrast to be used throughout the remainder of the experiment.

In the main experimental block, subjects completed 800 trials (100 trials for each of the 8 SOAs). SOAs were distributed across trials randomly. Every 100 trials, subjects received a self-terminated break lasting up to 60 seconds.

Subject selection

In order to maximize the suitability of the data for model fitting, we omitted from analysis all subjects who performed below chance levels at any of the SOAs (n=16), any who performed perfectly at any of the SOAs (n=3), and any whose mean visibility rating was lower than 5% of the maximum possible value at any SOA (n=1). Most subjects were excluded due to having at least one SOA with below chance levels of performance, which is perhaps not surprising given that we performed the thresholding procedure on only the 0 ms SOA and subjects had many chances at each of the other SOAs to perform considerably worse, potentially recording average performance below chance. Nonetheless, we kept strict inclusion criteria in order to optimize model fitting.

For the remaining 39 subjects, we quantified the extent to which each subject exhibited a dissociation between objective task performance and subjective visibility ratings across SOA as follows. For each subject, we ran a least-squares regression between d' and mean visibility rating at all but one SOA. Empirically observed visibility at the left-out SOA was then subtracted from the "expected" visibility predicted by the regression on the other SOAs. We defined the absolute value of this difference between observed and expected visibility for the left-out SOA as the "dissociation score" for that SOA. We calculated the dissociation score for each SOA and defined each subject's "dissociation index" as the maximum dissociation score across all SOA from that subject's data. Each subject's dissociation index provides a measure of the extent to which that subject exhibited a dissociation between task performance and visibility ratings.

We performed a median split on the dissociation index, selecting the 20 subjects who exhibited the highest such value for model fitting. For these 20 subjects, the mean dissociation index was 0.57 and was greater than zero, $p < .001$. For all 39 subjects, the mean dissociation index was slightly weaker but still evident at 0.39 ($p < .001$). Without excluding any subjects at all (n = 59), a similar mean value of 0.41 obtains ($p < .001$).

Note that these procedures were performed in order to improve the quality of the data analysis. Omitting subjects with noisy data reduces the noisiness of model fits. Selecting the subjects who show the strongest dissociations between task performance and stimulus visibility provides a more stringent test for the models and thus provides a sharper way to compare their efficacy in characterizing the data. All subject selection procedures were performed *a priori,* prior to any model fitting analysis.

One might worry that selecting only those subjects with the highest dissociation index creates a skewed or biased sample. However, for data sets that do not exhibit a dissociation, the traditional Single Channel SDT model is sufficient to characterize the data, and the Dual Channel and Hierarchical models can mimic Single Channel model behavior by appropriate adjustment of the model parameters. Thus, data that do not exhibit the dissociation are not informative with respect to the model selection. The logic of this approach is not to claim that one model or another is the "correct" one in the broadest sense, but rather to answer the following question: in those subjects who exhibit a dissociation between objective task performance and subjective ratings of perceptual clarity, what kind of processing structure best accounts for this dissociation? If one model structure clearly outperforms the other, this suggests that we must posit that the supported model structure describes some aspect of human perception, although it leaves open the possibility that other structures may be necessary to account for other kinds of perceptual phenomena.

Model assumptions

In each model, we made standard signal detection theory assumptions, as summarized in Figure 1-3: (1) the two stimuli used in the experiment gave rise to internal signals normally distributed along some decision axis; (2) perceptual decisions were made by comparing the signal on some decision axis to a criterion; and (3) visibility judgments were made by comparing the signal on some decision axis to multiple criteria, corresponding to the multiple visibility ratings available to subjects in this experiment.

In order to further constrain model fitting, we made one further assumption: (4) criteria for perceptual decisions and visibility ratings were set in the same way for each stimulus-mask SOA. That is, we assumed that subjects did not use different standards for deciding a stimulus's identity or visibility across the different SOAs. This assumption is justified by previous psychophysical findings. Gorea and Sagi (2000) found that when stimuli that are easy and difficult to perceive are interleaved randomly,

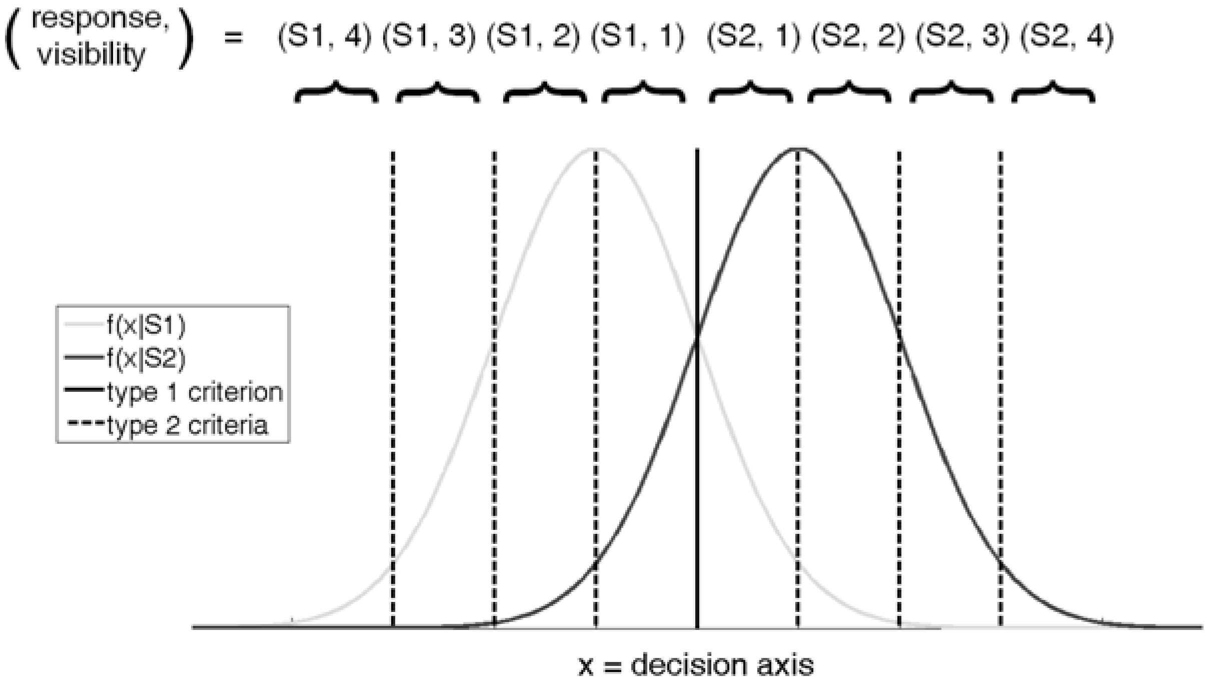
(response, visibility) = (S1, 4) (S1, 3) (S1, 2) (S1, 1) (S2, 1) (S2, 2) (S2, 3) (S2, 4)
f(x|S1)
f(x|S2)
type 1 criterion
type 2 criteria
x = decision axis

Figure 1-3. The standard signal detection theory model. All models under consideration built upon the foundation of the standard signal detection theory model. This model assumes that stimulus categories S1 and S2 each generate normal distributions of perceptual evidence along an internal decision axis. The observer segments the decision axis into discrete regions using a type 1 criterion (for making a stimulus classification response) and a set of type 2 criteria (for rating subjective levels of decision confidence or percept visibility). The stimulus classification and subjective rating reported by the observer on any given trial are determined by which region of the decision axis contains the perceptual evidence observed on that trial, as illustrated in the figure. The probability with which the observer produces a given (response, visibility) pair upon being shown stimulus SN is equal to the area under the curve f(x|SN) in the region of the decision axis corresponding to that response pair. For a more in-depth treatment, see Appendix A.

subjects do not judge stimulus classes with separate criteria, but rather use a single, non-optimal criterion for both. In our experiment, task difficulty varied across SOA, but SOAs were presented randomly, and thus task difficulty changed randomly across trials as it did in Gorea and Sagi (2000). If subjects cannot maintain separate sets of criteria for only two classes of randomly interleaved stimuli, it is highly unlikely that they could maintain seven distinct sets of criteria corresponding to the seven SOAs used in the current experiment.

Furthermore, in a study on the dynamics of criterion shifting, Brown and Steyvers (2005) found that criterion shifting is a slow process. In their experiment, task difficulty changed every 40 trials, requiring subjects to shift their decision criteria in order to maintain optimal task performance. However, even with this predictable block design, and even when subjects were forewarned that task difficulty would change during the experiment, subjects required about 8 - 22 trials (each trial lasting about 3.2 sec) to change their decision criteria. In the current experiment, task difficultly changed randomly and rapidly from trial to trial. The results of Brown and Steyvers suggest that this rapid and random shift in stimulus difficulty would far outstrip subjects' ability to slowly adjust their decision criteria. Taken together, these experimental results suggest that it is unlikely that subjects could have used different sets of decision criteria for each SOA, thus justifying our fourth modeling assumption.

Model descriptions

All models conformed to the broad specifications listed above, but differed from each other in overall model structure (Single Channel, Dual Channel, or Hierarchical). Because there are many ways each model structure can be implemented, we compared multiple kinds of implementations for each model type. In total we fit 4 Single Channel models, 10 Dual Channel models, and 12 Hierarchical models. In the following we give brief descriptions of each model tested. The names of the models in this section correspond to the model names used in Table 1-1.

Single Channel models

Single Channel

parameters: μ_{diff} (8), c (7)

The simplest model we tested was this basic SDT model. We suppose that the distance between the evidence distributions, μ_{diff}, changes for each of the 8 stimulus-mask SOAs. The observer must set 7 decision criteria in order to partition the decision axis into 8 regions, which correspond to the 8 kinds of responses the observer can give on a given trial (2 stimulus classifications * 4 levels of subjective visibility; Figure 1-3; Appendix A). For all models, we suppose that the decision criteria are constant across SOA.

Single Channel CV ("changing variance")

parameters: μ_{diff} (8), σ (8), c (7)

This is a modification of the Single Channel model which supposes that SOA affects not only the absolute distance between the stimulus distributions μ, but also their common standard deviation σ.

Other CV models

For every model listed below, we analyzed versions which did and did not allow the standard deviation of the stimulus distributions σ to vary across SOA. Every model following the naming format "Model X CV" is identical to the simpler model "Model X" with the exception that it has 8 added parameters in order to allow σ to vary with SOA.

Decision Noise

parameters: μ_{diff} (8), σ_c (8), c (7)

This model supposes that the type 2 criteria (the six decision criteria used to evaluate subjective visibility) are not constant from trial to trial, but in fact are drawn from a normal distribution with some standard deviation σ_c, where σ_c can vary with SOA. This model is based on Mueller and Weidemann (2008).

Dual Channel models

Dual channel models suppose that two separate information processing streams accruing noise from independent sources contribute to the perceptual decision making process. In SDT terms, these models posit the existence of two decision axes, one of which corresponds to conscious processing and the other, unconscious processing. The versions of these models considered here differ on how they suppose information from the conscious and unconscious processing channels are combined.

Independent Dual Channel

parameters: $\mu_{\text{diff C}}$ (8), $\mu_{\text{diff U}}$ (8), c_C (6), c_U (1)

The distance between stimulus distributions is modulated by SOA for both the conscious (μ_C) and unconscious (μ_U) decision axes. The conscious decision axis is only used to categorize stimuli that have a visibility of at least 2 or higher, i.e. it is not used to classify stimuli with visibility = 1. For this reason, only 6 decision criteria c_C are set on the conscious decision axis. For stimuli whose visibility is only rated as 1, the stimulus classification is made by doing signal detection on the unconscious decision axis using the criterion c_U. This model is based on Del Cul et al. (2009). (See Appendix B for an explicit comparison between our Independent Dual Channel model and the model used in Del Cul et al.)

Modulated Dual Channel N (N = 1, 2, 3)

parameters: $\mu_{\text{diff C}}$ (8), $\mu_{\text{diff U}}$ (8), c_C (6), c_U (1)

These models are identical to the Independent Dual Channel model, with one exception. Modulated Dual Channel N has a provision for altering subjective reports of visibility made from the conscious decision axis when its stimulus classification conflicts with the stimulus classification provided by the unconscious channel. Specifically, if visibility > 1 and visibility ≤ N+1, and if the stimulus classification of the conscious and unconscious channels disagree, then the classification from the conscious channel is used but the report of subjective visibility is reduced to 1.

Weighted Dual Channel

parameters: $\mu_{\text{diff C}}$ (8), $\mu_{\text{diff U}}$ (8), c_C (6), c_{TOT} (1)

Rather than treat information from the conscious and unconscious channels separately, the observer combines them into a new decision axis by computing a weighted average. The weight given to evidence arising from the conscious channel is $w_C = d'_C / (d'_C + d'_U)$, where $d' = \mu_{\text{diff}} / \sigma$ and $\sigma = 1$ for the non-CV models. This formula can give results outside of [0, 1] if negative d' values are entered. As a correction for this possibility, if the computation yields $w_C < 0$ then w_C is set to 0, and if it yields $w_C > 1$ then w_C is set to 1.

If visibility = 1, the stimulus is classified using the combined channel. If visibility > 1 and the conscious channel and combined channel agree on stimulus classification, then stimulus classification is given with the level of visibility dictated by the conscious channel. But if visibility > 1 and the conscious channel and combined channel disagree on stimulus classification, then the classification from the conscious channel is used but the report of subjective visibility is reduced to 1.

(Although it would be optimal to always use the stimulus classification provided by the combined channel, implementing this in the model would allow the nonsensical result that reports of stimulus classification could conflict with reports of subjective visibility, e.g. “the stimuli were squares, and I very clearly saw that the stimuli were diamonds.”)

Hierarchical models

Hierarchical models suppose that stimulus classification occurs according to Single Channel SDT principles, but that the perceptual evidence used to do stimulus classification changes before it is used to report subjective visibility, becoming weaker and / or noisier.

Decay Only

parameters: μ_{diff} (8), k (8), c (7)

The perceptual evidence used for performing stimulus classification is multiplied by a factor of k before it is used for reporting subjective visibility, where $0 \leq k \leq 1$. k varies across SOA.

Noise Only

parameters: μ_{diff} (8), σ_h (8), c (7)

Mechanisms for reporting subjective visibility access a noisier version of the perceptual evidence used for performing the stimulus classification task. The extra noise is sampled from a normal distribution with mean 0 and standard deviation σ_h. σ_h varies across SOA.

Noise + Decay

parameters: μ_{diff} (8), σ_h (8), k (8), c (7)

A combination of the Decay Only and Noise Only models.

Noise + Constant Decay

parameters: μ_{diff} (8), σ_h (8), k (1), c (7)

Same as Noise + Decay, but the signal decay parameter k is constrained to be constant across SOA.

Constant Noise + Decay

parameters: μ_{diff} (8), σ_h (1), k (8), c (7)

Same as Noise + Decay, but the hierarchical noise parameter σ_h is constrained to be constant across SOA.

Constant Noise + Constant Decay

parameters: μ_{diff} (8), σ_h (1), k (1), c (7)

Same as Noise + Decay, but the hierarchical noise parameter σ_h and signal decay parameter k are constrained to be constant across SOA.

Model fitting

Past efforts to fit signal detection theory parameters to rating data have used the following approach (Ogilvie & Creelman, 1968; Dorfman & Alf, 1969). First, we make two simplifying assumptions: (1) responses on each trial are independent from one another; (2) the probability of each response type associated with each stimulus class is constant across trials. When these assumptions are met, the likelihood of a set of signal detection model parameters given the observed data can be calculated using the multinomial model. Formally,

$$L(\theta \mid data) \propto \prod\nolimits_{r,v,s} Prob_{\theta}(resp = r, vis = v \mid stim = s)^{n_{data}(resp=r, vis=v \mid stim=s)}$$

where “resp” indicates stimulus classification response (square or diamond), “vis” indicates subjective rating of stimulus clarity (1 – 4), and “stim” indicates objective stimulus identity (square or diamond). r, v, and s, are indeces ranging over all possible values for the response, visibility, and stimulus variables, respectively. $Prob_{\theta}(resp = r, vis = v \mid sim = s)$ denotes the probability with which the subject produces the response “r” and visibility rating “v” after being presented with the stimulus “s”. This probability is determined by the SDT model specified with parameters θ. The set of parameters θ for each SDT model under consideration is listed above in the section titled “Model descriptions.” $n_{data}(resp = r, vis = v \mid sim = s)$ is a count of how many times a subject actually produced a response “r” and visibility rating “v” for the stimulus “s”.

The set of parameters θ that maximizes the probability of the data is referred to as the maximum likelihood estimate of the parameters θ, given the observed data. The signal detection models under consideration in this study differ in the distributions of $Prob_{\theta}(resp = r, vis = v \mid stim = s)$ values they can produce, which in turn determines the extent to which the different models can fit the data well and achieve a high maximum likelihood in the multinomial model.

Note that the models were not fit to summary statistics of performance such as percent correct or average visibility. Rather, models were fit to the full distribution of probabilities of each response and visibility rating contingent on each stimulus type. From this full behavioral profile of stimulus-contingent response probabilities, we can derive various summary statistics such as percent correct and average visibility (Figure 1-4), as well as type 2 performance (Figure 1-5). Thus, the behavioral data shown in these figures are not the data upon which the models were explicitly fit, but rather are different ways of highlighting aspects of that data.

We fit all models under consideration to the observed data by finding the maximum-likelihood parameter values θ. Maximum likelihood fits were found using a simulated annealing procedure (Kirkpatrick, Gelatt, & Vecchi, 1983). Model fitting was conducted separately for each subject’s data.

Formal model comparison

The maximum likelihood associated with each model characterizes how well that model captures patterns in the empirical data. However, comparing models directly in terms of likelihood can be misleading; greater model complexity can allow for tighter fits to the data but can also lead to undesirable levels of overfitting, i.e. the erroneous modeling of random variation in the data. The Akaike Information Criterion (AIC), motivated by considerations from information theory, provides a means for comparing models on the basis of their maximum likelihood fits to the data while correcting for model complexity (Burnham & Anderson, 2002). We used AICc, a variant of AIC which corrects for finite sample sizes:

$$AIC_c = -2\log L(\theta \mid data) + 2K\left(\frac{n}{(n-K-1)}\right)$$

where K is the number of parameters in the model and n is the number of observations being fit. In this data set, the number of observations is the number of response probabilities being estimated, so n = 2 (stimulus type) * 2 (response type) * 4 (visibility rating) * 8 (SOA) = 128. Lower values of AICc are desirable because they indicate a higher model likelihood and/or a lower model complexity (lower number of parameters).

We use the likelihood of each model, given the data, as a basis for model comparison:

$$L(model_i \mid data) \propto e^{-\frac{1}{2}\left(AICc_i - AICc_{min}\right)}$$

$AICc_i$ is the AICc for model i and $AICc_{min}$ is the lowest AICc exhibited by all models under consideration. These model likelihoods can be scaled to sum to 1, and the resulting "Akaike weights" reveal the relative weight of evidence for each model as evaluated by its fit to the data, correcting for model complexity.

$$w_i = \frac{e^{-\frac{1}{2}\left(AICc_i - AICc_{min}\right)}}{\sum_{m=1}^{M} e^{-\frac{1}{2}\left(AICc_m - AICc_{min}\right)}}$$

The foregoing analysis can be replicated using the Baysian Information Criterion (BIC) in place of AICc, where

$$BIC = -2\log L(\theta \mid data) + K \log n$$

In this case, calculating the analogue of the Akaike weights gives an estimate of the posterior probabilities of each model, assuming uniform prior probabilities (Burnham & Anderson, 2002):

$$Prob(model_i \mid data) = \frac{e^{-\frac{1}{2}(BIC_i - BIC_{min})}}{\sum_{m=1}^{M} e^{-\frac{1}{2}(BIC_m - BIC_{min})}}$$

Results

Model fitting results

Complete model comparison results are listed in Table 1-1. To simplify analysis, we focus on comparing the best-performing models in each model class. These are the models titled “Single Channel CV,” “Weighted Dual Channel,” and “Constant Noise + Decay.” Details of model specifications can be found in Materials & Methods under the heading “Model descriptions.”

Figure 1-4 displays the fits of these models to stimulus classification accuracy and mean visibility ratings at each SOA. The same data are re-plotted in the bottom panel to show mean visibility as a function of accuracy, so as to emphasize the strong dissociation between the two found in the

Table 1-1. Complete model comparison results.

Class	Model name	log L	# param	Akaike weight	Bayesian posterior probability
Single channel	Single Channel	-1243.4525	15	0.013	0.1025
Single channel	**Single Channel CV**	**-1212.9751**	**23**	**0.1572**	**0.1517**
Single channel	Decision Noise	-1329.3143	23	0	0
Single channel	Decision Noise CV	-1334.7274	31	0	0
Dual channel	Independent Dual Channel	-1233.6579	23	0	0.0001
Dual channel	Independent Dual Channel CV	-1204.2176	31	0.0001	0
Dual channel	Modulated Dual Channel 1	-1242.7603	23	0.0005	0.0001
Dual channel	Modulated Dual Channel 1 CV	-1212.6913	31	0	0
Dual channel	Modulated Dual Channel 2	-1272.0836	23	0	0
Dual channel	Modulated Dual Channel 2 CV	-1244.7851	31	0	0
Dual channel	Modulated Dual Channel 3	-1299.3272	23	0	0
Dual channel	Modulated Dual Channel 3 CV	-1271.9443	31	0	0
Dual channel	**Weighted Dual Channel**	**-1223.7024**	**23**	**0.1391**	**0.1226**
Dual channel	Weighted Dual Channel CV	-1201.0265	31	0.083	0.001
Hierarchical	Decay Only	-1215.7354	23	0.0119	0.0135
Hierarchical	Decay Only CV	-1209.3047	31	0	0
Hierarchical	Noise Only	-1222.3827	23	0.0002	0.0028

Hierarchical	Noise Only CV	-1199.8421	31	0.0763	0.0316
Hierarchical	Noise + Decay	-1199.6372	31	0.0525	0.0443
Hierarchical	Noise + Decay CV	-1196.0596	39	0	0
Hierarchical	Noise + Constant Decay	-1221.6126	24	0.0001	0.0001
Hierarchical	Noise + Constant Decay CV	-1198.2169	32	0.0086	0.0001
Hierarchical	**Constant Noise + Decay**	**-1206.93**	**24**	**0.3012**	**0.3627**
Hierarchical	Constant Noise + Decay CV	-1201.4194	32	0.0006	0
Hierarchical	Constant Noise + Constant Decay	-1233.0909	17	0.0507	0.0654
Hierarchical	Constant Noise + Constant Decay CV	-1204.0892	25	0.1052	0.1014

“Class” denotes model category (see Figure 1-1). Descriptions of each model listed under “Model name” are available in Materials and Methods, Model descriptions. “log L” is the quantitative measure of goodness of fit for each model, the log of the likelihood of the observed empirical data given the model structure and optimal parameter values. Larger values indicate better fit. “# param” lists the number of parameters for each model, a measure of model complexity. “Akaike weight” and “Bayesian posterior probability” are measures of overall model quality, taking into account goodness of fit and model complexity. Larger values indicate better models, and the data is scaled such that both measures sum to 1. For more details on these measures see Materials and Methods, Formal model comparison. The best models in each model class are highlighted in boldface.

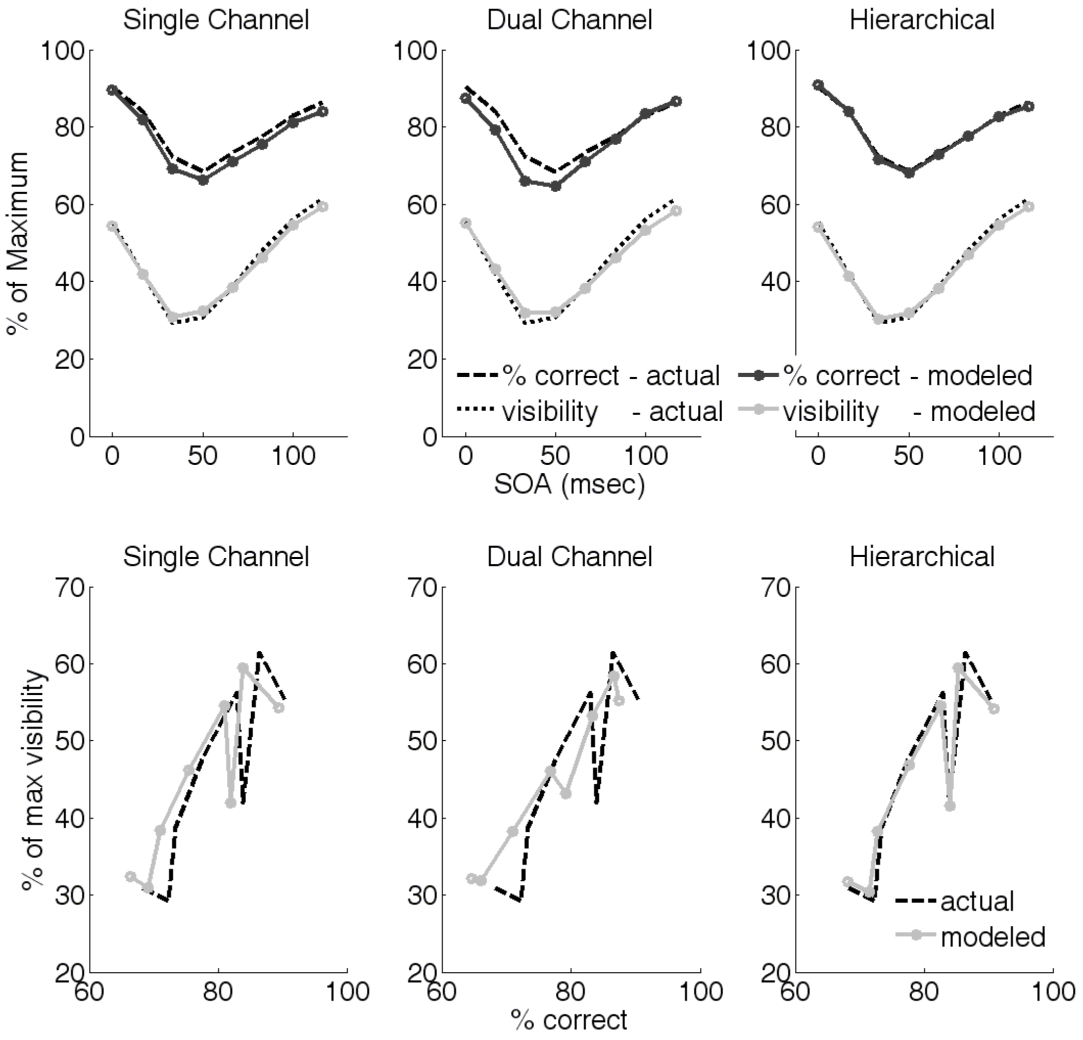

Figure 1-4. Model fits for task performance and reported visibility. Three categories of models (Single Channel, Dual Channel, and Hierarchical) were fitted to the behavioral data from the metacontrast masking paradigm. We tested multiple versions of each category of model (see Materials and Methods for details). Shown here are the best-fitting models from each category, selected according to formal model comparison techniques (Figure 1-6). The Hierarchical model performed best at capturing the dissociation between task performance and reported levels of stimulus visibility. This dissociation is made readily apparent by plotting visibility reports against task performance, as depicted in the bottom row of figures; the relationship is not monotonic, but exhibits a sharp spike at around 80-85% correct, reflecting that short SOAs had lower visibility than long SOAs in spite of having similar task performance.

behavioral data. Visual inspection suggests that the best Single Channel model slightly but systematically overestimates visibility as a function of accuracy, whereas the best Dual Channel model only produces a relatively small dissociation between accuracy and visibility. By contrast, the Hierarchical model provides a close fit to the data.

Another way of probing the relationship between objective task performance and subjective visibility rating is to analyze the behavior of subjective ratings conditioned on accuracy, what has been called “type 2” analysis to distinguish it from the “type 1” analysis of basic stimulus identification performance (Clarke et al., 1959; Galvin et al., 2003). In the top panel of Figure 1-5 we show model fits to type 2 hit rate (HR; p(high visibility | correct)) and type 2 false alarm rate (FAR; p(high visibility | incorrect)), where “high visibility” is defined for each subject as a visibility rating greater than that subject’s median visibility rating across all trials. In the bottom panel we show area under the type 2 ROC curve (estimated using A_g; Pollack & Hsieh, 1969), a measure of how well subjective ratings discriminate between correct and incorrect trials. In general, the basic signal detection theory model predicts that as stimulus classification performance improves, type 2 HR and type 2 FAR should diverge, and area under the type 2 ROC curve should increase (Galvin et al., 2003; Appendix A). However, in this data set type 2 performance is generally lower at longer SOAs than at shorter SOAs, even though task performance is similar at these SOAs. This pattern is difficult for both the Single and Dual Channel models to reconcile; as depicted in Figure 1-5, both overestimate type 2 performance, particularly at long SOAs. The Hierarchical model gives a reasonably good overall fit to the type 2 data, although even this fit did not produce a difference between type 2 performance at short and long SOAs as pronounced as in the data.

The results reported in Figure 1-5 are easy to intuit. For the Single Channel model, area under the type 2 ROC curve is already largely determined by specifying type 1 task performance, i.e. the perceptual evidence distributions and the type 1 criterion (Galvin et al., 2003; Appendix A). The strong

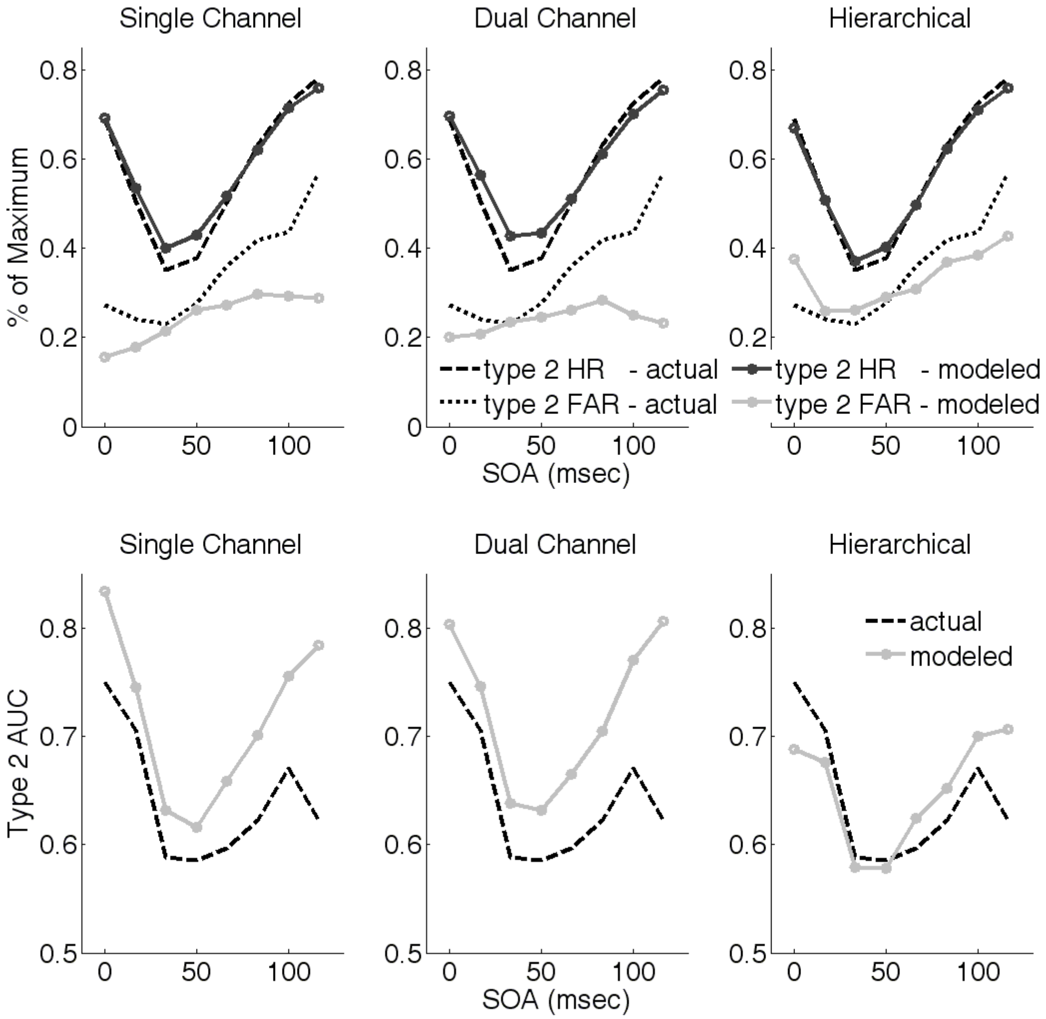

Figure 1-5. Model fits for type 2 data. In addition to the distinctive dissociation between task performance and visibility (Figure 1-4), the behavioral data also included a set of type 2 data that provided a challenge for model fitting. By "type 2 data" we refer to the probability of giving different levels of visibility ratings conditional upon task performance. (Top panel) Type 2 hit rate (HR; probability of high visibility for correct responses) and type 2 false alarm rate (FAR; probability of high visibility for incorrect responses) as a function of SOA. (Bottom panel) Area under the type 2 ROC curve as a function of SOA. In general, signal detection theory models predict that as task performance increases, type 2 hit rate and type 2 false alarm rate should diverge and area under the type 2 ROC curve should increase. In this data set, type 2 performance at long SOAs was worse than at short SOAs even though task performance was similar, a pattern difficult to reconcile for standard signal detection theory and its Dual Channel modification, but more closely approximated by the Hierarchical model.

relationship between stimulus classification performance and type 2 performance is essentially due to the fact that they are based on the same underlying information; there is no additional process by means of which the quality of information available to type 1 and type 2 mechanisms could differ. Thus, this fundamental assumption of the Single Channel models makes them somewhat inflexible in capturing the relationship between type 1 and type 2 data, particularly like those in the current experiment where area under the type 2 ROC curve exhibited a dissociation from task accuracy across SOA (compare Figure 1-4 and Figure 1-5). In principle, Single Channel models can reduce type 2 performance without affecting classification accuracy by supposing that type 2 criterion setting is a noisy process, such that the placement of the criteria varies randomly from trial to trial (Mueller & Weidemann, 2008), but this class of models gave poor overall fits to the current data set (Table 1-1).

One may expect the Dual Channel model to fare better because it postulates two different processes. However, this was not the case. The reason is that the “conscious” channel essentially acts like a Single Channel model, supposing a tight relationship between task performance and subjective visibility, and the “unconscious” channel is limited in the extent to which it can interfere with fully “conscious” processing. (For instance, one would not expect an observer to make a subjective report of having a distinct visual awareness of seeing squares and yet simultaneously claim that the stimuli presented on the screen were diamonds.) It is possible that Dual Channel models featuring more extensive and complicated interactions between the two channels could fare better, but such models would potentially constitute a departure from the fundamental dichotomy between “conscious” and “unconscious” processing streams that arguably is the main conceptual motivation for proposing the Dual Channel class of models. As it stands, the best Dual Channel model we tested already posits that in cases of conflict in the stimulus classification response, the “unconscious” channel can modulate visibility ratings made by the “conscious” channel; simpler Dual Channel models that better respected the distinction between “conscious” and “unconscious” processing performed worse (Table 1-1).

By contrast, the dissociation between type 1 and type 2 performance is more naturally captured by Hierarchical models, as they stipulate a less restrictive relationship between the quality of information available for type 1 and type 2 decision making. Changing the degree to which the evidence becomes degraded at the second stage of processing provides a means of changing the patterns of subjective rating without affecting basic task performance, which is determined by the first stage of processing.

As models become more complex, in general they become better able to capture real patterns in data, but also become more prone to erroneously capture noise in the data (overfitting). Thus, one approach to conducting formal model analysis involves using metrics like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), which reward models for closeness of fit to observed data while punishing them for complexity (number of parameters). In Figure 1-6 we present model comparison results based on a finite-sample correction of AIC, AICc (Burnham & Anderson, 2002). Overall, the hierarchical category of models collectively outperformed the Single Channel and Dual Channel models (top panel), and this pattern held up when comparing only the best models in each category (bottom panel). Essentially identical results are found using BIC (Table 1-1). Thus, the superior goodness of fit for the Hierarchical model evident in Figures 4 and 5 cannot be written off to overfitting. In fact, the three best models in each model category, though visibly differing in quality of data fitting, had essentially the same number of parameters (Single Channel and Dual Channel, 23; Hierarchical, 24).

Parameter values for the model fits

We can derive further insight into the way the best models in each category captured the data by investigating their parameter values (Figure 1-7).

The fit for the Single Channel model indicates a U-shaped curve for σ, the standard deviation of the perceptual evidence distributions, such that σ takes on higher values at longer SOAs. When criteria

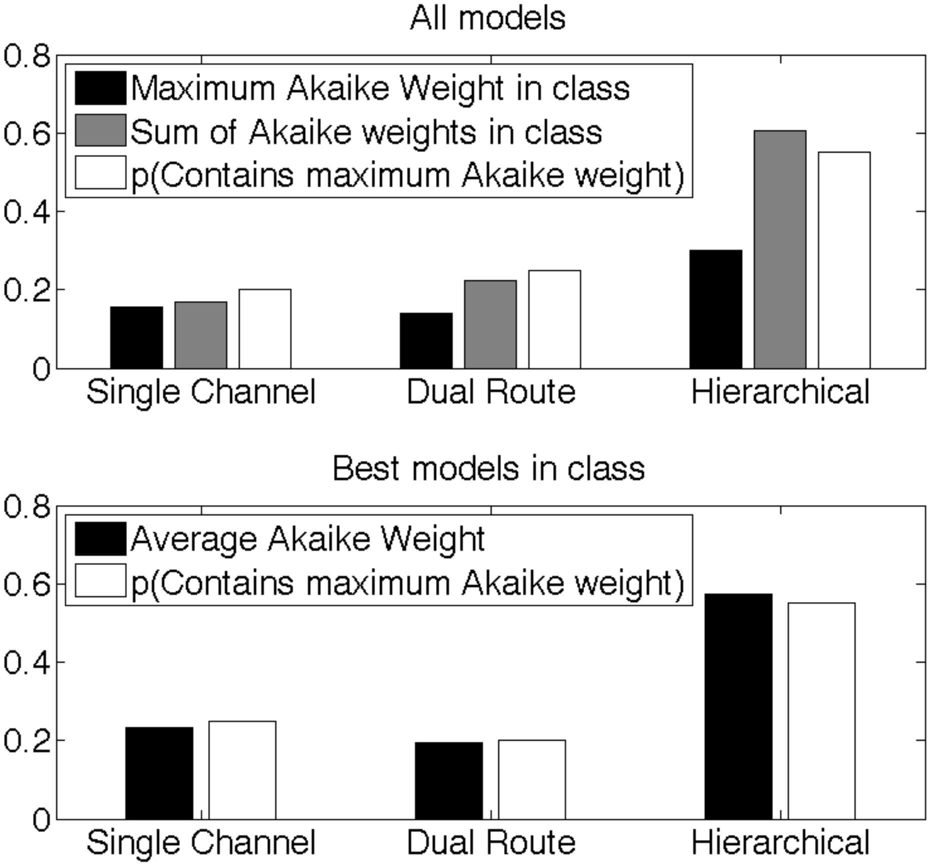

Figure 1-6. Model selection results. Formal model comparison was conducted using a finite sample size correction of the Akaike Information Criterion (AICc), which rewards models for closely fitting observed data while punishing models for the degree of complexity (i.e. number of free parameters; for list of free parameters for all models please see Materials and Methods). For ease of interpretation we display a transformation of AICc values into Akaike weights, which quantify the information theoretic evidence in favor of each model such that the weights sum to 1 (Burnham and Anderson, 2002). (Top panel) Model selection on all 26 models. The best hierarchical model had an average Akaike weight roughly twice as large as those of the best single channel and dual channel models. Similarly, the best model for each subject was roughly 2 – 3 times as likely to belong to the hierarchical class than to the other two model classes. The average Akaike weight summed across all hierarchical models was roughly 3 – 4 times as great as the Akaike weight sums for the other two classes. (Bottom panel) Similar results were found when restricting the analysis to the best models in each model class. Model comparison results were nearly identical when using the Bayesian information criterion to estimate the posterior probability of each model (Table 1-1).

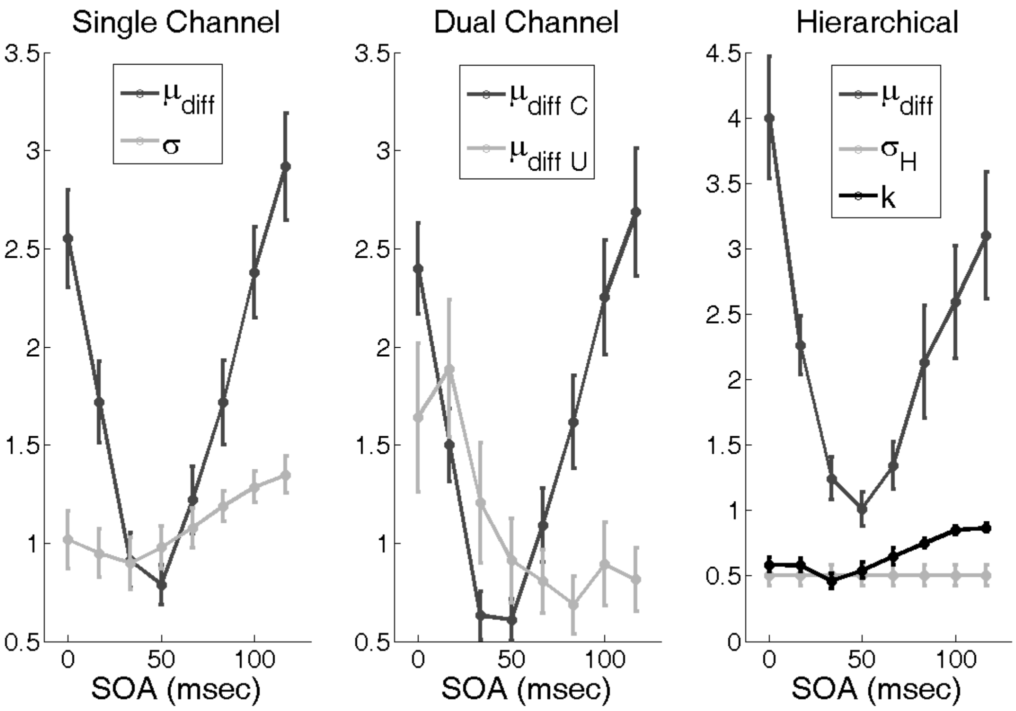

Figure 1-7. Parameter values from model fits. For descriptions of model structure and parameters, see Methods.

are held constant across SOA (a stipulation for all models, see Materials and Methods), larger values of σ entail higher levels of mean visibility rating (see e.g. Figure 1-3), and yet the model can predict similar levels of task performance at short and long SOAs since task accuracy depends on d' = μ_{diff} / σ. In this way, provided that the standard deviations of the evidence distributions can vary independently from their distance, the Single Channel model can capture the accuracy / visibility dissociation in the behavioral data (Figure 1-4). Thus, in order for the Single Channel model to capture this data, it must assume that the variance of the internal signal is highest at long SOAs where task performance and visibility are maximal. Although such a Poisson-like correlation of signal and noise is not in itself implausible, the specific patterns predicted are some cause for doubt. For instance, the model predicts that on average, the perceptual signal μ_{diff} at SOAs 0 ms and 100 ms is roughly equal, and yet the simultaneous presentation of stimulus and mask is less noisy than when their presentation is separated by a full 100 ms. It seems more likely that, controlling for the magnitude of the absolute signal, stimulus representations should be noisier when the mask is presented simultaneously than when the mask is presented 100 ms later.

The Dual Channel model predicts that perceptual sensitivity is greater in the “unconscious” channel than in the “conscious” channel for several short SOAs. Because this model resets visibility ratings to 1 when the two channels disagree on stimulus classification, setting the sensitivity of the “unconscious” channel higher at the short SOAs has the effect of increasing the frequency of disagreements between the two channels, thus reducing visibility at those SOAs without having a drastic effect on task performance. This allows visibility to be lower at shorter SOAs than at longer ones even though task performance at those SOAs is similar. However, the model only manages to produce a somewhat weak dissociation (Figure 1-4). Furthermore, it seems unlikely that processing in an unconscious channel could be so robustly high and consistently superior than conscious processing across several SOAs.

The Hierarchical model predicts that perceptual evidence decays in the second stage of “subjective” processing more readily at short than at long SOAs, thus leading to lower overall levels of visibility at the short SOAs in spite of similar stimulus discrimination sensitivity. By contrast, the model supposes that noise at the late processing stage is independent of SOA. This seems plausible if we imagine that signal transmission from early to later stages of perceptual processing depends in part on the processing that occurs in early sensory areas, whereas the noise intrinsic to later processing stages is independent of the noise in earlier stages.

The structure and parameter values of the Hierarchical model are also consistent with previous empirical findings from experiments focusing on the dissociation between objective task performance and subjective ratings of visibility. For instance, Lau and Passingham (2006) used a similar metacontrast masking paradigm as in the present study, and in the fMRI scanner they focused on a short and a long SOA where task performance was matched, and yet the subjective ratings of visual awareness differed. Higher subjective ratings of visibility at the long SOA were associated with higher level of activity in the dorsolateral prefrontal cortex. Interestingly, no significant difference in level of fMRI activity was found in posterior sensory areas. This is compatible with the Hierarchical model if we assume that the prefrontal activity reflects the hierarchical model's late stage process. Indeed, according to the parameter values of the best Hierarchical model (Figure 1-7), reported visibility was higher at the long SOAs than it was at the earlier, performance-matched SOAs due to a superior transmission of perceptual evidence to the late processing stage (i.e., higher values for the parameter k). This corroborates well with the fMRI result.

The two key-press design of the task did not favor the Hierarchical models

One might worry that the design of the current experiment is biased in favor of the Hierarchical model. We required subjects to report stimulus visibility after they reported stimulus identity, with a second key press. Perhaps signal degradation did occur between the "objective" and "subjective" decisions, in a fashion predicted by the Hierarchical model, but only because the design forced subjects to report visibility *after* reporting their perceptual decisions. This timing difference between the two key presses could trivialize our findings.

However, the implicit reasoning behind this argument is that signal degradation could be artificially introduced by increasing response time. The longer the subject takes to respond, the more degraded a signal presumably becomes. If this deflationary account of the modeling results were true, we might expect that the Hierarchical model's estimated values of signal decay and late processing noise should correlate with the time separating the stimulus classification key press from the subjective rating key press (henceforth, "rating RT"). However, the across-SOA correlation between estimated signal decay and rating RT was not significant for any subject (*p*s > .15), and the average correlation did not differ from zero (Fisher's r-to-z transform, *p* = .4; Fisher, 1915).

Since the parameter for late processing noise was constant across SOA for the best Hierarchical model, we cannot compute within-subject correlations of this parameters with rating RT. We did find that across subjects, the estimated amount of late noise correlated with average rating RT, *r* = -.48, *p* = .03. However, this result is in the opposite direction of that proposed by the trivializing critique regarding two separate key presses. That is, longer rating RTs were associated with smaller, rather than larger, estimates of late-stage processing noise.

Finally, we note that rating RT was not modulated by SOA (*p* = .4) and that the average rating RT was relatively small (426 ms). This suggests that the time between the first and the second key presses was mainly for motor preparation, i.e. subjects probably made both objective and subjective decisions, and then pressed two keys to reflect them in quick succession without much "thinking" in between. In our subjective experience this is how one would perform the task as well. Taken together, these results suggest that the success of the Hierarchical model in fitting the data cannot be trivially attributed to the two key press design of the task.

Discussion

In order to compare models of how subjective reports of visibility relate to objective perceptual processing, we collected data from a metacontrast masking paradigm that has been shown to induce dissociations between stimulus classification accuracy and reported levels of visibility across different levels of stimulus-mask onset asynchrony (SOA) (Lau & Passingham, 2006). We reasoned that the unusual, nonlinear relationship between accuracy and visibility across SOA (Figure 1-4) would pose a challenge to models of perceptual decision making, and thus prove useful for distinguishing amongst them. The data contained another dissociation that also proved difficult for the models to capture:

visibility ratings were more predictive of task accuracy at short than at long SOAs (Figure 1-5), even though stimulus classification accuracy at these SOAs was similar. Overall, the Hierarchical model provided the best and most parsimonious fit to the data. The model parameters it used to fit the data also seem plausible (Figure 1-7), and overall the model seems compatible with previous empirical findings (Lau & Passingham, 2006).

Why was the Hierarchical model successful where the Single Channel and Dual Channel models were not? The best-performing Hierarchical model (Constant Noise + Decay) was able to accommodate the relative dissociation between task performance and visibility ratings by supposing that early-stage perceptual processing is better transmitted to late-stage processing at long than at short SOAs. Because the early stage governs task performance and the late stage governs subjective reports, this allows for long SOAs to have higher subjective visibility than short SOAs in spite of having similar task performance.

This concept of differential transmission of information from earlier to later stages of processing bears some similarity to notions of processing bottlenecks in multi-stage processing hierarchies. For instance, Chun and Potter (1995) accounted for the attentional blink with a two-stage model wherein the processing of an initial stimulus interferes with the access of subsequent stimuli to late-stage processing responsible for mechanisms of conscious access and report. However, the attentional blink concerns interrupted access for a *subsequent* stimulus presented 150 – 400 ms *after* an initial stimulus, whereas here we are concerned with the interrupted processing of an *initial* stimulus presented ~16 ms *prior to* a mask. Furthermore, whereas the attentional blink involves interruption of objective stimulus processing, the effect we are focusing on in the metacontrast masking paradigm concerns suppressed subjective processing of a stimulus whose objective processing is otherwise intact.

The best-performing Single Channel model (Changing Variance) was able to accommodate this pattern to some extent by supposing that perceptual processing becomes more variable at long SOAs, thus producing sensory signals that more frequently exceed the observer’s criteria for producing high visibility ratings. However, although this model captured the gist of the performance-visibility dissociation, it sometimes produced too-high estimates of visibility ratings or too-low estimates of task performance (Figure 1-4, lower left panel).

By comparison, none of the Dual Channel models we considered appeared to capture the performance-visibility dissociation particularly well. Our SDT implementation of the Independent Dual Channel model (which most closely followed the model of Del Cul et al. (2009); see Appendix B) essentially acts like a Single Channel model with added flexibility for adjusting task performance at the lowest level of subjective visibility. This provides only a relatively limited mechanism for adjusting the relationship between task performance and visibility; holding the parameters of the “conscious” channel constant, changes in the “unconscious” channel can only influence task performance to the extent that subjects report the lowest level of subjective visibility. Thus, this model can accommodate only relatively small differences in task performance for conditions with similar mean levels of reported visibility. Additionally, because task performance at higher (presumably conscious) visibility levels cannot be affected by changing parameters of the “unconscious” channel, this model makes the relatively strong prediction that whatever differences in task performance do occur for visibility-matched conditions, they should arise purely from differences in task performance for trials with the lowest visibility rating. The best-performing Dual Channel model (Weighted Dual Channel) was somewhat more flexible, but still did not adequately capture the dissociation (Figure 1-4, bottom center panel).

In addition to the performance-visibility dissociation across SOA, we also found that the models differed in their ability to capture the degree to which visibility ratings were diagnostic of accuracy on a trial to trial basis. Visibility ratings for incorrect responses at short and long SOAs were generally higher than the model fits (Figure 1-5 top row), and the ability of visibility ratings to predict accuracy was generally lower than the model fits (Figure 1-5 bottom row). The Hierarchical model performed best at capturing these data because it posits that the sensory signal accrues additional noise at late processing stages. This reduces the information that such sensory signals carry regarding task performance on the trial level, which manifests as lower area under the type 2 ROC curve. By contrast, the Single Channel model posits that the same sensory evidence is used to make both the objective response and the visibility rating, and thus is considerably less flexible in the relationships it allows between task performance and type 2 accuracy (Maniscalco & Lau, 2012; Appendix A). Dual Channel models behaved similarly to Single Channel models in this respect, as they primarily differed with respect to processing at low levels of visibility.

All models we tested were constructed using signal detection theory (SDT) as a basis (Figure 1-3; Appendix A; Green & Swets, 1996; Macmillan & Creelman, 2005). In this work, SDT provided an ideal basis to compare overall model architectures in a simple but powerful framework. SDT is sufficiently powerful to be able to dissociate perceptual sensitivity from response bias—essential for the study of perceptual decision making and subjective reports of visibility—while also being sufficiently general as to be readily adapted to different model architectures. Using the same SDT framework for all models also facilitated direct model comparison by minimizing idiosyncratic computational differences between the models. Because our SDT models captured the core computational principles lying behind broadly divergent theories of how perceptual decision making and subjective visibility are related, the model comparison analysis sheds light on these broad conceptual issues.

One limitation to this approach is that the conclusions we have drawn may be somewhat specific to the particular SDT implementations we have used. (However, see Appendix B for evidence that our SDT implementation of the Independent Dual Channel model behaves similarly to the dual channel accumulation model in Del Cul et al. (2009).) Nonetheless, the relative simplicity of the SDT models we have chosen, in conjunction with the broad differences in the model classes being compared (Figure 1-1), would seem to mitigate such concerns. We have also endeavored to perform an unusually comprehensive analysis that directly compares a wide range of models' ability to account for the data, rather than simply demonstrating that a single model can produce reasonable fits to the data.

We also acknowledge that this analysis is driven by the current data set and is thus limited in its generalizability. For instance, it is possible that a Dual Channel model may perform better for capturing other kinds of empirical phenomena. Future work employing similar formal comparison strategies needs to be performed in these cases.

Are the models biologically realistic?

On the face of it, the models we considered depict a purely feedforward style of information processing. What of the fact that anatomically, the most related brain regions are linked by both feedforward and feedback connections? For instance, for the Hierarchical model it is perhaps natural to think of the first stage as representing processing in the early sensory regions in the brain, and the second stage as representing processing in higher regions such as the prefrontal cortex. In this sense, the model ignores the presence of top-down modulations from prefrontal cortex to early sensory areas. However, formally the model does not necessarily commit to such anatomical identifications. Strictly speaking, the model is agnostic as to whether the late stage is mediated by a feedforward or feedback process; late stage simply means it is late in the stream of information processing and thereby inherits the noise of earlier stages.

Even on the plausible and intuitive interpretation that in the Hierarchical model the first stage reflects early sensory processes and the second stage fronto-parietal processes, the model does not deny the existence of feedback connections. Nor does it deny the existence of parallel pathways as intuitively depicted by the Dual Channel model. The Hierarchical model suggests that *with respect to explaining* the relationships and potential dissociations of objective stimulus responses and subjective visibility ratings, the essential relevant structure of processing is hierarchical. This does not mean that the Hierarchical model explains all facts regarding brain processes or subjective experience. It is for the same reasons that the Single Channel model cannot be rejected on the grounds that the brain is clearly more complex than a single-stage processor.

Implications for theories of visual awareness

One currently popular theory suggests that feedback, and specifically feedback from extrastriate to primary visual cortex, is essential for visual awareness (Lamme, 2006; Block, 2007). One might take the point of view that the feedforward wave of processing from primary visual cortex to extrastriate areas represents an early stage of processing, and that feedback represents a second stage of processing, such that together they form a hierarchy.

Another dominant theory of visual awareness is the global workspace theory (Dehaene et al., 2003; Dehaene et al., 2006), according to which early sensory processing itself does not support conscious experience. In order to enter consciousness, the early perceptual signal must propagate into a second stage of processing mediated by a global workspace structure located in prefrontal and parietal cortices. Considerations like these may give the impression that both theories of visual awareness discussed above are compatible with the Hierarchical model.

However, it is important to emphasize that the present work focuses on the dissociation between objective task performance and subjective reports. According to the Hierarchical model, manipulation of the second stage of processing changes subjective reports but not task performance. But the feedback model and the global workspace model would not make such predictions. In these models, the supposed second stage of processing supports both subjective experience as well as amplification of the perceptual signal itself, which is essential for objective task performance. Thus, according to these theories, if the second stage of processing (feedback to striate cortex, or global workspace activity) is disrupted, both objective task performance and subjective reports will be affected. Therefore, these models bear more functional resemblance to the Single Channel models than the Hierarchical models. In order for such theories to obtain a reduction in subjectively reported level of awareness while keeping task performance constant, one natural solution would be that the perceptual signal from a separate, unconscious channel (e.g. a subcortical route) would need to be increased to compensate for the signal loss in the "conscious" channel. In other words, a Dual Channel model would need to be stipulated.

Therefore, as far as dissociations between task performance and subjective reports are concerned (e.g. when we are specifically trying to explain the kind of performance-matched difference in subjective rating and type 2 performance depicted in Figures 1-4 and 1-5), both aforementioned theories are more congenial with Single Channel and Dual Channel models than with Hierarchical models (Del Cul et al., 2009; Lau, 2011). The present results are thus surprising, or maybe even problematic, for these theories.

Viability of the metacontrast masking paradigm for dissociating objective and subjective processing

Recent research has called into question the viability of the metacontrast masking paradigm used here and previously (Lau & Passingham, 2006) for the purposes of dissociating ratings of awareness from objective task performance. These objections are based upon a putative difference in the nature of stimulus processing at short and long SOAs, such that a direct comparison between the two is problematic.

Jannati and Di Lollo (2012) argue that the "criterion contents" used to guide behavioral responses at short and long SOAs in Lau and Passingham's stimulus set differ. At short SOAs, the square / diamond target may perceptually 'fuse' with the metacontrast mask. In Lau and Passingham's stimuli, square and diamond targets were always presented at the same contrast, and the mask was similar to the mask displayed in Figure 1-2, with the exception that the 'star' shape of the mask was embedded within a solid-colored circle, rather than being a sparse line drawing as in the current experiment. Thus, at short SOAs, the square/diamond target may have perceptually fused with the circular mask, such that the salient perceptual feature for distinguishing between squares and diamonds would be the small gaps between target and mask, rather than the luminance-defined shape of the target itself (see Jannati and Di Lollo's Figure 1). Thus, the salient perceptual feature used to evaluate the stimulus—the "criterion content"—may have differed at short (orientation of target-mask gaps) and long (luminance-defined target shape) SOAs. To support their interpretation that the performance / awareness dissociation occurs due to differing criterion contents at short and long SOAs, Jannati and Di Lollo performed a separate metacontrast masking experiment in which they argue that the criterion content is the same at short and long SOAs. In this experiment, they failed to find short and long SOA pairs in which task performance was the same and reports of awareness significantly differed. If the criterion content does in fact differ between short and long SOAs, then a direct comparison between the two is problematic. If reports of stimulus visibility qualitatively differ, e.g. in regards to the perceptual feature which they evaluate, then direct quantitative comparison between them may not be meaningful.

However, although Jannati and Di Lollo failed to find performance-matched SOAs that exhibited differences in awareness in their revised experiment, there was nonetheless a Measure (Performance/Awareness) x SOA interaction, significant at the $p < .001$ level, due to the fact that the awareness curve was lower than the performance curve at short but not long SOAs. Thus, contrary to their interpretation of the findings, their results in the revised, single criterion content experiment seem to be consistent with the general pattern we have observed here and previously. As a result, their interpretation that the dissociation found in Lau and Passingham is due to a criterion content confound, such that removing this confound also removes the dissociation, does not seem justified by their data.

Additionally, our stimuli differ in important respects from those used in Lau and Passingham (2006) and Jannati and Di Lollo (2012). The metacontrast mask used in the current experiment was significantly less perceptually salient due to being a 1-pixel wide line drawing rather than a shape embedded in a solid-colored circle. Additionally, we adjusted stimulus contrast in order to control task performance, such that the stimulus contrast for each subject (mean Weber contrast = -0.11) was substantially lower than the mask contrast (Weber contrast = -1). Thus, even at short SOAs where the target and mask may have perceptually fused, luminance-defined shape of the target remained a highly salient perceptual cue for performing the shape discrimination task, and in this respect the same criterion content (luminance-defined target shape) was perceptually available across all SOAs.

Sackur (2013) used a multidimensional scaling analysis to investigate the perceptual dimensions underlying perceptual performance in metacontrast masking tasks. He argued that his analysis revealed three salient perceptual dimensions: one coding for SOA, and the other two coding separately for perception at short and long SOAs. Sackur argued that this finding upholds the idea that criterion content for metacontrast masking stimuli differs at short and long SOA. However, in Sackur's metacontrast masking task, the target was a square, and the mask was a thick square frame with the same contrast as the target. Unlike the masks we have discussed to this point, the inner perimeter of this square-frame mask perfectly traced the contours of the square target, leaving no gaps. Thus, at short SOA when the target and mask were presented simultaneously, the target was literally no longer discernable as such; what was physically presented on the screen was simply an unusually large square, i.e. the thick frame of the mask with its center filled in by the target. Indeed, in Sackur's analysis, the dimension coding for perception at short SOA time scales primarily differentiated the two SOA during which the target and mask physically overlapped in this way (SOA = 0 ms and 10 ms) from the remaining SOA (Sackur's Figure 5). Thus, the difference in perception at short and long SOAs found in Sackur's study is readily attributable to the fact that an unusually large version of the target stimulus (large square mask filled in with target) was briefly presented at short SOAs but not at long SOAs. No such disanalogy between short and long SOA is present in the stimuli used in the current experiment, and so

Sackur’s results do not seem to provide reason to believe that criterion content differed as a function of SOA with these stimuli.

Potential relations to the memory literature

It has been proposed that there are two distinct and dissociable memory systems, one supporting explicit, “conscious” recollection, and the other more relevant for vaguer judgments of familiarity or feelings of knowing, or unconscious priming behavior (e.g. Jacoby, 1991; Hintzman & Curran, 1994). However, it has also been argued that a single system view is more parsimonious (Squire, Wixted, & Clark, 2007; Wixted, 2007; Berry, Shanks, & Henson, 2008), and that the apparent dissociation between conscious recollection and unconscious memory is due to different levels of activation within the same system. Our results may contribute to this controversy, because the paradigms used in some of these memory studies are conceptually very similar to the paradigm used here: subjects make an objective judgment about the state of the world (identity of visual stimulus, or whether an item has been presented previously or not), and then make a subjective judgment about how they subjectively feel about the first-order process (high vs low visibility, or “Remember” vs “Know” in some memory studies). Here we offer a third alternative to this debate between a single system versus two dissociated systems: it could be that there are two processes that work in hierarchy. Future studies may employ the same model comparison method to arbitrate which is the best model for memory function by fitting the models to experimental data where the objective memory performance and the subjective reports of recollection experience dissociate.

Conclusion

Here we introduce a distinction between different signal processing architectures supporting the generation of subjective reports of visual awareness. Above we discussed some limitations of this approach, such as that it depends on the specific fitted dataset. Regardless of whether these results hold true, one important message is that we can go beyond the traditional assumption that perception depends on a single decision making process (Green & Swets, 1966; Macmillan & Creelman, 2005). These simple single process models have enjoyed great success in explaining many aspects of perception, and remain powerful contenders because of their simplicity, as shown in our model comparison analysis (which punishes complex models). But in cases where objective task performance and subjective reports dissociate, it may be important to consider perceptual decision models that postulate more than a single process, at least as possibilities. Our investigation suggests that, of the two models which postulate two processes, the Hierarchical model is superior to the Dual Channel model.

Chapter 2

Manipulation of working memory contents impairs relative metacognitive sensitivity in a concurrent visual discrimination task

Introduction

As we showed in Chapter 1, a Hierarchical modeling structure provides a good explanation for the behavioral phenomenon of relative blindsight discovered by Lau and Passingham (2006), whereby objective stimulus discrimination performance and subjective reports of stimulus visibility can dissociate. The functional structure of early and late processing stages of the Hierarchical model is suggestive of a mapping onto the anatomical structure of neural sites situated in earlier and later stages in the processing stream of visual information. Consistent with this notion, Lau and Passingham (2006) found that when subjective reports of visibility differed for target-mask stimulus onset asynchronies in which objective stimulus discrimination was matched, the elevated reports of subjective visibility were associated with enhanced activity in dlPFC but not in earlier, more posterior stages responsible for visual processing.

Several additional lines of evidence in the literature link higher-level portions of the frontal cortex function (including dorsolateral prefrontal cortex (dlPFC), rostrolateral prefrontal cortex (rlPFC), and anterior prefrontal cortex (aPFC)) to visual metacognition. Activations in dlPFC and rlPFC have been found to inversely correlate with reports of confidence in visual and memory tasks (Henson, Rugg, Shallice, & Dolan, 2000; Fleck, Daselaar, Dobbins, & Cabeza, 2006; Fleming, Huijgen, & Dolan, 2012), and rlPFC activations have been found to directly (in a memory task; Yokoyama et al., 2010) and indirectly (in a visual task; Fleming et al., 2012) correlate with metacognitive sensitivity. Individual differences in gray matter volume in aPFC positively correlate with visual metacognitive sensitivity (Fleming et al., 2010; McCurdy et al., 2013), and single unit recording activity in macaque aPFC has been shown to increase following correct decisions in a cuing task, even before task feedback is provided (Tsujimoto, Genovesio, & Wise, 2010). Finally, as we will show in Chapter 3, transcranial magnetic stimulation to bilateral dlPFC can impair metacognitive sensitivity.

dlPFC is also involved in working memory (WM) performance. Multiple lines of evidence implicate dlPFC particularly in the active processing of WM contents, rather than the mere storage of WM contents, which is typically attributed to more posterior brain regions, e.g. parietal and occipital cortex (Petrides, 2000; Miller & Cohen, 2001; Curtis & D'Esposito, 2003). For instance, dlPFC activations during delay periods in WM tasks increase when the task requires WM contents to be manipulated (D'Esposito, Postle, Ballard, & Lease, 1999), and other studies have found that dlPFC does not preferentially activate during delay periods, but rather its activation profile reflects the specific process of response selection performed on the basis of WM contents (Rowe, Toni, Josephs, Frackowiak, & Passingham, 2000; Rowe & Passingham, 2001; Rowe, Friston, Frackowiak, & Passingham, 2002). Basic short-term memory performance can be spared in patients with bilateral prefrontal damage (Petrides, 1989; Owen, Morris, Sahakian, Polkey, & Robbins, 1996), but dlPFC lesions impair performance on tasks that require active monitoring and manipulation of WM contents in humans (Petrides & Milner, 1982) and macaques (Petrides 1995). dlPFC also becomes more activated in WM tasks in which a cognitive strategy allows WM contents to be "chunked" into higher-level units, even though such chunking strategies effectively reduce the number of "items" in WM (Bor, Duncan, Wiseman, & Owen, 2003). This finding again suggests that dlPFC is more closely linked to strategic monitoring and manipulation of WM contents than it is to the overall difficulty of the memory task or to the number of items that need to be stored in WM.

Given that PFC is recruited in both metacognition and executive processing of WM contents, it is possible that common underlying mechanisms are at play in both kinds of cognitive functions. If so, we might expect that metacognitive performance would be selectively impaired by concurrently manipulating WM contents, especially in light of general processing capacity limits and bottlenecks in PFC (Marois & Ivanoff, 2005). Here we test this hypothesis in a dual-task paradigm. While holding a letter string in memory and alphabetizing it, subjects performed a simple 2-interval forced choice visual task and provided confidence ratings. After the visual task, a probe assessed memory for the alphabetized string. We analyzed metacognitive performance under low and high WM load. Within the high WM load condition, we further distinguished between trials that placed low and high manipulation demand (i.e. strings requiring little or extensive alphabetization). To anticipate, we found that metacognitive performance was selectively impaired under high WM load with high manipulation demand, suggesting that a common mechanism contributes to metacognitive evaluation of perceptual decision making and active manipulation of working memory contents.

Methods

Experiment 1

Participants

Twenty-three Columbia University students participated in the experiment. Participants gave informed consent and were paid $10 for approximately one hour of participation. The research was approved by the Columbia University's Committee for the Protection of Human Subjects.

One participant was omitted from data analysis, due to producing outlying data in the perceptual metacognitive task under high working memory load (Figure 2-4).

Experimental procedure

Subjects were seated in a dimmed room 60 cm away from a computer monitor. Stimuli were generated using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) in MATLAB (MathWorks, Natick, MA) and were shown on an iMac monitor (LCD, 24 inches monitor size, 1920x 1200 pixel resolution, 60 Hz refresh rate).

On every trial, a working memory task was performed concurrently with a visual discrimination task (Figure 2-1). At the start of the trial, an uppercase letter string in black font was displayed on a gray background for 2000 ms. The letter string could consist of either one letter (low WM load) or four letters (high WM load). The across-trial sequence of one- and four-letter string presentations was randomized, such that each string size occurred with equal frequency. Letters in the four-letter strings were presented in random alphabetical order. Letters were chosen randomly from the following letter bank: {F, G, H, J, K, L, M, N, P, Q, R, S, T}. Vowels and letters early and late in the alphabet were omitted to increase memorization and alphabetization difficulty. Subjects were instructed to hold the letter string presented at the start of the trial in memory and to alphabetize it, since memory for the alphabetized string would be probed at the end of the trial.

After the letter string was presented, a crosshair (.35° wide) was presented centrally for 500 ms, and then the stimuli for the visual discrimination task were presented. Two stimuli were presented simultaneously for 33 ms, one 4° to the left of fixation and one 4° to the right. Each stimulus was a circle (3° diameter) consisting of randomly generated visual noise. The target stimulus contained a randomly oriented sinusoidal grating (2 cycles per degree) embedded in the visual noise. After stimulus presentation, subjects provided a forced-choice judgment of whether the left or the right stimulus contained a grating. The grating location was determined randomly on each trial, and gratings appeared equally often on the left and right. Following stimulus classification, subjects rated their confidence in the accuracy of their response on a scale of 1 through 4. Subjects were encouraged to use the entire confidence scale. If the confidence rating was not registered within 5 s of stimulus offset, the trial proceeded as if a confidence rating had been entered. Such trials were omitted from all analyses. There

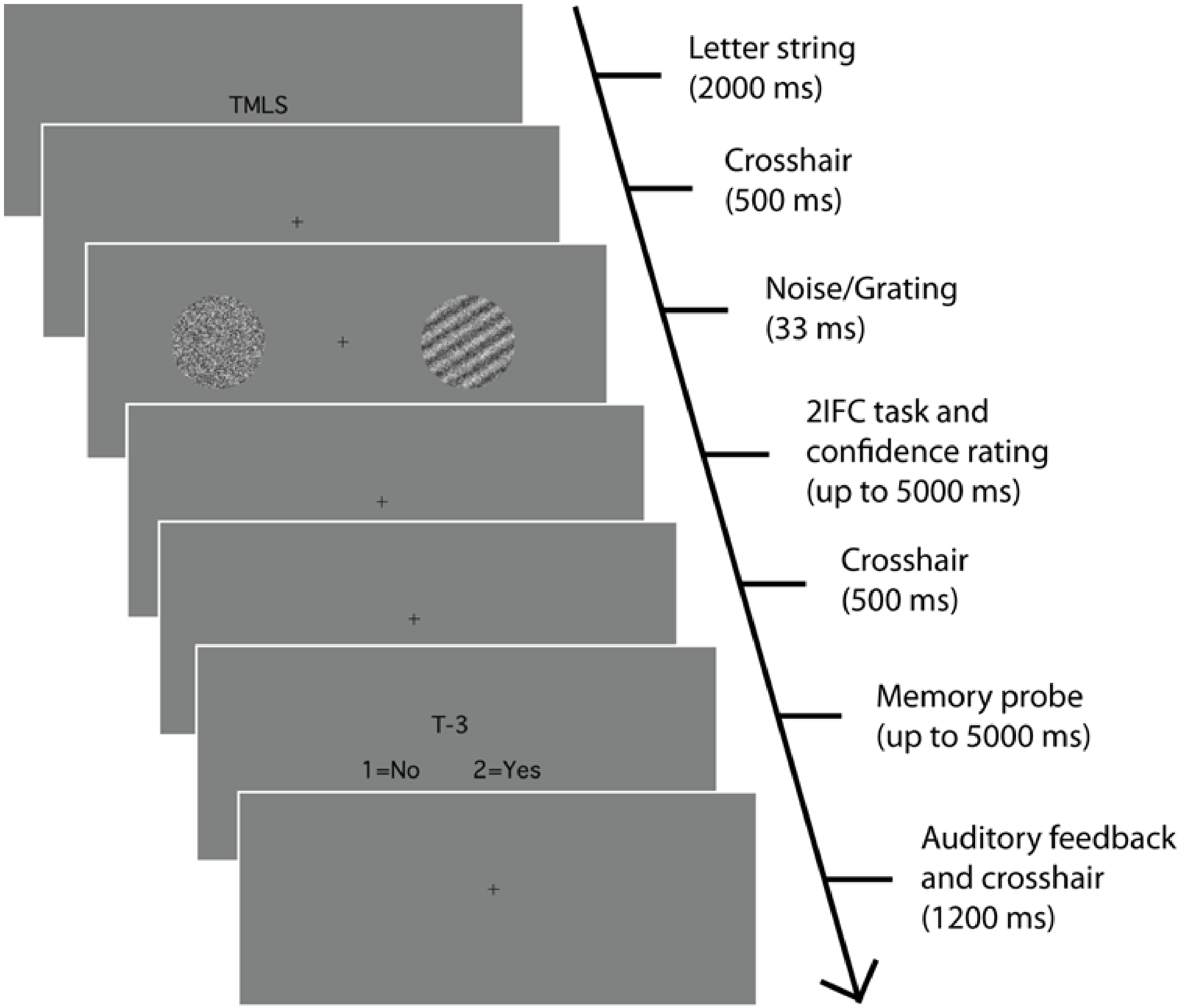
TMLS
+
+
+
+
T-3
1=No 2=Yes
+
Letter string (2000 ms)
Crosshair (500 ms)
Noise/Grating (33 ms)
2IFC task and confidence rating (up to 5000 ms)
Crosshair (500 ms)
Memory probe (up to 5000 ms)
Auditory feedback and crosshair (1200 ms)

Figure 2-1. Experimental design. Subjects performed a working memory (WM) task concurrently with a perceptual decision making task. At the start of the trial, a letter string was presented. Subjects were informed to hold the string in memory and sort it into alphabetical order. Strings could be either one letter long (low WM load) or four letters long (high WM load). Due to randomization of the four letter strings, these could be either easy to alphabetize (high load/easy alphabetization or “high/easy”) or difficult (high load/hard alphabetization or “high/hard”). Subsequently, subjects performed a 2-interval forced choice discrimination task. Two noisy stimuli appears to the left and right of fixation, and one of these contained a sinusoidal grating. Subjects indicated which side the grating appears on and rated decision confidence on a scale of 1 – 4. Finally, subjects performed the WM task. A memory probe consisting of a letter-number pair inquired as to whether the probe-letter was located at probe-number position of the alphabetized string. Experiments 1 and 2 used this same basic design, with slight modifications between them (see Methods).

was a 500 ms interval between the entry of confidence rating and the presentation of the memory probe.

The memory probe consisted of a letter and a number, e.g. T-3. Subjects judged whether it was true that the letter of the memorized and alphabetized string picked out by the probe number matched the probe letter. For instance, suppose that the initial letter string was TMLS, and the memory probe was T-3. The T-3 probe would pose the question, "is it true that the 3rd letter in the alphabetized letter string is a T?" Subjects indicated either "yes" or "no" in response to the probe. In this example, the correct answer is "no," since the alphabetized string is LMST, and the third letter of this string is S, not T. Probe letters were always selected randomly from one of the letters contained in the original letter string. As a consequence of this policy, for one-letter strings, the correct answer was always "yes." For four-letter strings, the probe letter was chosen randomly. For half of all trials, the probe number corresponded to the true index of the probe letter in the alphabetized string. For the remaining half of all trials, the probe number was chosen randomly from one of the three remaining indeces. Thus, for four-letter strings, the correct answer was "yes" for half of all trials. Grating location, letter string size, and correct answer for four-letter strings ("yes" or "no") were counterbalanced.

If no memory response was entered within 5 s of probe onset, the trial proceeded as if a response had been entered. Such trials were omitted from all analyses. After entry of the memory response, a crosshair was presented centrally for 1200 ms, after which time the next letter string was presented. At the beginning of this interval, a 200 ms tone indicated accuracy for the working memory task—a brief high-pitched tone indicated a correct memory response, and a brief low-pitched tone indicated an incorrect response.

At the start of each experimental session, subjects completed 2 practice blocks (20 trials each) and 1 calibration block (120 trials). In the calibration block, performance on the 2-interval forced choice grating localization task was adjusted continuously between trials on the basis of the subject's task performance using the QUEST threshold estimation procedure (Watson & Pelli, 1983). In order to shorten trial length, no letter strings or memory probes were presented during the calibration block; each trial consisted only of presentation of the visual stimuli, followed by the subject's key presses indicating the grating location and decision confidence. Target stimuli were defined as the sum of a grating with Michelson contrast $C_{grating}$ and a patch of visual noise with Michelson contrast C_{noise}. The total contrast of the target stimulus, $C_{target} = C_{grating} + C_{noise}$, was set to 0.9. The non-target stimulus containing only noise was also set to a Michelson contrast of 0.9. The QUEST procedure was used to estimate the ratio of the grating contrast to the noise contrast, $R_{g/n} = C_{grating} / C_{noise}$, which yielded 72% correct performance in the 2IFC task. Three independent threshold estimates of $R_{g/n}$ were acquired, with 40 randomly ordered trials contributing to each, and the median estimate of these was used to create stimuli for the main experiment.

In the main experiment, subjects completed 8 blocks of 50 trials each, for a total of 400 trials. After each block, subjects were provided with a self-terminated rest period lasting up to one minute.

Experiment 2

Participants

Thirty Columbia University students participated in the experiment. Participants gave informed consent and were paid $10 for approximately one hour of participation. The research was approved by the Columbia University's Committee for the Protection of Human Subjects.

One participant was omitted from data analysis, due to producing outlying data in the perceptual metacognitive task under high working memory load (Figure 2-4).

Experimental procedure

The experimental procedure was identical to that of Experiment 1, with three exceptions.

First, in the working memory task, the probe letter was now allowed to differ from the original letter string for one-letter strings. For one-letter strings, the probe letter matched the original letter for half of all trials, and thus the correct answer for the memory task was “yes” on half of all trials. However, as in Experiment 1, the probe letter for four-letter strings was always randomly selected from one of the letters contained in the initially presented string.

Second, in the visual discrimination task, two levels of grating contrast were used. The higher level of contrast was determined in the calibration block, as in Experiment 1. The lower level of grating contrast was set equal to half the value of the higher grating contrast. As with the high grating contrast stimuli, the low grating contrast stimuli were defined by adding the low-contrast grating to a white noise pattern, such that the contrast of the grating+noise stimulus as a whole was set to 0.9. Contrast level was counterbalanced with grating location, letter string size, and correct answer for the memory task (“yes” or “no”).

Third, the presentation of one- and four-letter strings was now blocked, rather than randomly interleaved across trials. For 14 subjects, the first 4 blocks (200 trials) of the main experiment contained only one-letter strings, and the last 4 blocks (200 trials) contained only four-letter strings. For the remaining 16 subjects, the order was reversed. Assignment of subjects to the low-load-first and high-load-first conditions was randomized.

Data analysis for the perceptual task

We measured perceptual and metacognitive performance in the visual task using signal detection theory (SDT) analysis (Green & Swets, 1966; Macmillan & Creelman, 2005; Appendix A). We defined hit rate (HR) as the probability that the subject reported that the grating was on the right, given that the grating was on the right, and false alarm rate (FAR) as the probability that the subject reported that the grating was on the right, given that the grating was on the left. We calculated $d' = z(HR) - z(FAR)$ and used *d'* to quantify sensitivity in the visual discrimination task.

We similarly quantified metacognitive sensitivity, i.e. the efficacy with which confidence ratings discriminate between a subject's own correct and incorrect responses, with meta-*d′* (Maniscalco & Lau, 2012; Appendix A). Specifically, for each WM condition of each subject's data, we found the value of meta-*d′* that jointly maximized the likelihood of the response-specific type 2 ROC curves, where response-specific type 2 ROC curves are derived from "type 2" probabilities of the general form P(confidence = c | stimulus = s and response = r). Maximization of likelihood was achieved using the Optimization Toolbox in MATLAB (MathWorks, Natick, MA). Essentially, estimating meta-*d′* in this analysis amounts to fitting the SDT model to the type 2 probabilities of every subject/condition for every possible permutation of stimulus, response, and confidence level. Please see Appendix A for a more in-depth treatment of the methodology for estimating meta-*d′*.

According to SDT, perceptual sensitivity and metacognitive sensitivity are directly correlated; as an observer becomes better at performing a perceptual tasks, it theoretically follows that metacognitive sensitivity also improves (Galvin et al, 2003; Maniscalco & Lau, 2012). Meta-*d′* is defined such that, if an observer with perceptual sensitivity *d′* exhibits metacognitive performance exactly in line with the SDT prediction, then meta-*d′* = *d′*. However, if the observer underperforms SDT expectation, then meta-*d′* < *d′*.

As suggested in Maniscalco & Lau (2012), these observations suggest a useful conceptual distinction between *absolute* and *relative* metacognitive sensitivity. Absolute metacognitive sensitivity concerns how well confidence ratings discriminate correct from incorrect responses overall. Relative metacognitive sensitivity concerns how well confidence ratings discriminate correct from incorrect responses, *relative to* how informative we might expect those confidence ratings to be in light of the observer's perceptual performance. Whereas absolute metacognitive sensitivity can be measured straightforwardly with meta-*d′*, relative metacognitive sensitivity can be measured by means of a numerical comparison between meta-*d′* and *d′*. Relative metacognitive sensitivity is a useful construct in that it allows us to take the theoretical relationship between perceptual and metacognitive performance into account when evaluating metacognitive performance, which in turn facilitates discovery of "genuinely" metacognitive effects, as opposed to differences in absolute metacognitive sensitivity that can potentially be attributed to differences in the underlying perceptual task performance.

In Experiment 2, the low contrast condition led to unexpectedly low levels of performance in the perceptual task. Average *d'* for the low contrast stimuli was .25, which for an unbiased observer corresponds to a rate of 55% correct responding. One sample t-tests revealed that both *d'* and meta-*d'* were significantly greater than zero (i.e., the chance level of responding) in the low contrast condition (*p*s < .05). However, at these near-chance levels of performance, data are noisy and subject to floor effects. For these reasons, we will focus on analyzing only the high contrast condition of Experiment 2. Nonetheless, the data from the low contrast condition are useful to the extent that they demonstrate that meta-*d'* scales directly with *d'* and can be significantly better than chance even when *d'* is itself close to chance performance. In turn, this suggests that the function relating *d'* and meta-*d'* has a y-intercept approximately equal to zero. On the assumption that the function relating *d'* and meta-*d'* is linear (with a zero y-intercept), the slope of this line would then indicate the quality of metacognitive performance relative to perceptual performance. But the slope of a line with zero y-intercept is just the ratio of y to x, as e.g. for two points on a line, $y_1 = mx_1$ and $y_2 = mx_2$, it follows that $m = y_1 / x_1 = y_2 / x_2$. Therefore, computing the ratio meta-*d'* / *d'* is akin to measuring the slope of the line relating meta-*d'* and *d',* and thus provides a means of measuring the quality of metacognitive performance, relative to perceptual performance. We therefore compute the ratio M = meta-*d'* / *d'* to measure relative metacognitive sensitivity.

Results

Due to the similarities in experimental design and empirical outcomes in Experiments 1 and 2, we will present the results from these experiments concurrently. The primary analysis of interest is to assess performance in the perceptual task as a function of difficulty of the working memory (WM) task. We therefore distinguish between low WM load (one-letter memory string) and high WM load (four-letter memory string) conditions.

We further categorize the four-letter strings by the degree to which these randomly created strings were initially presented in alphabetical order. Since subjects were required not only to hold the strings in memory but also to alphabetize them, the degree to which the strings were initially well-alphabetized or scrambled could further modulate the resources or processes required to perform the memory task. We classified string alphabetization by counting how many of the three consecutive letter pairs in each four-letter string were in alphabetical order. For instance, in the string ADBC, two consecutive letter pairs are in alphabetical order (AD and BC) but one is not (DB). Strings with two or three letter pairs in alphabetical order were considered to be well alphabetized, and strings with zero or one letter pair in alphabetical order were considered to be poorly alphabetized. We refer to well alphabetized four-letter strings as “high WM load / easy alphabetization” or “high/easy” for short, and poorly alphabetized strings as “high WM load / hard alphabetization” or “high/hard” for short.

Working memory performance

Overall, the WM load manipulation was successful in presenting a challenging working memory task (Figure 2-2), as revealed by separate 2 (WM load: high, low) x 2 (Experiment: 1, 2) mixed-measures ANOVAs on accuracy and reaction time in the WM task. Compared to the low load condition, high WM load decreased accuracy (main effect of WM load, $p < .001$; Experiment 1: low load mean = 97.4% correct, high load mean = 77.8% correct; Experiment 2: low load mean = 88.4% correct, high load mean = 77.7% correct) and increased median reaction time (WM load, $p < .001$; Expt 1: low load mean = 609

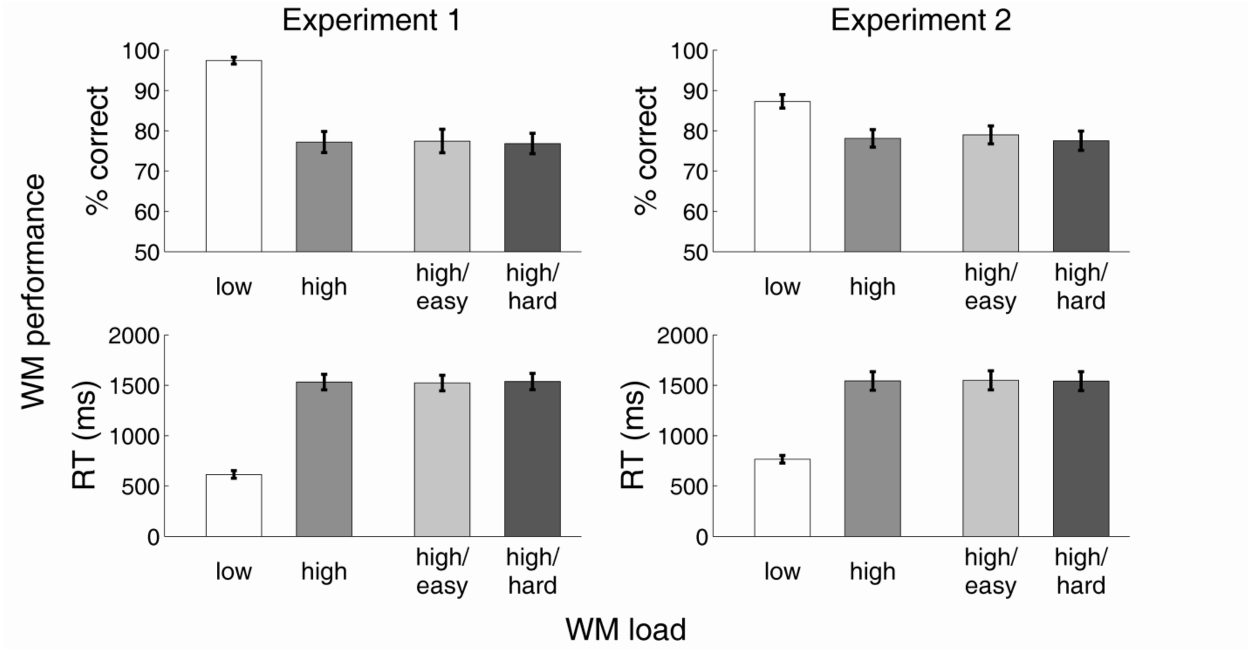

Figure 2-2. Working memory performance. As expected, the memory task was significantly more difficult under high WM load than under low load, exhibiting significantly lower rates of correct responding and longer reactions times. However, within the high WM load condition, the distinction between easy and difficult alphabetization did not manifest as an observable change in WM task performance. Error bars represent 1 SEM.

ms, high load mean = 1537 ms; Expt 2: low load mean = 761 ms, high load mean = 1519 ms). However, alphabetization difficulty did not affect accuracy (*p* = .4) or median reaction time (*p* > .9) in the memory task.

The main effect of WM load on task performance was modulated by a significant WM load x Experiment interaction (*p* = .008). The source of this interaction was that memory performance under low load was significantly better in Experiment 1 than in Experiment 2 (independent samples t-test on % correct, *p* < .001). (Median reaction time on the memory task was also faster in Experiment 1, although the WM load x Experiment interaction for median RT did not achieve significance, *p* = .19.) This difference was due to the fact that the low load task was trivial in Experiment 1, as the probe letter was always the same as the one-letter string, whereas in Experiment 2, the probe only matched the one-letter string on half of all trials and thus posed a simple but non-trivial memory demand (see Methods). However, the structure of the memory task under high load was identical for the two experiments, and here memory performance did not differ for either accuracy or reaction time (*p*s > .8).

Perceptual task performance as a function of WM load

We plot *d′* and meta-*d′* as a function of WM load and alphabetization difficulty in Figure 2-3. We analyzed this data with separate 2 (WM load: high, low) x 2 (Experiment: 1, 2) mixed design ANOVAs for *d′* and meta-*d′*. In both experiments, WM load impaired perceptual sensitivity (*d′*; WM load, *p* < .001; WM load x Experiment, *p* > .9; Expt 1: low load mean = 1.88, high load mean = 1.63; Expt 2: low load mean = 1.96, high load mean = 1.72) and metacognitive sensitivity (meta-*d′*; WM load, *p* = .004; WM load x Experiment, *p* = .6; Expt 1: low load mean = 1.26, high load mean = .95; Expt 2: low load mean = 1.25, high load mean = 1.03).

However, the reduction in meta-*d′* due to WM load is qualified by the fact that WM load also reduced *d′*. Since *d′* and meta-*d′* correlate (Galvin et al., 2003; Maniscalco & Lau, 2012), the reduction in meta-*d′* under high WM load might be attributable merely to the reduction in *d′*, rather than to a direct, independent effect on metacognitive performance per se. If WM load impaired metacognitive performance over and above its impairment of perceptual performance, we might expect that WM load would decrease the ratio meta-*d′* / *d′*, which we shall hereafter refer to as M. Although M was numerically lower under high load (Expt 1: low load mean = .74, high load mean = .61; Expt 2: low load mean = .68, high load mean = .63), these differences were not statistically significant in the WM load x Experiment ANOVA (WM load, *p* = .16; WM load x Experiment, *p* > .5). Thus, overall WM load did not appear to reduce relative metacognitive sensitivity, as measured by M.

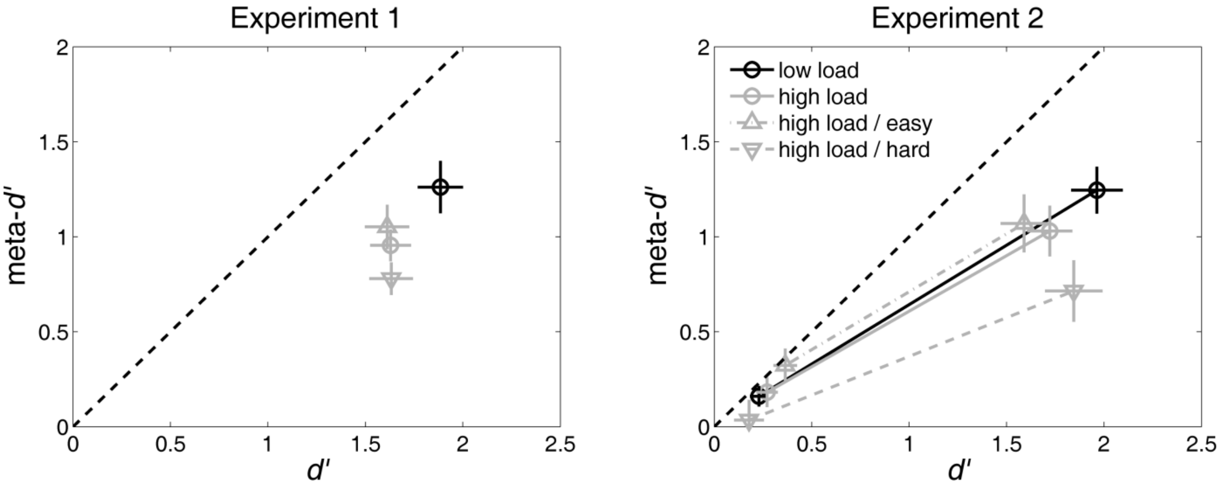

Figure 2-3. Perceptual and metacognitive performance as a function of WM load. We measured perceptual sensitivity with the signal detection theory (SDT) measure *d′* (Green & Swets, 1966) and metacognitive sensitivity with the SDT measure meta-*d′* (Maniscalco & Lau, 2012). If subjects perform according to SDT expectations, data should fall along the dashed line of unity, meta-*d′* = *d′*. Here, subjects' metacognitive sensitivity underperformed SDT expectation. Overall, under high WM load, *d′* and meta-*d′* were equally impaired. Crucially, well-scrambled WM strings were associated with an impaired *ratio* of meta-*d′* to *d′*, suggesting that the process of manipulating the contents of working memory had a selective deficit on relative metacognitive sensitivity. On these plots, this result manifests as the data for the "high load / hard alphabetization" condition occupying a lower region on the y-axis of the meta-*d′* vs *d′* plot than the other data points in spite of having a similar x-axis value. Error bars represent 1 SEM.

Perceptual task performance as a function of WM load and alphabetization difficulty

In order to take into account the effect of alphabetization difficulty, we calculated M separately for the high load / easy alphabetization and high load / hard alphabetization conditions. Scatterplots relating M for these conditions as well as M under low load are displayed in Figure 2-4. One subject in each of Experiments 1 and 2 produced outlying data on these plots, and were therefore excluded from all analyses. Inspection of the remaining data suggests that M was lower under high load / hard alphabetization than in the other conditions.

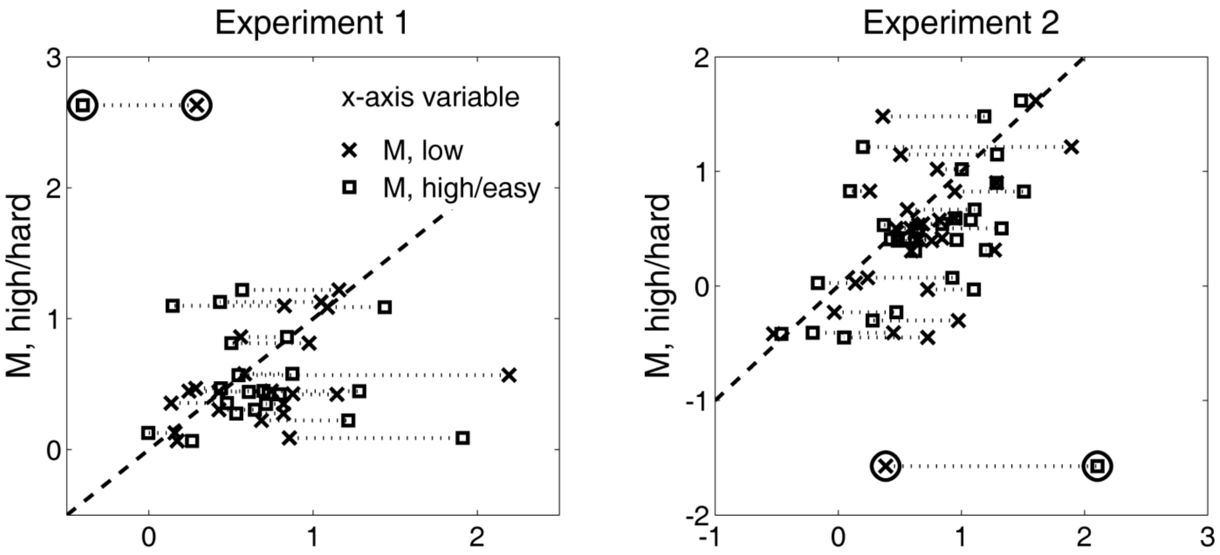

Figure 2-4. Scatterplots of M under the different WM conditions. M, as the ratio of meta-*d′* to *d′*, measures how well the subject performed metacognitively (meta-*d′*) in relation to perceptual performance (*d′*). For subjects behaving according to signal detection theory expectation, M = 1, whereas M < 1 indicates metacognitive performance that is suboptimal relative to SDT expectation. In the scatterplots, dashed horizontal lines connect the two data points generated by single subjects. Most points fall below the line of unity, suggesting that M is impaired in the "high/hard" condition compared to the "low" and "high/easy" conditions. Data shown in circles were considered to be outliers, and data from these two subjects was omitted from all analysis.

To investigate this possibility, we conducted a 2 (WM demand: low, high/hard) x 2 (Experiment: 1, 2) mixed-design ANOVA on M, where we use the factor name "WM demand" rather than "WM load" to highlight the fact that this factor now subdivides the high load condition according to alphabetization difficulty. Indeed, we found that, compared to low WM load, high load / hard alphabetization impaired M in both experiments (WM demand, p = .003; WM demand x Experiment, p = .9; Expt 1: low load mean = .74, high/hard mean = .54; Expt 2: low load mean = .68, high/hard mean = .46). By stark contrast, high load / easy alphabetization strings did not impair M relative to low WM load (WM demand, p > .9; WM demand x Experiment, p > .7; Expt 1: low load mean = .74, high/easy mean = .71; Expt 2: low load mean

= .68, high/easy mean = .71). M under high load / hard alphabetization was also significantly lower than under high load / easy alphabetization (WM demand, *p* = .006; WM demand x Experiment, *p* > .6). Thus, relative metacognitive sensitivity in the perceptual task was not affected by the overall memorization load placed upon WM, but rather was selectively impaired by the need to perform extensive alphabetization on high load WM strings. These findings are portrayed in Figure 2-5.

We pursued these findings further by investigating the separate effects of alphabetization difficulty under high WM load on *d′* and meta-*d′*. A 2 (WM demand: high/easy, high/hard) x 2 (Experiment: 1, 2) mixed-design ANOVA on *d′* did not reveal any significant effects (WM demand, *p* = .12; WM demand x Experiment, *p* = .19; Expt 1: high/easy mean = 1.61, high/hard mean = 1.63; Expt 2: high/easy mean = 1.59, high/hard mean = 1.84). However, a similar ANOVA for meta-*d′* did reveal an effect of alphabetization difficulty for both experiments (WM demand, *p* = .004; WM demand x Experiment, *p* > .6; Expt 1: high/easy mean = 1.05, high/hard mean = 0.78; Expt 2: high/easy mean = 1.07, high/hard mean = 0.71). Thus, whereas overall WM load impaired both *d′* and meta-*d′*, the added component of alphabetization difficulty within the high load condition did not modulate *d′*, but did impose a selective deficit for meta-*d′*.

Confidence as a function of accuracy and WM demand

Metacognitive sensitivity is determined by how an observer places confidence ratings for correct and incorrect responses. There are several ways in which high WM demand may have impaired metacognitive performance—e.g. by reducing confidence for correct responses, increasing confidence for incorrect responses, or both. To investigate, we performed a 2 (Accuracy: correct, incorrect) x 2 (WM demand: low, high/hard) x 2 (Experiment: 1, 2) mixed-design ANOVA on confidence in the perceptual task. In addition, a significant main effect of Accuracy on confidence (*p* < .001), reflecting higher confidence for correct responses, there was also a significant Accuracy x WM demand interaction (*p* =

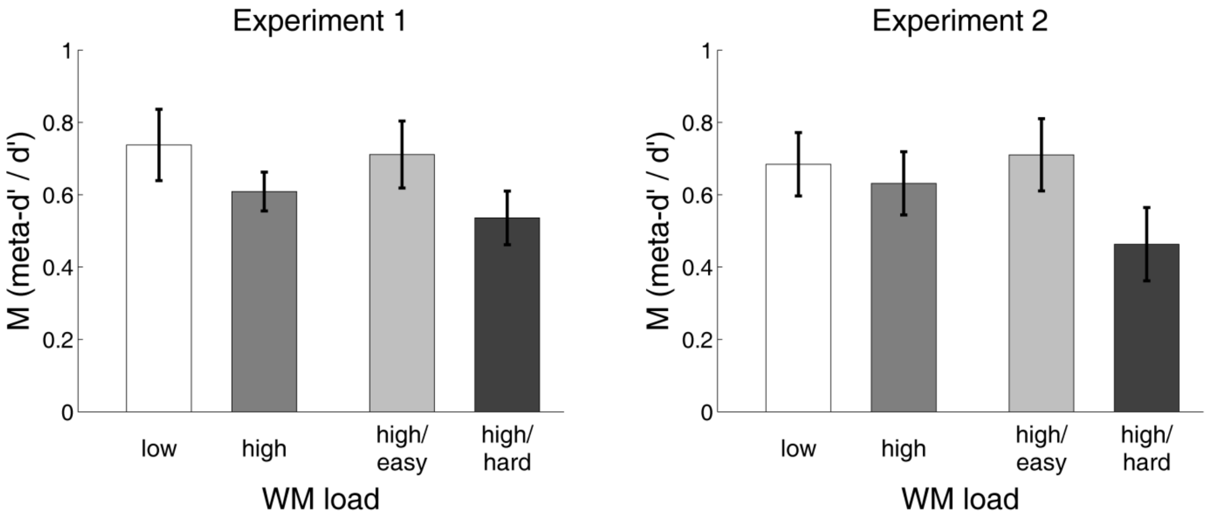

Figure 2-5. Average values of M across WM load conditions. In both experiments, although M was numerically lower under high overall WM load than under low load, the difference was not significant. However, M under high WM load and difficult alphabetization was significantly lower than it was under low load. Error bars represent 1 SEM.

.006) which was not modulated by Experiment (Accuracy x WM demand x Experiment, *p* > .6). The Accuracy x WM demand interaction reflects the fact that under high/hard WM demand, confidence for correct responses decreased whereas confidence for incorrect responses increased (Figure 2-6). This pattern can also be seen in the pooled type 2 ROC curves described in more detail in the following section (Figure 2-7), as under high/hard WM demand, type 2 false alarm rates increased whereas type 2 hit rates decreased relative to the low WM load condition.

Pooled type 2 ROC curve analysis

One potential concern with the foregoing analyses is that trial counts were somewhat low, a concession necessary in the task design due to the relatively long duration of each trial. In Experiment 1,

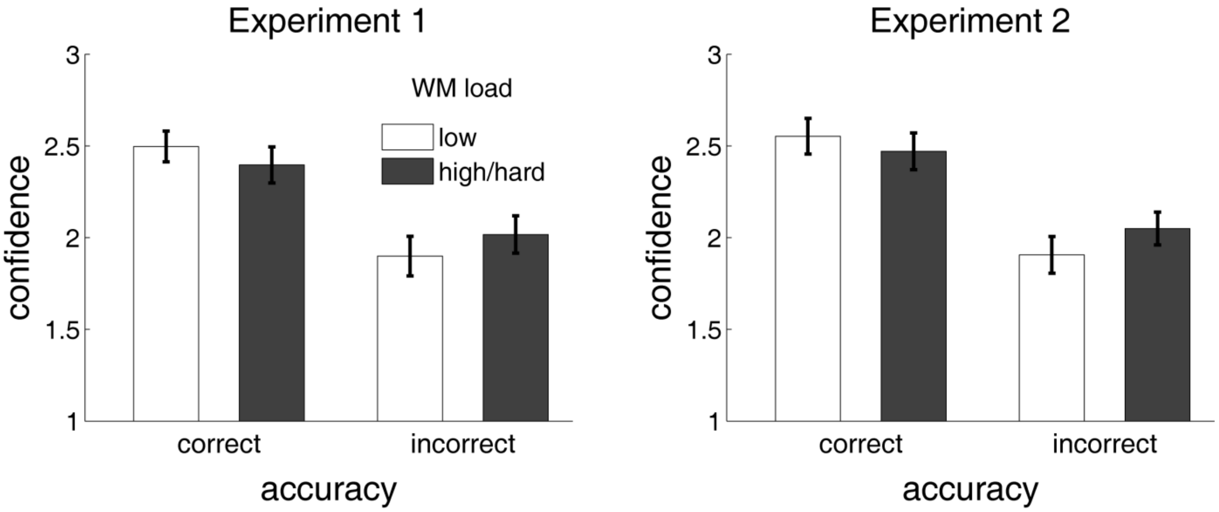

Figure 2-6. Mean levels of confidence as a function of accuracy in the perceptual task and WM load. Overall levels of confidence did not differ for low and high/hard WM load. A qualitatively similar pattern arose in both experiments, whereby under high/hard WM load, confidence for correct decisions decreased and confidence for incorrect decisions increased. Error bars represent 1 SEM.

200 trials contributed to the low load condition, and roughly 100 trials contributed to each of the high/easy and high/hard conditions. In Experiment 2, for each level of grating contrast, 100 trials contributed to the low load condition, and roughly 50 trials contributed to each of the high/easy and high/hard conditions. In order to lend further support to the findings described above, we therefore performed a complementary analysis that pooled data across subjects.

The ideal approach to performing an SDT analysis is to calculate metrics such as *d'* separately for each subject, using their individual hit rate and false alarm rate data. But in cases where within-subject trial counts are a concern but there is ample between-subject data, an alternative approach is to average hit rates and false alarm rates across subjects, and use this *pooled* data to perform SDT analysis on the group as a whole (Macmillan & Kaplan, 1985; Macmillan & Creelman, 2005). This pooling approach is a legitimate way to analyze the data; for instance, it was used extensively in a classic article demonstrating SDT's ability to characterize a wide variety of empirical ROC curves (Swets, 1986a). Although the pooling approach potentially underestimates sensitivity if subjects have very different values for sensitivity or response bias (Macmillan & Kaplan, 1985), such concerns are mitigated for the present purposes, as we are primarily concerned in analyzing the *difference* in metacognitive sensitivity between two conditions, rather than the overall level of metacognitive sensitivity in a single condition.

For the present purposes, we wish to compare metacognitive performance in the low WM load and the high/hard WM load conditions. Thus, we pooled data across subjects to construct pooled type 2 relative operating characteristic (ROC) curves. The type 2 ROC curve is a plot of type 2 hit rate (i.e. probability of high confidence for correct responses) against type 2 false alarm rate (i.e. probability of high confidence for incorrect responses) (Galvin et al., 2003). The "type 2" designation indicates the task of classifying response accuracy with confidence ratings, in contradistinction to the "type 1" task of performing an objective classification of the stimuli. Because subjects rated confidence on a scale of 1 through 4, three (type 2 FAR, type 2 HR) pairs could be calculated for each subject by separately considering "high confidence" to consist in all confidence ratings greater than 1, all ratings greater than 2, or all ratings greater than 3 (Macmillan & Creelman, 2005). We computed the (type 2 FAR, type 2 HR) pairs for each subject in the low WM load and high/hard WM load conditions and averaged these across subjects. The resulting ROC curves are displayed in Figure 2-7.We similarly computed the across-subject average (FAR, HR) for the visual discrimination task, and computed a group *d'* from this pooled data. We used this value of pooled *d'* to construct the ideal pooled type 2 ROC curve, assuming unbiased responding in the visual discrimination task (Maniscalco & Lau, 2012).

Visual inspection of the pooled type 2 ROC curves confirms that metacognitive performance was worse under high/hard WM load than under low WM load, as under this condition the type 2 ROC curve lies closer to the line of chance metacognitive performance, i.e. the line where type 2 FAR = type 2 HR. In order to quantify this observation, we performed a bootstrap analysis (Mooney & Duval, 1993). In the

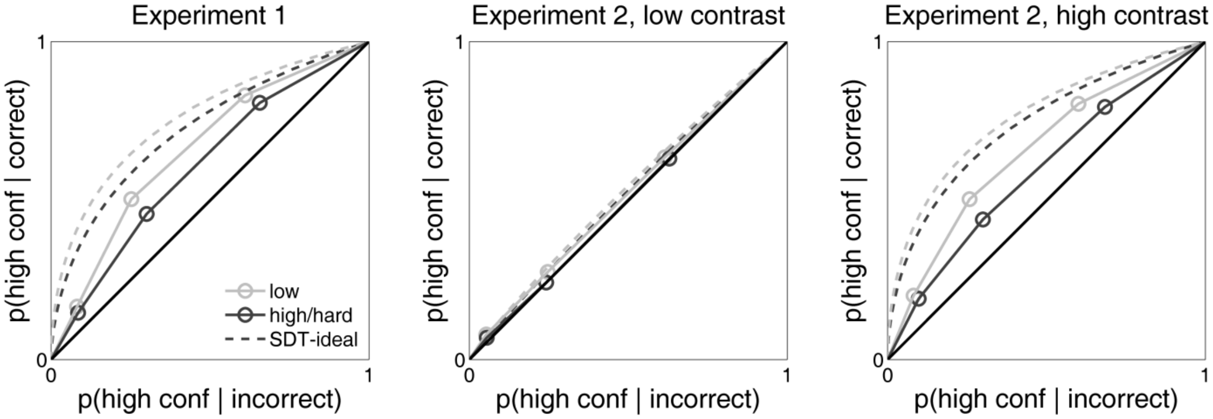

Figure 2-7. Pooled type 2 ROC curves. In the analyses depicted in the previous figures, *d′* and meta-*d′* were computed separately for each subject. We supplemented this analysis by pooling together (averaging) type 2 hit rates (p(high conf | correct)) and type 2 false alarm rates (p(high conf | incorrect)) across subjects and using the averaged data to construct the pooled type 2 ROC curves displayed here. Similar features are evident in the pooled analysis—the empirical type 2 ROC curves are closer to the diagonal line of chance metacognitive performance than are the SDT-ideal dashed curves (echoing the finding that M < 1), and the type 2 ROC curve is closer to chance in the high/hard WM load condition than it is under the low WM load condition (echoing the finding that M is lower for “high/hard” is lower than M for “low”). A bootstrap analysis provided quantitative statistical support for these qualitative observations (see Results).

bootstrap procedure, the sampling distribution for a variable is estimated by repeatedly resampling with replacement from the original data set and computing the value of the variable for each such bootstrap sample. We constructed 1000 bootstrap samples of the type 1 and type 2 hit rate and false alarm rate data for each WM load condition. For each bootstrap sample, we calculated *d′* and estimated meta-*d′* by finding the least-squares fit of the meta-*d′* model to the type 2 ROC curve (Maniscalco & Lau, 2012). We then analyzed the distribution of values for $M_D = M_{low} – M_{high/hard}$. For Experiment 1, the mean M_D was .21 and only 3.9% of all bootstrap samples had $M_D < 0$. For Experiment 2, the mean M_D was .28 and only 2% of all bootstrap samples had $M_D < 0$. Thus, this complementary bootstrap analysis of the pooled type

2 ROC data provides converging evidence for the claim that metacognitive performance was impaired under the high WM load / hard alphabetization condition.

Analysis of sequential dependencies in confidence rating

One possible way in which high/hard WM load might have impaired metacognitive sensitivity is by adding noise to type 2 criterion setting (Mueller & Weidemann, 2008; Benjamin, Diaz, & Wee, 2009). According to the SDT model, confidence ratings are created by comparing the magnitude of evidence for a perceptual decision to a set of criterion values (Macmillan & Creelman, 2005; Appendix A). (We refer to “type 2 criteria” to distinguish these decision criteria from the “type 1 criterion” that determines the observer’s perceptual decisions about the stimulus.) If an observer uses different values for the type 2 criteria across trials, the net effect is that metacognitive sensitivity is reduced (Mueller & Weidemann, 2008). One tell-tale sign of noise in the criterion setting process is trial-to-trial dependencies in perceptual decisions (Mueller & Weidemann, 2008). If an observer’s responses from trial to trial are correlated in spite of stimulus strength across trials being randomized, this is evidence that criterion setting drifts over the course of the experiment and is therefore not perfectly stationary. However, the converse inference does not hold: failure to find trial-to-trial response dependencies does not indicate the absence of noise in criterion setting, since e.g. if criterion values were corrupted with noise drawn from a random distribution on each trial, such random noise in the criterion setting process would not produce systematic across-trial response dependencies.

We thus tested the hypothesis that high/hard WM load reduces metacognitive sensitivity by inducing sequential dependencies in confidence rating. We limited the analysis to Experiment 2, since this experiment used a block design for WM load which thus facilitated analysis of performance under the various WM load conditions in sequential trials. We limited analysis to sequential trial pairs satisfying the following conditions: both trial *i* and trial *i* – 1 had high contrast grating stimuli; the subject successfully entered a perceptual decision and confidence rating for both trial *i* and trial *i* – 1; and the perceptual decision on both trial *i* and trial *i* – 1 was correct. We enforced these conditions in order to minimize extraneous sources of variance in confidence ratings for each sequential trial pair. (There were an insufficient number of trials to conduct a similar analysis restricted only to incorrect perceptual decisions.) For each subject, we computed (1) the mean of the differences in confidence for each trial *i* and trial *i* – 1; (2) the standard deviation of the differences in confidence for each trial *i* and trial *i* – 1; (3) the Pearson's correlation coefficient for confidence on trial *i* and trial *i* – 1.

Paired t-tests did not reveal a significant difference between low WM load and high/hard WM load in terms of either the mean or the standard deviation of the difference in confidence for sequential trials (*p*s > .4). We compared the sequential trial correlations in confidence for low and high/hard WM load by transforming each subject's Pearson's *r* value into a normal deviate *z* value using Fisher's r-to-z transform (Fisher, 1915). A paired t-test on these *z* values did not reveal a significant difference in the trial-to-trial correlations in confidence for low and high/hard WM load (*p* > .6). Taken together, these results suggest that the decrease in metacognitive sensitivity due to high/hard WM load was not mediated by the kind of noisy type 2 criterion setting that would result in sequential dependencies in confidence rating. However, it remains possible that more random forms of noise in the type 2 criterion setting process might be a candidate mechanism for this effect.

Discussion

In summary, we found that when experimental subjects had to perform a working memory task concurrently with a perceptual decision making task, performance on the two tasks interacted in interesting ways. First, there was an overall effect of WM load whereby both perceptual (*d'*) and metacognitive (meta-*d'*) sensitivity in the perceptual task decreased when longer letter strings had to be maintained in memory. Second, there was a specific effect of the manipulation demand imposed by WM contents on perceptual metacognition. For letter strings that were initially poorly alphabetized, stronger manipulation demand was imposed upon subjects, as they had to perform more mental operations upon WM contents in order to arrive at a properly alphabetized string. This manipulation demand had selective effects upon relative metacognitive sensitivity, as measured by M = meta-*d′* / *d′*. When manipulation demand for four-letter WM strings was low (i.e. the “high/easy” condition), M did not differ for high and low WM load. But when manipulation demand for four-letter strings was high (i.e. the “high/hard” condition), M was significantly lower than in the low load and high/easy load conditions. Thus, relative metacognitive sensitivity was insensitive to overall WM load, but was selectively impaired when extensive manipulation of WM contents was required.

It is important to interpret these results in light of the theoretical distinction between *absolute* and *relative* metacognitive sensitivity introduced in Maniscalco & Lau (2012). Absolute metacognitive sensitivity refers to the overall efficacy with which confidence ratings discriminate between correct and incorrect responses, as measured e.g. by area under the type 2 ROC curve (Galvin et al., 2003; Fleming et al., 2010). Relative metacognitive sensitivity evaluates the empirically observed level of absolute metacognitive sensitivity with respect to the *expected* level of absolute metacognitive sensitivity, given an observer’s performance on the primary stimulus classification task. Such an expectation can be derived by the theoretical machinery of signal detection theory (Galvin et al., 2003), with the important features that (1) task performance should place a theoretical limit on metacognitive performance, and (2) as task performance improves, so should metacognitive performance. These theoretical predictions have been validated in empirical data (Maniscalco & Lau, 2012).

The SDT measure of absolute metacognitive sensitivity, meta-*d′* (Maniscalco & Lau, 2012; Appendix A), was designed with an eye towards providing a straightforward way to measure relative metacognitive sensitivity. Meta-*d′* is defined such that, for an observer whose performance conforms to SDT assumptions, meta-*d′* = *d′*. Thus, relative metacognitive sensitivity can be operationalized as a direct numerical comparison between meta-*d′* and *d′*, e.g. a subtraction or division. In this study, we found evidence that the y-intercept of the function relating meta-*d′* and *d′* is zero (Figure 2-3), thus implicitly supporting the usage of meta-*d′* / *d′* as the appropriate measure of relative metacognitive sensitivity for these data. For instance, if we suppose that the true function relating meta-*d′* and *d′* for a given observer is meta-*d′* = .8 * *d′*, then the ratio meta-*d′* / *d′* would have a constant value of 0.8 regardless of the value of *d′*, whereas the value of the difference meta-*d′* – *d′* would differ depending on the value of *d′*.

In the current data set, although meta-*d′* decreased under high WM load, *d′* also decreased to a similar extent. Given the known theoretical and empirical dependence of meta-*d′* upon *d′* (Galvin et al., 2003; Maniscalco & Lau, 2012), it is therefore possible to attribute the decline in meta-*d′* under high WM load to the co-occurring decline in *d′*, rather than supposing that WM load had a direct effect upon overall metacognitive performance. Indeed, the relative measure of metacognitive sensitivity, meta-*d′* / *d′*, did not significantly differ as a function of WM load. Thus, while high WM load imposed an overall deficit in performance on the perceptual task, perhaps due to reduced attentional allocation to the visual stimuli under high load, we did not find strong evidence that WM load produced a selective deficit upon metacognitive processing in and of itself.

By contrast, we found that relative metacognitive sensitivity in the perceptual task was indeed impaired in the specific case where the contents of WM required a substantial degree of manipulation (alphabetization). This finding is in keeping with prior empirical investigations on the neural bases of visual metacognition and WM performance. Various higher-level regions of the human and monkey prefrontal cortex, including dorsolateral, rostrolateral, and anterior prefrontal cortex, have been linked to metacognitive performance in visual and memory tasks (Henson et al., 2000; Fleck et al., 2006; Fleming et al., 2010; Yokoyama et al., 2010; Tsujimoto et al., 2010; Fleming et al., 2012; McCurdy et al., 2013). Similarly, dorsolateral prefrontal cortex (dlPFC) in particular has been linked to performance in

WM tasks, with a strong line of evidence that dlPFC is involved particularly with the *manipulation* and *selection* of WM contents, rather than just the passive maintenance of items in WM (Petrides & Milner, 1982; Petrides, 1989; Petrides, 1995; Owen et al., 1996; D'Esposito et al., 1999; Petrides, 2000; Rowe et al., 2000; Miller & Cohen, 2001; Rowe & Passingham, 2001; Rowe et al., 2002; Curtis & D'Esposito, 2003; Bor et al., 2003). In the current study, the fact that relative metacognitive sensitivity was impaired not by overall WM load, but rather by the specific requirement to extensively manipulate WM contents, suggests the working of a common, limited neural resource in dlPFC that contributes to both the manipulation of WM contents and the metacognitive evaluation of visual task performance.

We note that a somewhat similar finding to the current study was previously reported in the context of visual search tasks by Han & Kim (2004). In that study, the time required to find a visual target in a cluttered display as a function of display set size ("search slope") was compared for concurrent WM tasks that either did or did not require active manipulation of WM contents (backwards counting for number items, or alphabetization for letter items). Search slope was significantly steeper than in a control condition when subjects had to manipulate WM contents, but not when subjects had to passively maintain a number or letter string in WM. Thus, as in the present study, Han & Kim found that aspects of processing in a visual task could undergo selective impairment due to the requirement to manipulate WM contents. Han & Kim concluded that aspects of executive functioning, as reflected in manipulation of WM contents, may be required to perform visual search. However, it is unclear to what extent this impairment in visual search is related to the impairment on relative metacognitive sensitivity observed in the current study.

How might it be the case that a cognitive/neural mechanism that contributes to manipulation of WM contents also contributes to metacognitive evaluation of visual perception? The explanation cannot be an overly general mechanism, such as supposing that additional attentional resources required for the WM task would leave fewer attentional resources for the visual task. Such general mechanisms would presumably induce global changes in visual task performance, affecting both *d′* and meta-*d′*, rather than being specific to meta-*d′* / *d′*.

One potential mechanism might have to do with strategies for representing and re-representing stimuli. In WM tasks, lateral PFC activation has been associated with encoding strategies for WM contents—when items are presented in a way that facilitates their reorganization into higher-level units or "chunks," lateral PFC becomes more activated (Bor et al., 2003). Some views hold that metacognition similarly involves the construction of higher-order re-representations or meta-representations of cognitive/neural processing occurring at lower levels in the processing hierarchy (Nelson & Narens, 1990; Schooler, 2002; Cleeremans, Timmermans, & Pasquali, 2007; Pasquali et al., 2010). If so, it is possible that the same processes involved in manipulating and re-organizing WM contents might also be involved in manipulating and re-organizing sensory representations for the purposes of metacognitive evaluation. Presumably, occupation of such a resource in the manipulation of WM contents would detract from its active employment in the metacognitive evaluation of visual processing, thus impairing relative metacognitive sensitivity.

Another possible set of common underlying mechanisms concerns response selection and the maintenance and flexible adaptation of decision rules. Response selection, defined by Curtis and D'Esposito (2003) as "the operation by which information in short-term storage becomes the focus of attention such that it can be maintained and eventually used to choose an appropriate motor response" (p. 421), has been tied to dlPFC activity in the context of WM tasks (Rowe et al., 2000; Rowe & Passingham, 2001; Rowe et al., 2002; Curtis & D'Esposito, 2003). More broadly, PFC has been theorized to support varying levels of sophistication and abstraction in the control and organization of behavior as a function of stimuli and environmental context, action contingencies, currently active goals, and so on (Koechlin & Summerfield, 2007; Badre, 2008). By way of comparison, the SDT model posits that perceptual classification of stimuli and confidence ratings are the outcomes of cognitive decision processes that are not the rigid outcome of low-level perceptual processing but rather can be flexibly adjusted according to the prevailing task instructions, stimulus context, and reward contingencies (Tanner & Swets, 1954; Green & Swets, 1966; Macmillan & Creelman, 2005). According to SDT, perceptual and metacognitive decisions are determined by defining a set of decision criteria which determine the rules according to which graded and ambiguous internal perceptual evidence is mapped onto discrete perceptual decisions and motor outputs (Green & Swets, 1966; Macmillan & Creelman, 2005; Appendix A). A common mechanism in PFC underlying the processes of selecting, evaluating, and manipulating WM contents in the WM task and the processes of metacognitive criterion setting in the perceptual task could potentially explain the results of the current study.

Regardless of the specific manner in which manipulation of WM contents influences metacognitive performance, the results of this study demonstrate a dissociation between perceptual and metacognitive sensitivity, suggesting that these depend on separate underlying mechanisms. Indirect evidence for such a position comes from anatomical (Fleming et al., 2010; McCurdy et al., 2013) and fMRI (Henson et al., 2000; Fleck et al., 2006; Yokoyama et al., 2010; Fleming et al., 2012) studies in humans, and single-unit recordings in monkeys (Tsujimoto et al., 2010), that associate metacognitive performance with high-level structures in PFC rather than earlier visual processing regions. Here we provide stronger evidence for a perceptual/metacognitive dissociation than these prior associational studies by demonstrating the existence of an experimental intervention that selectively disrupts metacognitive performance. These results closely echo those discussed in Chapter 3, in which we demonstrate that transcranial magnetic stimulation to bilateral dlPFC can selectively impair metacognitive, but not perceptual, sensitivity.

Chapter 3

Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs relative metacognitive sensitivity in a visual discrimination task

Introduction

In Chapter 1, we showed that a Hiearchical model best accounts for the dissociation between objective task performance and subjective reports of visibility in the metacontrast masking paradigm. Comparison of these modeling results with imaging results on the same paradigm (Lau & Passingham, 2006) is suggestive that the late processing stage posited by the Hierarchical model might correspond to dorsolateral prefrontal cortex (dlPFC). In Chapter 2, we demonstrated that when subjects perform a working memory task concurrently with a perceptual task, metacognitive sensitivity in the perceptual task can be selectively disrupted when extensive manipulation of the contents of working memory is required. These results are similarly suggestive of the recruitment of a common neural resource housed in dlPFC. However, in both cases the link between visual metacognition and dlPFC is indirect.

In the current chapter, we more directly probe the influence of dlPFC upon visual metacognition by assessing the impact of bilateral transcranial magnetic stimulation (TMS) to dlPFC on objective and subjective perceptual performance. We required volunteers to perform a 2-interval forced-choice visual task, identifying the spatial arrangement of two visual stimuli (a square and a diamond, Figure 3-1 A). At the same time, they also rated the subjective visibility of the stimuli ('clear' or 'unclear'). Subjects performed these tasks before and after TMS applied to bilateral dlPFC (Figure 3-1 B). We used theta-burst stimulation (TBS), a recently developed protocol that is known to effectively depress cortical excitability by mimicking the action of long-term potentiation and long-term depression in cortical tissues (Huang, Edwards, Rounis, Bhatia, & Rothwell, 2005). One advantage of this technique is that the effect of 20 seconds of stimulation is known to last for up to 20 minutes, which means we had the opportunity to depress both sides of the dlPFC by stimulating them sequentially. We opted for bilateral stimulation as this has been suggested to be critical: Sahraie et al. (1997) have suggested that one reason visual defects do not seem to frequently follow prefrontal lesions may be that such lesions have to be large and bilateral. Using this sequential method to depress the dlPFC bilaterally, we found that the metacognitive sensitivity of reported visual awareness was reduced after TMS.

Methods

Participants

Twenty healthy volunteers with normal or corrected-to-normal vision and no history of neurological disorders or head injury were recruited from the database of volunteers at the Functional Imaging Laboratory, Institute of Neurology, University College London, UK. Written informed consent was obtained from all participants. The study was approved by the joint ethics committee for the National Hospital for Neurology and Neurosurgery (UCLH NHS Trust) and the Institute of Neurology (UCL).

Experimental procedure

Subjects were asked to perform a 2-interval forced-choice task (Figure 3-1 A). Testing was performed in a darkened room. Stimuli were presented against the white background of a CRT monitor refreshing at 120 Hz. The monitor was placed 40 cm away from the subjects' eyes.

On each trial, a diamond and square were presented on either side of a central crosshair for 33 ms. The stimuli had sides measuring 0.8 degrees of visual angle and were centered 1 degree to the left and right of the central crosshair. 100 ms after stimulus onset, a metacontrast mask was displayed for 50 ms in order to enhance task difficulty. The two possibilities for the sequence of stimuli (square on the

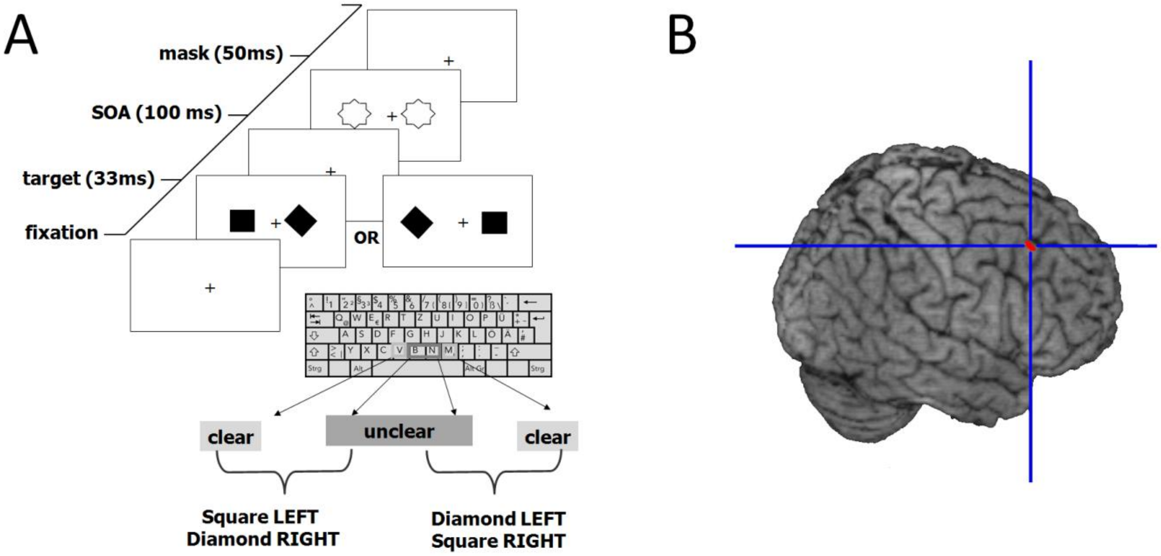

Figure 3-1. Experimental design. (A) Visual task and stimuli. Participants were required to perform a 2-interval forced-choice visual task, identifying the spatial arrangement of two visual stimuli (square on the left and diamond on the right, or the other way round). They rated the subjective visibility ('clear' or 'unclear') at the same time. Thus, in every trial subjects had 4 options as to which key to press in order to respond. **(B) Site of stimulation.** The dorsolateral prefrontal cortex (dlPFC) was the targeted site of stimulation, and was chosen because neural activity from this area has been shown to reflect a difference in the subjective ratings of visibility even when performance in a forced-choice visual task was matched (Lau & Passingham, 2006). The image showing the site of stimulation is based on magnetic resonance brain scans of 6 of the 20 subjects in this study. The scans were collected after completion of the TMS experiments. Right and left dlPFC coordinates were [37 26 50] and [-41 18 52], with standard deviations [4.6 5.6 5.3] and [4.3 5.1 3.8] respectively.

left and diamond on the right, and vice versa) were presented with equal probability in a pseudo-random order.

The subjects' task was to identify which stimulus sequence had just been presented, square left / diamond right or vice versa. At the same time, subjects gave subjective ratings of stimulus visibility ('clear' or 'unclear'). Subjects were instructed to make the visibility judgment in a relative manner, to distinguish between stimuli that were relatively more or less visible. Since stimulus contrast was adjusted so as to yield threshold performance on the stimulus classification task, stimuli used in this experiment were somewhat difficult to see. Nonetheless, subjects were instructed to judge stimulus visibility on each trial relative to the context of stimuli used in this experiment. For instance, a subject might judge that the stimulus on a certain trial was more readily visible than the majority of stimuli seen in the experiment up until that point, even if its visibility was poor by everyday standards. Subjects were encouraged to judge such stimuli as exhibiting "high clarity," i.e. having relatively high clarity compared to other stimuli observed in the experimental context.

Performance level was controlled to be at approximately 75% correct throughout the experiment by titrating the contrast of the stimuli, using a standard up-down transformed-response staircasing procedure (Macmillan & Creelman, 2005). Each trial was randomly designated as belonging to staircase A or staircase B. For staircase A, contrast on the current trial was increased if the subject responded incorrectly on the previous 'A' trial, whereas contrast on the current trial was decreased if the subject responded correctly on the two previous 'A' trials. Staircase B worked in a similar manner, except it required 3 consecutive correct responses on 'B' trials in order to reduce contrast.

Subjects attended two separate testing sessions, both preceded by a demonstration and a practice phase of 100 trials intended to familiarize the subjects with the task and to allow them to reach a stable level of performance. After practice, subjects underwent an initial ('pre') block of 300 trials to measure forced-choice task performance and subjective ratings of visibility. On average this took 10.9 minutes, excluding brief breaks after every 100 trials. After completing this block, two real (or sham) continuous theta-burst stimulation (cTBS) conditioning stimulations, one to the left and one to the right, were delivered to the dorsolateral prefrontal area. The two stimulations were separated by a one minute inter-train interval. Following real (or sham) stimulation, subjects did another ('post') block of 300 trials. On average this took 10.4 minutes, excluding brief breaks after every 100 trials. Session order by type of cTBS (real versus sham) was counterbalanced across subjects.

Theta-burst stimulation

In each TBS session, 600 biphasic stimuli, at a stimulation intensity of 80% of active motor threshold (AMT) for the right first dorsal interosseous (FDI) hand muscle, were given over the left and right DLPFC area using a Magstim Super Rapid stimulator (Whitland, Wales, UK) connected to four booster modules. The conditioning cTBS stimuli were delivered in two separate 20-second trains of 300 cTBS pulses, one for the left and one for the right, separated by an inter-train interval of 1 minute. A similar bilateral procedure has been used in a recent clinical study (Arfeller, Vonthein, Plontke, & Plewnia, 2009).

A standard figure-of-eight-shaped coil (Double 70mm Coil Type P/N 9925; Magstim) was used for both real and sham cTBS. Real cTBS was delivered with the coil placed tangentially to the scalp with the handle pointing posteriorly. In sham cTBS sessions, the coil was placed perpendicularly to the scalp, an ineffective position for the delivery of conditioning pulses, which provided comparable acoustic stimuli to the real cTBS condition. The coil was positioned with the handle at 45° to the sagittal plane. The current flow in the initial rising phase of the biphasic pulse in the biphasic pulse induced a posterior-to-anterior current flow in the underlying cortex.

The basic TBS pattern was a burst containing 3 pulses of 50 Hz magnetic stimulation given in 200 ms intervals (i.e. at 5 Hz). In the continuous theta burst stimulation paradigm (cTBS), a 20 second train of uninterrupted TBS is given (300 pulses or 100 bursts). Physiological studies have shown that this produces a decrease in corticospinal excitability which lasts for about 20 minutes (Huang et al., 2005), when applied to the primary motor cortex, M1. This rTMS paradigm has the advantage of being a rapid and efficient method of conditioning, which has effects on corticospinal excitability that have been shown to involve similar mechanisms to long-term potentiation/depression (LTP/LTD) with NMDA dependence (Huang, Rothwell, Edwards, & Chen, 2008), as well as effects on behavior and learning (Huang et al., 2005; Talleli et al., 2007).

The site of cTBS stimulation was located 5 cm anterior to the 'motor hot spot' on a line parallel to the midsagittal line. This dlPFC location has been used in previous studies and can be shown consistently on structural scans (Mottaghy, Gangitano, Sparing, Krause, & Pascual-Leone, 2002; Rounis et al., 2006; Figure 3-1 B). The position of the motor hot spot was defined functionally as the point of maximum evoked motor response in the slightly contracted right FDI. The active motor threshold was defined as the lowest stimulus intensity that elicited at least five twitches in 10 consecutive stimuli given over the motor hot spot, while the subject was maintaining a voluntary contraction of about 20% of maximum using visual feedback.

The use of such low subthreshold intensity (80% AMT) had the advantage of decreased spread of stimulation away from the targeted site thus keeping the area that was stimulated with the conditioning pulses more focal (Pascual-Leone, Valls-Solé, Wassermann, & Hallett, 1994; Münchau, Bloem, Irlbacher, Trimble, & Rothwell, 2002). Also, a previous study on the prefrontal cortex that applied intensity above motor threshold reported unpleasant vagal reactions in subjects (Grossheinrich et al., 2009). However, even at that higher intensity there was no adverse effects on mood, seizure or epileptiform observed in the recorded EEG. This suggests that our stimulation at this lower intensity should be safe to our subjects.

Data analysis

Metacognitive sensitivity (i.e. the efficacy with which visibility ratings distinguish between correct and incorrect responses) was assessed using two separate methods. The first method followed previous studies (e.g. Kolb & Braun, 1995; Kornell, Son, & Terrace, 2007) in using the correlation between accuracy and subjective rating as a measure of metacognitive sensitivity. We used the correlation coefficient phi, which quantifies the degree of correlation between two binary variables, to calculate the correlation between task accuracy (correct/incorrect) and stimulus visibility (clear/unclear). Phi is equivalent to Pearson's *r* computed for two binary variables, and like *r* it ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation). We calculated phi for the 300 trials pre- and post- real and sham TMS for each subject. We predicted that TMS would hinder metacognitive sensitivity, and thus that there would be a TMS (real, sham) x Time (pre, post) interaction.

We also performed a signal detection theoretic (SDT) analysis to estimate metacognitive sensitivity by estimating meta-*d′* (Maniscalco & Lau, 2012; Appendix A). The need for performing a signal detection theory analysis is that phi can be shown to generate non-regular ROC (Receiver Operating Characteristic) curves, which in turn implies an underlying threshold model of detection (Swets, 1986b). The ROC profile and threshold model of phi are not in good agreement with the standard SDT model (Macmillan & Creelman, 2005), nor with theoretical derivations of type 2 ROC curves from the standard SDT framework (Galvin et al., 2003; Maniscalco & Lau, 2012; Appendix A). The consequence of this is that phi may confound sensitivity and response bias, rather than being a pure measure of sensitivity.

Thus, we also quantified metacognitive sensitivity, i.e. the efficacy with which confidence ratings discriminate between a subject’s own correct and incorrect responses, using meta-*d′* (Maniscalco & Lau, 2012; Appendix A). Specifically, for each TMS x Time condition of each subject’s data, we estimated meta-*d′* as follows. First, we estimated the SDT parameters *c′* (the stimulus classification criterion measured relative to *d′*) and *s* (the ratio of standard deviations of internal evidence for the two stimulus classes) (Macmillan & Creelman, 2005). Holding *c′* and *s* constant, we estimated the value for meta-*d′* that minimized the sum of squared errors (SSE) between observed and modeled type 2 ROC curve for trials in which the stimulus was classified as "square left/diamond right." We then estimated a separate meta-*d′* value in the same way, this time for trials in which the stimulus was classified as "diamond left/square right." Thus, we generated two estimates of meta-*d'*, corresponding to the subject's type 2

ROC curves conditional on each stimulus classification type. These two estimates were combined via a weighted average, where the weight of each meta-*d'* estimate was determined by the number of trials used to estimate it. The mean SSE corresponding to each meta-d' estimate was 9.1 x 10^{-5}, indicating that this approach provided an excellent fit to the observed type 2 ROC data. Minimization of SSE was achieved using the Optimization Toolbox in MATLAB (MathWorks, Natick, MA).

Because we are testing a directional hypothesis in a 2 x 2 factorial design (i.e. metacognitive sensitivity is reduced following real TMS more so than following sham TMS), we report halved *p*-values for the TMS x Time interaction on phi and meta-*d' – d'*.

Results

In the following we present ANOVA analyses with within-subject factors of TMS (real, sham) and Time (pre, post) for several independent variables of interest such as accuracy, response time for correct trials, etc. None of these analyses exhibited a main effect of TMS condition (Fs < 1.7), indicating that the real and sham TMS sessions were comparable on baseline task performance.

Stimulus contrast was adjusted online in order to control classification accuracy; thus, as expected, frequency of correct responses did not vary as a function of time or the TMS x Time interaction (*p*s > .05) (Figure 3-2 A). A more insightful measure of stimulus classification performance is the mean contrast required to keep classification accuracy constant. The stimulus contrast generated by the performance staircasing algorithm reduced over time (*p* < .001), suggesting a perceptual learning effect: over time, subjects required a lower level of contrast in order to maintain the same level of response accuracy. However, the TMS x Time interaction was not significant (*p* > .05), indicating that the TMS treatment had no effect on stimulus classification performance (Figure 3-2 B). Likewise, reaction time for correct trials improved over time (*p* = .016), but was not sensitive to TMS (*p* > .05) (Figure 3-2 C).

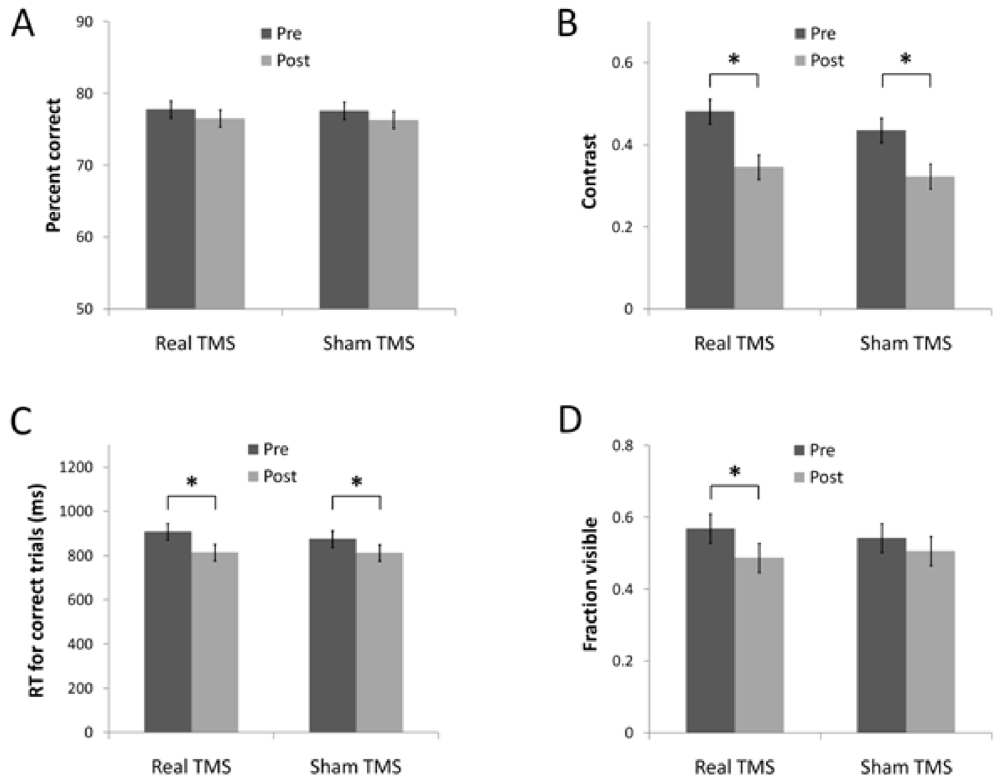

Figure 3-2. **Task performance. (A) Percent correct.** Percent correct was controlled by titration of stimulus contrast, such that stimulus judgments were about 75% correct throughout the experiment (see Methods). Therefore, the lack of any significant effects on these values is trivial. **(B) Mean stimulus contrast.** Stimulus contrast was determined online by the computer program (see Methods), such that if subjects performed better than 75% correct, the contrast was reduced, and if subjects performed worse than 75% correct, the contrast was increased. There was a main effect of time on contrast (p < .001), indicating a perceptual learning effect; had the computer not been programmed to adjust task difficulty online, subjects would have shown improved accuracy over time. However, perceptual learning was not affected by TMS (TMS x Time interaction F = 0.73). **(C) Reaction time for correct responses.** Perceptual learning was also evident in reaction time data. Subjects were quicker to make correct responses in the second half of the experiment (main effect of Time, *p* = .016). However, again, this learning effect was not modulated by TMS (TMS x Time interaction F = 0.79). **(D) Mean visibility ratings.** Visibility ratings decreased over time (*p* = .005), but the TMS x Time interaction on visibility was not significant (*p* = .4). See the discussion for caveats about the visibility rating analysis. 'Real pre': performance level before real TMS. 'Real post': after real TMS. 'Sham pre': before sham TMS. 'Sham post': after sham TMS. * *p* < .05. Error bars represent 1 SEM.

Similarly, mean visibility ratings decreased over time (*p* = .005), but independently of the TMS manipulation (*p* > .05) (Figure 3-2 D). We address this null finding more fully in the discussion.

As hypothesized, TMS significantly impaired metacognitive sensitivity. A TMS x Time interaction was evident for the correlation between accuracy and visibility, phi (*p* = .036) (Figure 3-3 A). Investigation of this interaction revealed that phi was lowered following real TMS (one-tailed paired t-test, *p* < .001) but not sham TMS (*p* > .05).

The bias-free SDT measure of metacognitive sensitivity, meta-*d′* – *d′*, also exhibited a TMS x Time interaction effect (*p* = .015) (Figure 3-3 B). The difference between observed and ideal type 2 sensitivity decreased following real TMS (one-tailed paired t-test, *p* = .006) but not sham (*p* > .05). Metacognitive sensitivity was significantly suboptimal following real TMS, i.e. meta-*d'* < *d'* (one-tailed t-test, *p* = .004) but not in any other TMS x Time condition (*p* > .05).

There are several ways in which TMS could have impaired metacognitive sensitivity. One possibility is that TMS reduced visibility for correct trials, which would amount to a kind of relative blindsight (Lau & Passingham, 2006). Alternatively, TMS may have increased visibility for incorrect trials, a kind of "hallucinatory" effect. A third possibility is that the reduction in metacognitive sensitivity was not specific to correct or incorrect trials. Thus, to better characterize the effect of TMS, we examined visibility ratings separately for correct and incorrect trials pre- and post-TMS (Figure 3-4 A). We found a significant Accuracy x Time interaction (*p* < .001), driven by the fact that TMS reduced visibility for correct responses (two-tailed paired t-test, *p* = .002) but not incorrect responses (*p* > .05). Thus, TMS impaired metacognitive sensitivity by selectively reducing the visibility of correctly classified stimuli.

Discussion

Our results show that theta-burst TMS applied to bilateral dlPFC can reduce metacognitive sensitivity, i.e. the efficacy with which subjective visibility ratings distinguish between correct and

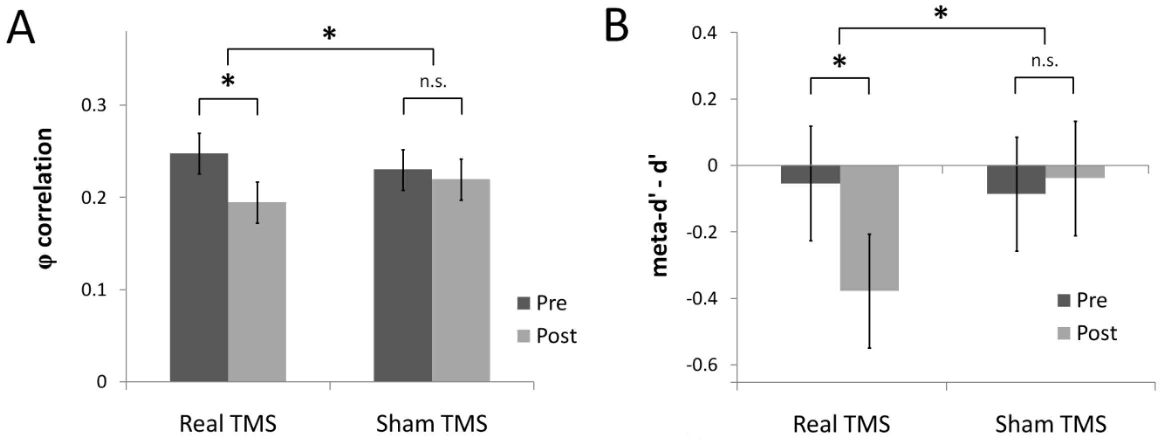

Figure 3-3. Effect of TMS on metacognitive sensitivity. (A) Correlation coefficient, phi. TMS significantly reduced phi, the correlation between stimulus classification accuracy and stimulus visibility. The effect of TMS is evident in a significant TMS x Time interaction, p = .036; phi was lower following real TMS (p < .001) but not sham TMS (p = .5). **(B) Divergence from optimal metacognitive sensitivity, meta-*d’* - *d’*.** A signal detection theory analysis revealed that subjects’ relative metacognitive sensitivity, as measured by meta-*d’* – *d’* (Maniscalco & Lau, 2012; Appendix A), was significantly impaired by TMS (TMS x Time, p = .015). Metacognitive sensitivity was lower following real TMS (p = .006) but not sham TMS (p = .7). Subjects exhibited significantly suboptimal metacognitive sensitivity following real TMS, i.e. meta-*d’* - *d’* < 0 (p = .004) but not in any other experimental condition (ps > .3). 'Real pre': metacognitive performance before real TMS. 'Real post': after real TMS. 'Sham pre': before sham TMS. 'Sham post': after sham TMS. * p < .05, n.s. denotes not significant. Error bars represent 1 SEM.

incorrect stimulus judgments. This effect was driven specifically by a reduction in visibility for correct trials, rather than by a specific elevation of visibility for incorrect trials or by a non-specific effect. In this sense, the direction of the effect is reminiscent of blindsight (Weiskrantz, 1997), where patients deny visual awareness even when they can perform visual discrimination tasks well above chance level. The effect of TMS was specific to metacognitive sensitivity; TMS did not disrupt stimulus classification

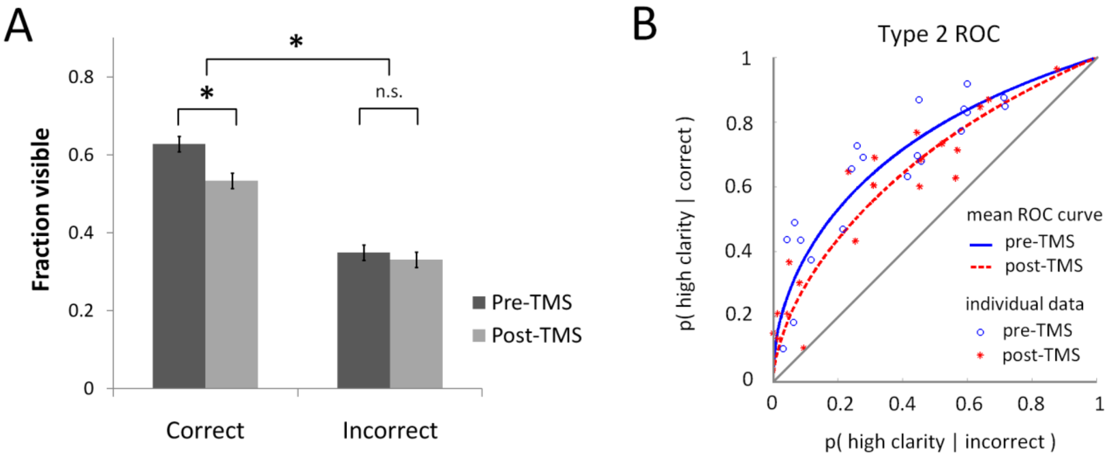

Figure 3-4. Nature of the TMS effect on metacognition. (A) Selective reduction of type 2 hit rate. Visibility ratings are displayed as a function of Time (pre/post TMS) and Accuracy (correct/incorrect) for the real TMS condition. TMS significantly reduced visibility for correct responses (two-tailed paired t-test, *p* = .002), but not for incorrect responses (*p* = .5). The Time x Accuracy interaction was significant, *p* < .001. These results suggest that TMS reduced metacognitive sensitivity (Fig 3-3) specifically by decreasing visibility ratings for correct responses (as opposed to increasing visibility ratings for incorrect responses). Thus, TMS induced a kind of relative blindsight, to the extent that TMS suppressed the reports of visibility for accurately processed stimuli. * *p* < .005, n.s. denotes not significant. Error bars represent 1 SEM. **(B) Type 2 ROC analysis.** Individual data points indicate the type 2 hit rates and false alarm rates for every subject pre- and post-TMS. Type 2 ROC curves were estimated for each subject using estimates of meta-*d'*, *c'*, and *s*; the average of these ROC curves is plotted for the pre- and post-real TMS conditions. The distribution of individual data points and the fitted ROC curves indicate that TMS influenced metacognitive sensitivity, rather than just response bias. Note that the ROC curve reflects meta-*d'*, and thus is not as sensitive to the effect of TMS as the measure used in the analysis, meta-*d'* – *d'* (Fig 3-3), since some variation in meta-*d'* is attributable merely to variation in *d'* (Maniscalco & Lau, 2012; Appendix A).

performance, as measured by contrast level (Figure 3-2 B) and reaction time for correct trials (Figure 3-2 C).

We did not find a significant effect of TMS on averaged stimulus visibility itself. However, note that the effect of TMS is at least partially characterized by a change in visibility ratings, in that TMS reduced metacognitive sensitivity precisely by reducing visibility for correctly classified stimuli while leaving visibility for incorrectly classified stimuli unaffected (Figure 3-4 A). Indeed, although the interaction was not significant, separate paired t-tests show a difference in visibility pre- and post- real TMS (two-tailed, t(19) = 3.09, *p* = .002) but no difference pre- and post- sham TMS (t(19) = 1.47). There are two reasons why the design of the current study may not have been ideal to statistically detect an effect of TMS on overall stimulus visibility. One reason is that stimulus visibility was affected by an experimental factor other than TMS, namely the contrast levels of the stimuli, which were adjusted on-line throughout the experiment in order to hold discrimination performance constant. Another reason is that subjects were instructed to use visibility ratings in a relative manner, in order to distinguish stimuli that were relatively more or less visible. The instruction to rate visibility in this relative way may have obscured the extent to which visibility ratings reflected absolute differences in stimulus visibility across experimental conditions. Nonetheless, these limitations are not important for the main focus of this study, which is the metacognitive sensitivity of visibility ratings.

One typical argument against studies of awareness is that the manipulation in question might have only changed subjects' criteria for producing subjective ratings, rather than changing awareness *per se*. A change in response criterion is not necessarily uninteresting-- but more importantly, this is not what we found. Our type 2 SDT analysis demonstrates that TMS reduced metacognitive sensitivity (i.e. the efficacy with which subjective visibility ratings discriminate between correct and incorrect judgments), rather than merely affecting metacognitive response bias (i.e. the overall propensity to give high visibility ratings). TMS reduced visibility for correct trials (type 2 HR) but not for incorrect trials (type 2 FAR) (Figure 3-4 A), a pattern that cannot be attributed solely to changes in response bias. Likewise, our measure of type 2 sensitivity, meta-*d′* – *d′*, is not sensitive to changes in type 2 response bias (Maniscalco & Lau, 2012; Appendix A). We also demonstrate this point graphically in Figure 3-4 B, which shows the type 2 ROC points for each subject, and mean fitted ROC curves, pre- and post-TMS. The distribution of type 2 ROC points and the fitted type 2 ROC curves differ, indicating lower type 2 sensitivity following TMS. If TMS only affected subjects' criteria for reporting high visibility, the type 2 ROC curves pre- and post-TMS should overlap (Macmillan & Creelman, 2005), contrary to our findings.

Our results extend previous work. Similarly to the present study, Del Cul et al. (2009) showed that prefrontal lesions can affect subjective reports of visual experience more than visual task performance. Slachevsky (Slachevsky et al., 2001; Slachevsky et al., 2003) has shown that lesion to the prefrontal cortex can affect awareness in the monitoring of actions or sensory-motor readjustments. Other studies show that visual processing can be affected by lesion (Latto & Cowey, 1971) or TMS (Grosbras & Paus, 2003; Ruff et al., 2006) to the frontal eye field. Turatto, Sandrini, and Miniussi (2004) showed that TMS to the dlPFC can affect performance in change blindness. These studies show that, contrary to what critics have argued (Pollen, 1995), disruption of activity in the prefrontal cortex can in fact influence awareness and visual processing. What is new in the present study is that it specifically highlights the role of the prefrontal cortex in supporting the metacognitive sensitivity of visual awareness.

The prefrontal cortex is associated with many important cognitive functions, and therefore our interpretation is not that it is completely specific to the metacognitive sensitivity of visual awareness. It is likely that bilateral theta-burst TMS to the dlPFC would impair performance in other tasks where metacognitive visual awareness is not required. Instead of applying the same TMS treatment to unrelated control tasks and hoping to show a negative result in those situations, we show that TMS impaired a specific *process* involved in our task, namely metacognitive awareness, but not other processes involved in the same task. It is important to note that performance in the stimulus classification task was not influenced by TMS under the stimulation parameters currently used. Thus, it is unlikely that TMS affected metacognitive sensitivity by means of non-specific disturbances such as reductions in visual attention or general arousal.

As in Del Cul et al. (2009), one limitation of the present study is that we did not show that a similar effect could not be obtained in a control anatomical site. The lack of such control conditions is unfortunate and largely constrained by logistics (e.g. we did not have ethical approval for every brain regions for this relatively new TMS protocol, and the leading authors have since relocated elsewhere). However, given that the TMS was applied offline (i.e. not during task), and that the effect did not change basic task performance, it is unlikely that the results we obtained were due to the general distraction due to TMS. It is likely that TMS applied to an unrelated region, such as the somatosensory area, would not lead to our metacognitive effect. However, it remains an open question whether TMS applied to parietal areas that are connected to dlPFC would lead to similar results.

In any case, our conclusion is not that the neural circuitry that supports metacognitive visual awareness is completely localized in the dlPFC. Rather, we conclude that disruption of activity in this area can impair the metacognitive sensitivity of visual awareness. The present results show that the prefrontal cortex is functionally relevant to visual awareness, in that manipulation of the former can affect the latter. Further, the data clarify in what way the prefrontal cortex might contribute. Activity in the dlPFC may play a relatively unimportant role in representing the visual signal itself, but it may be essential for some form of internal uncertainty monitoring that allows observers to be able to distinguish when visual processing is effective and when it is not. It is this introspective and metacogntive aspect of visual awareness for which the prefrontal cortex may be critical.

Chapter 4

Limited cognitive resources explain a tradeoff between perceptual and metacognitive vigilance

Introduction

In Chapter 2 and Chapter 3 we demonstrated the existence of cognitive and neural interventions that have selective effects on relative metacognitive sensitivity, findings that suggest that objective and subjective visual performance are dissociable and may depend on separate underlying mechanisms. In this chapter we demonstrate the existence of another dissociation between perceptual and metacognitive sensitivity that occurs naturally over the course of time, presumably as a result of tradeoffs in perceptual and metacognitive performance necessitated by the onset of fatigue.

As an observer continuously performs a perceptual task, the observer's perceptual sensitivity tends to decline over time, an effect known as the vigilance decrement (Davies & Parasuraman, 1982; Warm, 1984; See, Howe, Warm, & Dember, 1995). Research has suggested that limited cognitive resources (Kahneman, 1973; Matthews, Davies, Westerman, & Stammers, 2000; Wickens, 2002) become depleted as a vigil progresses, and so the vigilance decrement is better accounted for by resource exhaustion than by mindlessness or task disengagement (e.g. Grier et al., 2003; Helton & Warm, 2008; Helton et al., 2005; Warm, Parasuraman, & Matthews, 2008). Consistent with the resource depletion account, the vigilance decrement is exacerbated by increasing task demands such as stimulus degradation, rate of stimulus presentation, and memory load (See et al., 1995), and is associated with depleted ratings of energetic arousal, elevated reports of stress, and declines in cerebral blood flow velocity (Warm et al., 2008).

A seemingly unrelated line of research involves the relationship between perceptual sensitivity and perceptual metacognition (e.g., confidence ratings). Recent work has developed a signal detection theory (SDT) analysis of confidence ratings (Galvin et al., 2003; Maniscalco & Lau, 2012), allowing for a bias-free measure of metacognitive sensitivity (i.e. an observer’s ability to discriminate between his own correct and incorrect judgments, regardless of his tendency to report high confidence). Of particular interest is how such measures of metacognition are related to perceptual performance. A tacit assumption of the classical SDT analysis of confidence rating data is that perceptual decisions and confidence ratings are based on the same underlying process (Galvin et al., 2003; Macmillan & Creelman, 2005; Maniscalco & Lau, 2012), and this view has received some empirical support (Kepecs et al., 2008; Kiani & Shadlen, 2009; Kepecs & Mainen, 2012). Other findings suggest that metacognition is subserved by high-level prefrontal mechanisms and is therefore partially dissociable from perceptual performance (e.g. Fleming et al., 2010; Pleskac & Busemeter, 2010; McCurdy et al., 2013).

In this chapter, we bring these two lines of research together by investigating the joint behavior of SDT measures of perceptual and metacognitive sensitivity over time. If a single process generates perceptual and metacognitive decisions, we should expect declines in perceptual sensitivity to be associated with declines in metacognitive sensitivity (Maniscalco & Lau, 2012). Conversely, if distinct processes generate perceptual and metacognitive decisions, we might expect vigilance decrements in perception and metacognition to be dissociable.

To anticipate, we find a robust effect whereby changes in perceptual and metacognitive sensitivity over time are weakly or negatively correlated, contrary to the strong positive correlation predicted by a single-process view of perception and metacognition. Voxel based morphometry analysis suggests that this finding is mediated by a common cognitive resource housed in anterior prefrontal cortex, a region previously associated with visual metacognitive sensitivity (Fleming et al., 2010; McCurdy et al., 2013). Consistent with this account, we find that alleviating metacognitive task demands reduces the perceptual vigilance decrement. Thus, perception and metacognition appear to be distinct processes that can differentially access limited cognitive resources.

Methods

Experiment 1

Data from this experiment were originally reported in Maniscalco & Lau (2012).

Participants

Thirty Columbia University students participated in the experiment. Participants gave informed consent and were paid $10 for approximately one hour of participation. The research was approved by the Columbia University's Committee for the Protection of Human Subjects.

Four participants were omitted from data analysis. One exhibited perfect task performance. The other three used an extreme confidence rating (lowest / highest rating) more than 89% of the time, an extreme bias in reporting confidence that renders meaningful analysis of metacognitive sensitivity difficult.

Experimental procedure

Participants were seated in a dimmed room 60 cm away from a computer monitor. Stimuli were generated using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) in MATLAB (MathWorks, Natick, MA) and were shown on an iMac monitor (LCD, 24 inches monitor size, 1920x 1200 pixel resolution, 60 Hz refresh rate).

On every trial, two stimuli were presented simultaneously, one 4° to the left of fixation and one 4° to the right (Figure 4-1 A). Stimuli were presented on a gray background for 33 ms. Each stimulus was a circle (3° diameter) consisting of randomly generated visual noise. The target stimulus contained a randomly oriented sinusoidal grating (2 cycles per degree) embedded in the visual noise. After stimulus presentation, participants provided a forced-choice judgment of whether the left or the right stimulus contained a grating. Following stimulus classification, participants rated their confidence in the accuracy of their response on a scale of 1 through 4. Participants were encouraged to use the entire confidence

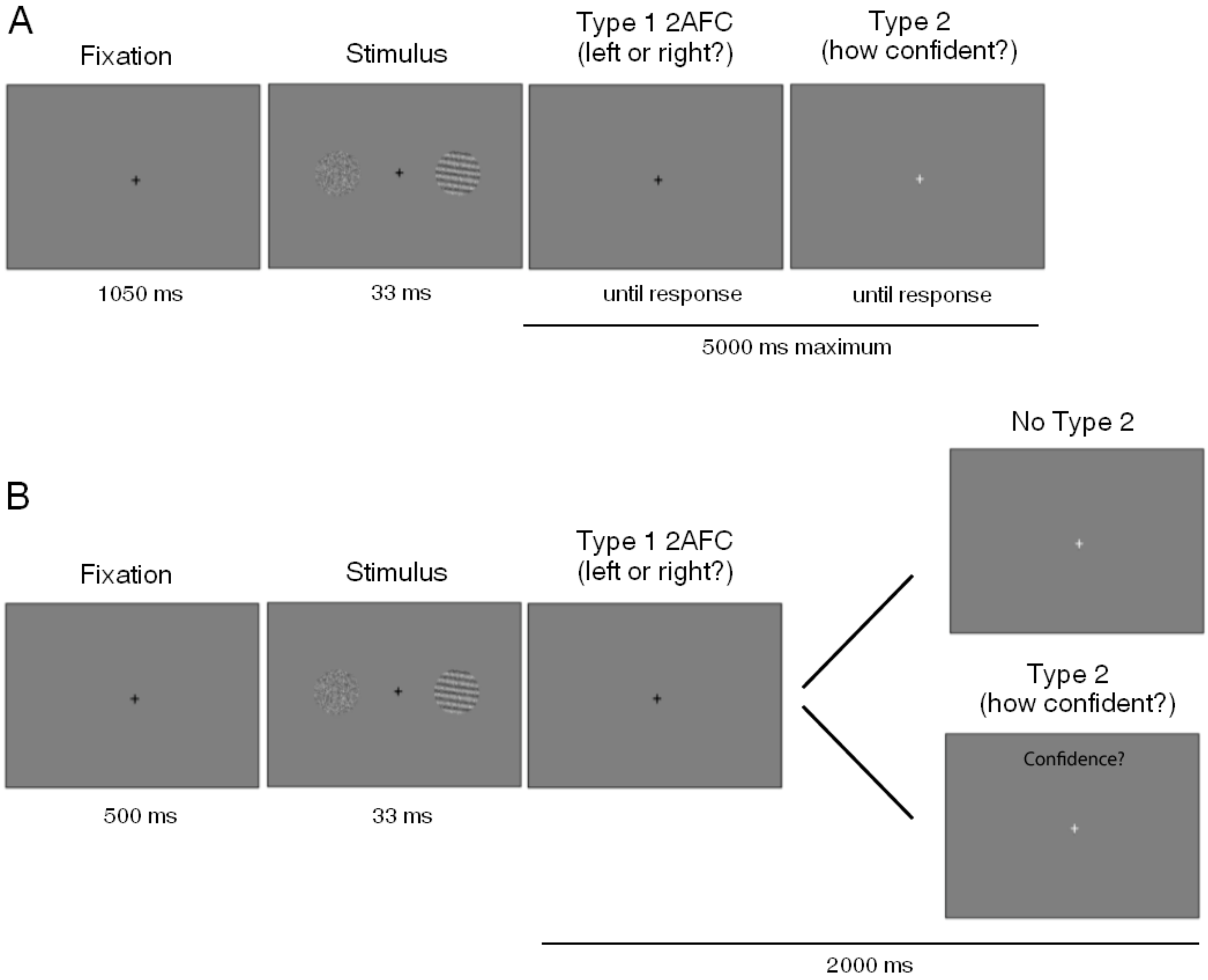

Figure 4-1. Design for Experiments 1 – 4. (A) Experiments 1 – 2. Subjects performed a spatial 2-interval forced choice task. On each trial, two patches of visual noise simultaneously appeared to the left and right of fixation. One of these patches contained an embedded sinusoidal grating. Subjects first indicated whether the left or right patch contained the grating, and then rated decision confidence on a scale of 1 – 4. Trial duration was determined by subject response time. **(B) Experiments 3 – 4.** Experiment 3 was similar to Experiments 1 – 2, except that in even-numbered blocks of trials ("partial type 2 blocks"), subjects were not required to rate confidence for the first half (50 trials) of the block. A written cue appeared above fixation on all trials where subjects were required to rate confidence. Trial duration was fixed to be 2.533 s. In Experiment 4, subjects wagered points rather than rating confidence, such that they won or lost the number of points wagered depending on the accuracy of the left/right decision. Subjects were also provided with feedback about wagering performance after each block.

scale. If the confidence rating was not registered within 5 seconds of stimulus offset, the next trial commenced automatically. (Such trials were omitted from all analyses.) There was a 1 s interval between the entry of confidence rating and the presentation of the next stimulus. Participants were instructed to maintain fixation on a small crosshair (.35° wide) displayed in the center of the screen for the duration of each trial.

At the start of each experimental session, participants completed 2 practice blocks (28 trials each) and 1 calibration block (120 trials). In the calibration block, the detectability of the grating in noise was adjusted continuously between trials on the basis of the participant's task performance using the QUEST threshold estimation procedure (Watson & Pelli, 1983). Target stimuli were defined as the sum of a grating with Michelson contrast $C_{grating}$ and a patch of visual noise with Michelson contrast C_{noise}. The total contrast of the target stimulus, $C_{target} = C_{grating} + C_{noise}$, was set to 0.9. The non-target stimulus containing only noise was also set to a Michelson contrast of 0.9. The QUEST procedure was used to estimate the ratio of the grating contrast to the noise contrast, $R_{g/n} = C_{grating} / C_{noise}$, which yielded 75% correct performance in the 2AFC task. Three independent threshold estimates of $R_{g/n}$ were acquired, with 40 randomly ordered trials contributing to each, and the median estimate of these was used to create stimuli for the main experiment.

Experiment 2

Data from this experiment were originally reported in McCurdy et al. (2013).

Participants

Forty-one Radboud University students participated in the experiment. Participants gave informed consent and were paid €8 for approximately one hour of participation. The research was approved by the local ethics committee where the experiment was performed (CMO region Arnhem-Nijmegen, the Netherlands).

Experimental procedure

Experimental design was identical to Experiment 1, with the following exceptions.

Blocks of the visual perception task were interleaved with blocks of a memory task. (Comparison of visual and memory task performance is explored in McCurdy et al. (2013); data from the memory task is not analyzed here.) Each participant completed two experimental sessions on two consecutive days. On day 1, participants completed two practice blocks of the visual task, a calibration block for the visual task, and two blocks of the visual task consisting of 102 trials each. On day 2, participants completed three more blocks of the visual task, using the stimulus settings acquired from the calibration block on day 1. As with Experiment 1, trial duration for the visual task was determined by response times, and participants experienced a self-terminated rest period of up to a minute between blocks.

Rather than using a single value for the ratio of grating and noise contrast ($R_{g/n}$), as in the previous experiment, three different levels of $R_{g/n}$ were used across trials. The calibration block determined the highest level of $R_{g/n}$, $R^*_{g/n}$, and the two lower levels of contrast ratio were determined by multiplying $R^*_{g/n}$ by 0.75 and 0.5. In this manuscript, all analyses for Experiment 2 collapse across contrast level in order to yield sufficient trials for SDT analysis.

Image acquisition

For thirty-two of the participants, a 1.5T Avanto MR-scanner (Siemens, Erlangen, Germany), using a 32-channel head coil, was used to acquire the T1-weighted anatomical MRI images (176 slices, echo time = 2.95ms, TR = 2250ms, voxel size 1 mm isotropic). The remaining nine participants were scanned using different scanning parameters, and this was included as a covariate in the multiple regression design matrix in SPM8.

Voxel-based morphometry analysis

VBM preprocessing was carried out using SPM8 (http://www.fil.ion.ucl.ac.uk/spm). Similar to the pre-processing protocol used by Fleming et al. (2010), the scans were first segmented into gray matter, white matter, and cerebral spinal fluid in native space. DARTEL (Ashburner, 2007) was used to increase the accuracy of inter-subject alignment by aligning and warping the gray matter images to an iteratively improved template. The DARTEL template was then registered to MNI space, and then gray matter images were modulated such that their tissue volumes were preserved. Images were smoothed using an 8 mm full width at half maximum Gaussian kernel. The resultant pre-processed gray matter images were analyzed using MarsBar v0.42 software (marsbar.sourceforge.net). ROIs were defined as a 10 mm sphere around each of the two peak voxel coordinates identified by McCurdy et al. (2013) (peak voxel coordinate for left aPFC = [-12 54 16]; peak voxel coordinate for right aPFC = [32 50 7]; both survived cluster family-wise-error correction) and the gray matter volumes in each sphere was calculated.

<u>Experiment 3</u>

Participants

Twenty-one Columbia University students participated in the experiment. Participants gave informed consent and were paid $10 for approximately one hour of participation. The research was approved by the Columbia University's Committee for the Protection of Human Subjects.

One participant was omitted from data analysis due to using the highest confidence rating on 96% of all trials, an extreme bias in reporting confidence that renders meaningful analysis of type 2 data difficult.

Experimental procedure

Experimental design was identical to Experiment 1, with the following exceptions.

The primary manipulation of interest in Experiment 3 was that in odd-numbered experimental blocks, participants did not provide confidence ratings in the first 50 of 100 trials in the block. Call these “partial type 2 blocks,” as opposed to “whole type 2 blocks” in which confidence ratings were provided on all trials. Before each block, participants were instructed which kind of block was about to be presented. For partial type 2 blocks, the instruction read as follows: “Upcoming block: There will be NO CONFIDENCE RATING for the first 50 trials. Do not enter confidence ratings until you are prompted to do so.” For whole type 2 blocks, the instruction read “Upcoming block: There will be confidence rating on EVERY trial.”

In order to clearly distinguish trials in which confidence ratings were and were not required, a text prompt reading “Confidence?” was displayed on every trial where confidence ratings were required. The prompt was displayed 6.4° above fixation and appeared after successful entry of the perceptual decision.

Because some trials did not require confidence ratings, partial type 2 blocks would be shorter in duration than whole type 2 blocks if trial duration depended on participant response times, as it did in Experiments 1 and 2. Therefore, in order to standardize the temporal duration of the experiment, the duration of each trial and the duration of each break period were set constant. In Experiment 1, participants entered both the stimulus judgment and confidence rating in 2 seconds or less for 92% of all trials. Therefore, following each stimulus presentation in Experiment 2, there was a fixed response period of 2 seconds, during which participants had to enter the required stimulus and confidence responses. After the response period and prior to the next stimulus presentation, a crosshair was displayed for 0.5 sec. Altogether, each trial lasted 2.533 s, a close match to the mean trial duration of 2.315 s in Experiment 1. Additionally, all break periods between blocks were set to 1 minute. When only 10 s of break time were left, three auditory tones alerted participants to prepare for the upcoming block, and a timer counting down the remaining seconds of the break period was presented on the screen.

Experiment 4

Participants

Thirty-three Columbia University students participated in the experiment. Participants gave informed consent and were paid $10 for approximately one hour of participation. The research was approved by the Columbia University's Committee for the Protection of Human Subjects.

Six participants were omitted from data analysis due to using the highest point wager (see below) on 93% of all trials, an extreme bias in wagering that renders meaningful analysis of type 2 data difficult.

Experimental procedure

Experimental design was identical to Experiment 3, with the following exceptions.

In Experiment 4, the confidence rating system was replaced with a point wagering system. Participants were instructed that, following each stimulus identification response, they would sometimes be prompted to wager points on their stimulus decision. Participants could wager between 1 and 4 points. For correct trials, the number of wagered points was added to a running point tally, whereas for incorrect trials, the number of wagered points was subtracted from the tally.

Participants were instructed that their goal was to maximize the number of points they received over the whole course of the experiment. They were given the following guidelines for maximizing points: (1) get as many stimulus decisions correct as possible; (2) although the optimal wagering strategy is to wager 4 points for correct trials and 1 point for incorrect trials, the participant does not have perfect knowledge of which trials are incorrect, and high wagers for incorrect responses are costly. Thus, participants were encouraged to wager points according to the best estimate of the likelihood that the stimulus response was correct, and so the entire wagering scale should be utilized in order to reflect variations in this estimated likelihood across trials.

For non-wagering trials, participants were instructed that correct trials would add 3 points to their tally and incorrect trials would subtract 3 points from their tally. Additionally, in order to incentivize participants to enter all required responses for each trial, they were informed that 10 points were subtracted from the tally for any trial where not all required responses were entered within the 2 second time limit.

During break periods, participants were provided with feedback on their wagering performance. They were shown how many points they had earned in the previous block, how many points they could have earned with an “optimal” wagering strategy (i.e., had they wagered 4 points for all correct responses and 1 point for all incorrect responses), and their overall wagering efficiency (the former quantity divided by the latter). The same information was provided for overall wagering performance across all blocks thus far completed.

The text prompts used in Experiment 3 to inform participants which kind of block was about to come up, and to prompt them to enter wagers on trials where wagers were required, were the same as in Experiment 2 except the word “confidence” was replaced by “wager.”

Data analysis

We measured perceptual and metacognitive performance in the visual task using signal detection theory (SDT) analysis (Green & Swets, 1966; Macmillan & Creelman, 2005; Appendix A). We defined hit rate (HR) as the probability that the subject reported that the grating was on the right, given that the grating was on the right, and false alarm rate (FAR) as the probability that the subject reported that the grating was on the right, given that the grating was on the left. We calculated *d′* = z(HR) – z(FAR) and used *d′* to quantify sensitivity in the visual discrimination task.

We similarly quantified metacognitive sensitivity, i.e. the efficacy with which confidence ratings discriminate between a subject's own correct and incorrect responses, with meta-*d′* (Maniscalco & Lau, 2012; Appendix A). Specifically, we found the value of meta-*d′* that jointly maximized the likelihood of the response-specific type 2 ROC curves, where response-specific type 2 ROC curves are derived from "type 2" probabilities of the general form P(confidence = c | stimulus = s and response = r). Maximization of likelihood was achieved using the Optimization Toolbox in MATLAB (MathWorks, Natick, MA). Essentially, estimating meta-*d′* in this analysis amounts to fitting the SDT model to the type 2 probabilities for every possible permutation of stimulus, response, and confidence level. Please see Appendix A for a more in-depth treatment of the methodology for estimating meta-*d′*.

Monte Carlo SDT simulations

We performed Monte Carlo SDT simulations in order to assess the extent to which observed changes in perceptual and metacognitive performance over time deviated from SDT expectation. We structured the simulation so as to closely mirror key features of the empirical data across Experiments 1 – 4.

For each subject in Experiments 1 – 4, we binned together all trials occurring in the first half of an experimental block and computed *d′* (call this d'_1), and similarly binned together all trials occurring in the second half of an experimental block and computed *d′* (call this d'_2). (For Experiments 3 and 4, data was gathered only from blocks in which confidence ratings or wagers were provided on every trial.) Visual inspection of the scatterplot of d'_2 vs d'_1 suggested that these variables were roughly distributed as a bivariate normal distribution. Therefore, we computed the mean $\begin{bmatrix}1.6464 & 1.5637\end{bmatrix}$ and covariance

$\begin{bmatrix} 0.4520 & 0.4105 \\ 0.4105 & 0.4590 \end{bmatrix}$ for *d′*$_1$ and *d′*$_2$ across all experiments, and used a bivariate normal distribution with this mean and covariance as the basis for subsequent statistical sampling.

In Experiment 1, 500 trials contributed to each estimate of *d′*$_1$ and *d′*$_2$, whereas this number was reduced to 255 trials in Experiment 2 and 250 trials in Experiments 3 and 4 (after limiting the analysis to blocks where confidence ratings were provided on every trial). Therefore, in all simulations, 250 simulated "trials" contributed to the estimate of each SDT parameter for each simulated subject. Because the average number of subjects entered into the analysis for Experiments 1 – 4 was 28.5, each simulation contained data for 30 simulated subjects.

Simulation procedure

Simulations proceeded as follows. We simulated 2000 experiments, where each experiment had 30 simulated subjects, with a total of 500 simulated trials for each subject.

For each subject, we first obtained "true" values for *d′*$_1$ and *d′*$_2$ by randomly sampling from the bivariate normal distribution described above. If this resulted in any negative values, the sampling procedure was repeated until both *d′* values were positive. These "true" *d′* values were used as the basis for subsequent sampling in order to obtain "simulated" values for *d′* and meta-*d′*, as described below.

We also created a unique set of decision criteria for each subject. Decision criteria were initialized to values of -2, -1.75, -.75, 0, .75, 1.75, 2. In order to create different decision criteria for different simulated subjects, a small amount of random noise from *N*(0, .5) was added to the initial values of the decision criteria. Decision criteria were then re-sorted to ensure they were in ascending order. Once the values of the decision criteria were determined for a simulated subject, these same criteria values were used for all simulated trials without any further variation.

For the first block half consisting of 250 trials, 125 simulated “S1” trials (corresponding to the experimental condition where the grating was on the left) generated 125 sensory samples drawn from the normal distribution $N(-d'_1/2,\ 1)$. Another 125 simulated “S2” trials (corresponding to the experimental condition where where the grating was on the right) generated 125 sensory samples drawn from $N(+d'_1/2,\ 1)$. (These reflect the normal distributions of sensory evidence contingent on stimulus presentation posited by SDT.) Each such sample was compared to the decision criteria, and this comparison determined the simulated subject’s response for each trial (Macmillan & Creelman, 2005). Responses for the perceptual task could be either “S1” (i.e. “grating was on the left”) or “S2” (i.e. “grating was on the right”), and responses for the metacognitive task were a confidence rating ranging from values of 1 through 4. A similar procedure was used to simulate sensory samples and behavioral responses for the second block half of 250 trials.

Now that each trial was associated with a “true” stimulus configuration as well as the simulated subject’s perceptual and metacognitive judgments, we were able to compute *d’* and meta-*d’* for the first and second block half for each simulated subject using standard SDT analyses (Macmillan & Creelman, 2005; Maniscalco & Lau, 2012).

Modulation of simulated results using aPFC data from Experiment 2

Analysis of Experiment 2 suggested a model whereby aPFC gray matter volume is positively associated with both meta-d'_1 and Δ*d’* (Results; Figure 4-5). In order to take these effects into account in the simulations, we used the following procedure. Using the data from Experiment 2, we applied a regression analysis to estimate the β values for the following equation:

(1) $\text{aPFC}_{\text{data}} = \beta_1 * \Delta d'_{\text{data}} + \beta_0$

For the analysis of metacognitive performance, we defined the ratio of meta-*d’* to *d’* in the first block half as

(2) $M_{1,\text{ data}}$ = meta-$d'_{1,\text{ data}}$ / $d'_{1,\text{ data}}$

Using data from Experiment 2, we applied another regression analysis to estimate the β values for the following equation:

(3) $M_{1,\text{ data}} = \beta_3 * \text{aPFC}_{\text{data}} + \beta_2$

On the basis of the β values obtained from the regression analysis on (1) and the simulated values of Δ*d'*, we assigned each simulated subject an aPFC volume:

(4) $\text{aPFC}_{\text{sim}} = \beta_1 * \Delta d'_{\text{sim}} + \beta_0$

We then used the obtained value of aPFC_{sim} to adjust the simulated subject's simulated value for M_1 as follows.

(5) $M_{1,\text{ adj}} = \beta_3 * \text{aPFC}_{\text{sim}} + M_{1,\text{ sim}}$

Since $\text{aPFC}_{\text{data}}$ was scaled in such a way that the mean value was 0, the coefficient β_2 derived from regression analysis on equation (3) codes the mean value of M_1 in the data, which was 0.865. However, in SDT simulations, the mean value of M_1 was 0.996, consistent with the SDT expectation that meta-*d'* = *d'* and therefore meta-*d'* / *d'* = 1 (Maniscalco & Lau, 2012). Thus, applying the β_2 coefficient to the simulated data would have been inappropriate. Instead, we replaced the β_2 coefficient (an estimate of the mean value of M_1 in the empirical data) with the actual value of M_1 derived for each simulated subject. This has the benefit of retaining natural between-subject sampling variation in the values of M_1 arising from the Monte Carlo simulation procedure when calculating the value for $M_{1,\text{ adj}}$.

Finally, we obtained a new value for meta-$d'_{1,\text{ sim}}$ using the following equation.

(6) meta-$d'_{1,\text{ adj}} = M_{1,\text{ adj}} * d'_{1,\text{ sim}}$

Results of the "SDT + aPFC" simulation displayed in Figure 7 are derived by taking the same simulated data used for the "SDT" simulation, with the exception that values of meta-$d'_{1,\text{ sim}}$ were replaced by the value of meta-$d'_{1,\text{ adj}}$ calculated for each simulated subject. This had the effect of modulating the simulated data set such that the relationships between simulated aPFC volume, Δ*d'*, and meta-d'_1 were similar to the relationships empirically observed in Experiment 2.

Analysis of correlations between Δmeta-d' and Δd'

Each of the 2000 simulated experiments contained 30 simulated subjects, and so 30 values of $\Delta d'_{\text{sim}}$ and Δmeta-d'_{sim}. For each simulated experiment, we calculated the Pearson's *r* correlation coefficient between $\Delta d'_{\text{sim}}$ and Δmeta-d'_{sim}. In order to mitigate the influence of outliers, we excluded data from all simulated subjects with any *d'* value lower than 0.25 or higher than 3.

This resulted in 2000 simulated values for Pearson's *r*. We used these 2000 values in order to estimate the sampling distribution of *r* with and without the aPFC modulation of the SDT simulations, as displayed in Figure 7C. Estimates of the sampling distribution in turn allowed us to characterize the likelihood of the empirically observed *r* values in Experiments 1 – 4 under the null SDT model and the SDT model augmented by the aPFC findings.

Regression of Δmeta-*d'* onto Δ*d'*

For Experiments 1 – 4, one analysis of interest was to characterize the empirical relationship between Δmeta-*d'* and Δ*d'*. Ideally, regressions between these variables should take into account that both are subject to sampling error. However, errors-in-variables approaches to regression typically require some knowledge or assumptions about the error structures of the dependent and independent variables.

We capitalized on the results of the Monte Carlo SDT simulations in order to characterize the error structures for these measures. As described above, for each simulated subject, we selected "true" values for $d'_{1,\text{ true}}$ and $d'_{2,\text{ true}}$, and then repeatedly performed a sampling procedure using the SDT model parameterized with $d'_{1,\text{ true}}$ and $d'_{2,\text{ true}}$ in order to obtain corresponding "simulated" values $d'_{1,\text{ sim}}$, $d'_{2,\text{ sim}}$, meta-$d'_{1,\text{ sim}}$, and meta-$d'_{2,\text{ sim}}$. For each simulated subject, we calculated the sampling error for Δ*d'* as

(7) $\text{error}_{\Delta d'} = \Delta d'_{\text{true}} - \Delta d'_{\text{sim}}$

Likewise, since on the basic SDT model used here, meta-*d'* = *d'*, it follows that $\Delta\text{meta-}d'_{\text{true}} = \Delta d'_{\text{true}}$. Thus, we calculated error for meta-*d'* as

(8) $\text{error}_{\Delta\text{meta-}d'} = \Delta d'_{\text{true}} - \Delta\text{meta-}d'_{\text{sim}}$

Sampling errors for Δ*d'* and Δmeta-*d'* were not correlated (Pearson's $r = -0.015$). Therefore, it was appropriate to use Deming regression to characterize their relationship (Deming, 1943). Deming regression requires knowing the value of the parameter δ, which is the ratio of the variances of error in the dependent and independent variables. On the basis of the simulation outcomes, we estimated that $\delta = \text{var}(\text{error}_{\Delta\text{meta-}d'}) / \text{var}(\text{error}_{\Delta d'}) = 2.1535$. Therefore, for all regressions of Δmeta-*d'* onto Δ*d'* reported in this paper, we used Deming regression with δ = 2.1535.

Results

According to SDT, for an ideal observer, perceptual performance should be related to metacognitive performance such that *d'* = meta-*d'* (Maniscalco & Lau, 2012). Deviations from this expectation due to sampling error and suboptimal metacognitive performance are to be expected, but a robust SDT prediction is that between and within subjects, *d'* and meta-*d'* should positively correlate. Substantiating this general prediction, we found that overall *d'* correlated positively with overall meta-*d'* in all four experiments (*r*s > .5, *p*s < .03), and also that *d'* correlated positively with meta-*d' within* each block-half for all experiments (*r*s > .4, *p*s < .02), with the lone exception of a non-significant correlation in the first block-half of Experiment 3 (*r* = .15, *p* = .5). It is thus surprising that, although *d'* and meta-*d'* were consistently positively correlated overall and *within* each block half, changes in these measures *across* block half consistently failed to positively correlate, as described below.

Experiment 1

In Experiment 1, we analyzed the time course of perceptual and metacognitive performance within brief experimental blocks of trials. Subjects completed 10 blocks, each containing 100 trials, and received a self-terminated rest period of up to one minute between blocks.

On each trial, two patches of visual noise were presented to the left and right of fixation, and in one of these patches a sinusoidal grating was present (Figure 4-1 A). Subjects provided a two interval forced choice discrimination on whether the grating was in the left or right stimulus, and then rated decision confidence on a scale of 1 through 4. Trial duration depended on response times; mean block duration across all participants was 231.5 s.

To analyze perceptual and metacognitive performance over the course of a block of trials, we binned the trials from the first and second half of each experimental block. Thus, bin 1 contained the first 50 trials from all 10 blocks, for a total of 500 trials, and bin 2 similarly contained the last 50 trials from all 10 blocks. For each block half, we measured stimulus identification performance independently of response bias by using the signal detection theoretic (SDT) measure *d′* (Macmillan & Creelman, 2005). We measured metacognitive sensitivity (i.e. how well confidence ratings track accuracy) independently of biases in confidence rating using the SDT-inspired measure meta-*d′* (Maniscalco & Lau, 2012). For convenience, we will sometimes refer to stimulus identification performance as “type 1 performance” and metacognitive sensitivity as “type 2 performance” (Clarke et al., 1959). Because meta-*d′* expresses type 2 performance on the same scale as the type 1 measure *d′*, numerical values of meta-*d′* and *d′* may be compared directly. Meta-*d′* is defined such that meta-*d′* = *d′* for an observer whose metacognitive performance conforms to the expectations of the classical SDT model (Maniscalco & Lau, 2012).

The results of this analysis are plotted in Figure 4-2 A. To assess the effects of time passage within a block of trials on task performance, we conducted a 2 (Task Type: type 1, type 2) x 2 (Time: 1st block half, 2nd block half) repeated measures ANOVA. The ANOVA revealed a significant Task Type x

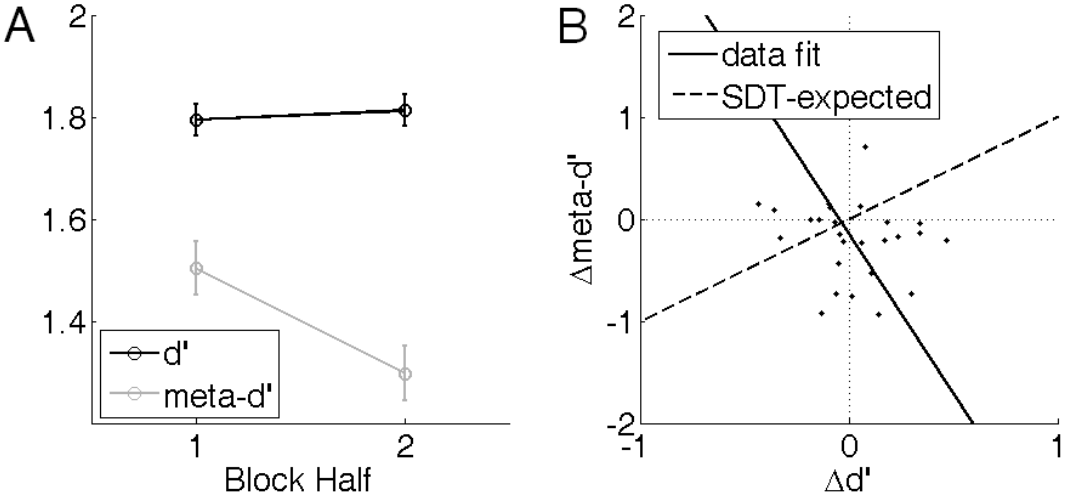

Figure 4-2. Results for Experiment 1. (A) Mean perceptual (*d'*) and metacognitive (meta-*d'*) performance over time. We analyzed the dynamics of subjects' task performance over the timecourse of blocks of 100 trials. Interestingly, we observed that meta-*d'* decreased over time whereas *d'* remained constant (Task Type x Time, *p* = .02), suggestive of a selective fatigue effect for metacognition. Error bars represent within-subjects standard errors (Morey, 2008). (**B) Between-subject correlation of changes in perceptual and metacognitive performance.** We computed the change in *d'* and meta-*d'* between the first and second half of all blocks (i.e. $\Delta d' = d'_{\text{2nd block half}} - d'_{\text{1st block half}}$; $\Delta\text{meta-}d' = \text{meta-}d'_{\text{2nd block half}} - \text{meta-}d'_{\text{1st block half}}$) and found that these measures were inversely related, in stark contrast to SDT expectation. This suggests a tradeoff effect whereby maintenance of perceptual performance comes at the expense of maintenance in metacognitive performance, and vice versa.

Time interaction (*p* = .024), driven by the fact that over time, *d'* remained constant (*p* = .7) whereas meta-*d'* decreased (*p* = .011).

We also assessed the relationship between changes in perceptual task performance and metacognitive performance over time. For each participant, we calculated *d'* and meta-*d'* using trials from the first and second halves of all blocks. We defined $\Delta d' = d'_2 - d'_1$ and $\Delta\text{meta-}d' = \text{meta-}d'_2 - \text{meta-}d'_1$, where subscripts indicate block half. Since the expectation under SDT is that meta-*d'* = *d'*, it follows that under SDT expectation, Δmeta-*d'* = Δ*d'*.

Contrary to SDT expectation, empirically Δ*d′* and Δmeta-*d′* exhibited an inverse relationship (Figure 4-2 B). The observed Pearson's *r* correlation was -.18, whereas the SDT-expected value for *r*, according to computational simulations, was .41 (see Figure4- 7 C and "Monte Carlo SDT simulations" in Methods). Under the null hypothesis that changes in *d′* and meta-*d′* are generated by an SDT process with an expected *r* = .41, we estimate that the empirically observed *r* = -.18 corresponds to a one-tailed *p*-value of 0.0015 (Figure 4-7 C). Thus, according to SDT, the observed inverse relationship between Δ*d′* and Δmeta-*d′* is highly unlikely. The Deming regression slope relating Δ*d′* and Δmeta-*d′* was -3.12, lower than the SDT-expected value of 1 (see "Regression of Δmeta-*d′* onto Δ*d′*" in Methods).

These findings suggest that the observed dynamics of perceptual and metacognitive performance over time violate SDT expectation in multiple respects. We found that over time, meta-*d′* decreased even as *d′* remained constant (Figure 4-2 A), suggesting that overall, perceptual and metacognitive performance may vary independently. We further found that changes in perceptual and metacognitive performance over time were negatively, not positively, correlated (Figure 4-2 B). This surprising result is suggestive of a tradeoff in vigilance for the two types of tasks. In order to shed further light on this finding, we used a similar experimental design in Experiment 2 and collected structural MRI data.

Experiment 2

In Experiment 2, we used a nearly identical task design as in Experiment 1, with minor adjustments (see Methods). For each subject, we collected estimates of gray matter volume and analyzed the relationship between task performance over time and brain structure using voxel-based morphometry.

As in Experiment 1, trial duration was not fixed, but determined by participant response times. On average, each block of 102 trials lasted 242.2 s. We performed a 2 (Task Type: type 1, type 2) x 2

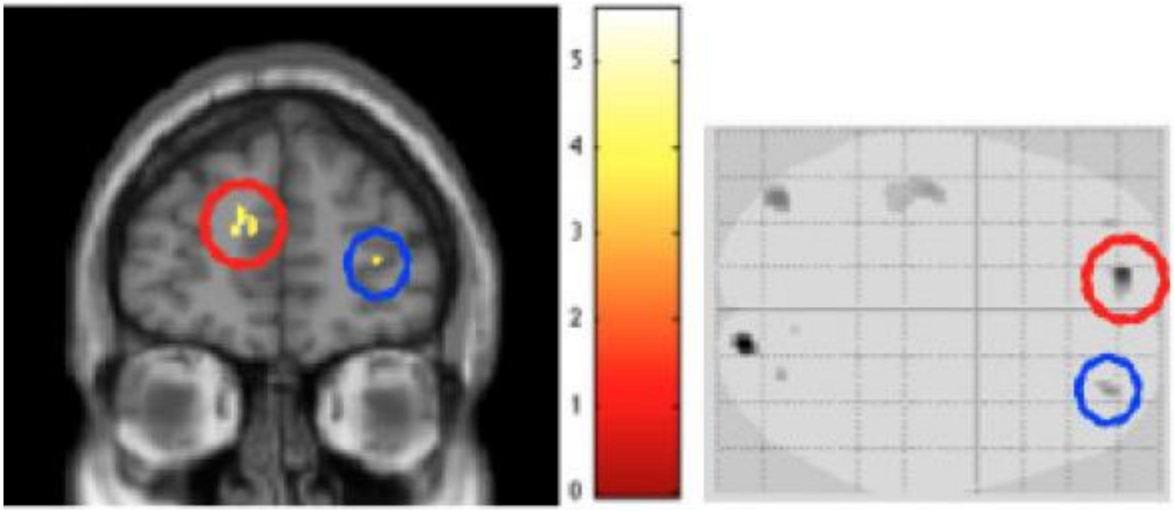

Figure 4-3. Brain regions selected for voxel based morphometry analysis. Two regions of interest in anterior prefrontal cortex (aPFC) were selected for analysis on the basis of positive correlations with metacognitive efficiency (meta-*d’* / *d’*). These regions were identified in a previous analysis of the data, conducted in McCurdy et al. (2013), and are consistent with previous findings relating metacognitive sensitivity to aPFC gray matter volume (Fleming et al., 2010). To obtain the most robust estimate of aPFC volume, we combined both aPFC clusters to produce an average volume, as in McCurdy et al. Peak voxel coordinate for left aPFC = [-12 54 16]. Peak voxel coordinate for right aPFC = [32 50 7]. Both survived cluster family-wise-error correction. Figure adapted from McCurdy et al.

(Time: 1st block half, 2nd block half) repeated measures ANOVA, but there was neither a main effect of Time (*p* = .3) nor a Task Type x Time interaction (*p* = .14; Figure 4-4 A). However, as in Experiment 1, the between-subject relationship between Δ*d’* and Δmeta-*d’* failed to conform to SDT expectation (Figure 4-4 B). The Pearson’s *r* correlation coefficient for Δ*d’* and Δmeta-*d’* was .07, whereas the SDT-expected value for *r*, according to computational simulations, was .41 (see Figure 4-7 C and “Monte Carlo SDT simulations” in Methods). Under the null hypothesis that changes in *d’* and meta-*d’* are generated by an SDT process with an expected *r* = .41, we estimate that the empirically observed *r* = .07 corresponds to a one-tailed *p*-value of 0.026 (Figure 4-7 C). The Deming regression slope relating Δ*d’* and Δmeta-*d’* was 0.18, lower than the SDT-expected value of 1 (see “Regression of Δmeta-*d’* onto Δ*d’*” in Methods). Thus,

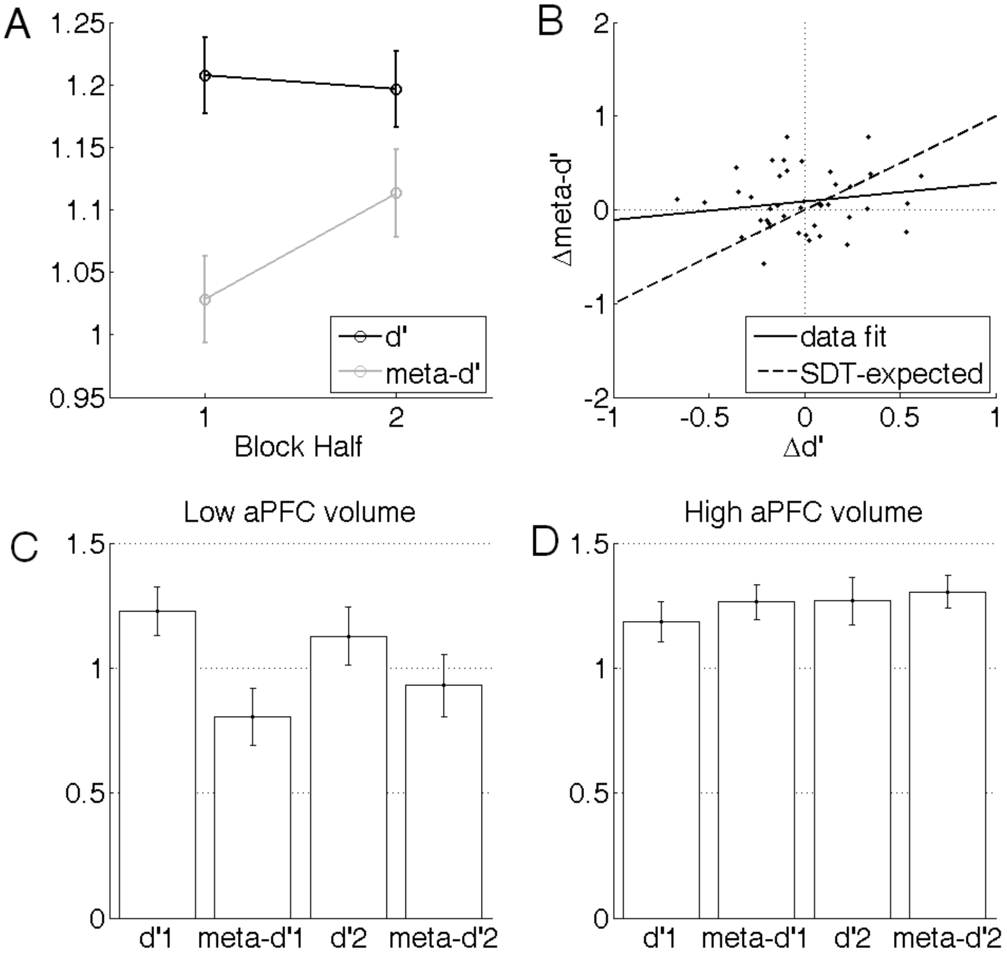

Figure 4-4. Results for Experiment 2. A) Mean perceptual (*d'*) and metacognitive (meta-*d'*) performance over time. On average, neither perceptual nor metacognitive performance changed over time (Time, *p* = .3; Task Type x Time, *p* = .14). **B) Between-subject correlation of changes in perceptual and metacognitive performance.** As in Experiment 1, the relationship between changes in *d'* and meta-*d'* failed to match SDT expectation. **C, D) Perception and metacognition as a function of aPFC volume.** A median split analysis revealed that subjects with lower aPFC volume tended to experience decreases in *d'* and increases in meta-*d'* (Task Type x Time x aPFC Volume, *p* = .03), contrary to the pattern of subjects with higher aPFC volume. This suggests that the between-subject inverse relationship between changes in *d'* and meta-*d'* may be partially accounted for by individual differences in aPFC volume. Error bars in panels A, C, and D represent within-subjects standard errors (Morey, 2008).

as in Experiment 1, the relationship between maintenance of perceptual and metacognitive performance was below SDT expectation.

We went on to relate this variability in metacognitive efficiency to inter-individual differences in brain structure. In a previous study (McCurdy et al., 2013), we defined a measure of metacognitive efficiency on a visual behavioral task (Figure 4-1 A) as the ratio meta-*d′* / *d′*; for SDT-ideal observers, this ratio should equal 1, and for metacognitively suboptimal observers it should be less than 1. Voxel-based morphometry analysis revealed that metacognitive efficiency was positively correlated with gray matter volume in regions in anterior prefrontal cortex (aPFC) (Figure 4-3; adapted from McCurdy et al.). In the present study, we focused on the two regions in the aPFC identified by McCurdy et al. as regions of interest (ROIs). (Peak voxel coordinate for left aPFC = [-12 54 16]; peak voxel coordinate for right aPFC = [32 50 7], both survived cluster family-wise-error correction.) The two clusters were used to define ROIs using the MarsBar toolbox (Brett, Anton, Valabregue, & Poline, 2002) and gray matter volume in the aPFC clusters was calculated. To obtain the most robust estimate of aPFC volume, we combined both aPFC clusters in the region to produce an average volume, as in McCurdy et al.; all subsequent analyses refer to this combined data as aPFC.

In order to assess how aPFC volume influenced the tradeoff effect, we performed a median split on aPFC volume and calculated *d′* and meta-*d′* over time for subjects with low and high aPFC volume (Figure 4-4 C-D). A 2 (Task Type: type 1, type 2) x 2 (Time: 1st block half, 2nd block half) x 2 (aPFC Volume: low / high) ANOVA revealed a significant Task Type x aPFC Volume interaction (*p* = .002) and a significant Task Type x Time x aPFC Volume interaction (*p* = .03). On average, subjects with high aPFC volume did not exhibit decreases in *d′* or meta-*d′* over time (Task Type x Time, *p* = .7), and were also metacognitively optimal in the sense that meta-*d′* was not significantly different from *d′* (Task Type, *p* = .4). By contrast, subjects with low aPFC volume were metacognitively suboptimal overall in the sense that meta-*d′* was significantly lower than *d′* (Task Type, *p* = .002). Crucially, low aPFC subjects also exhibited a numerical decrease in *d′* over time as well as an increase in meta-*d′*, such that the interaction was significant (Task Type x Time, p = .01). This pattern of changes in *d′* and meta-*d′* having the opposite sign for low aPFC subjects mirrors the tradeoff effect exhibited in Experiment 1 (Figure 4-2 B), Experiment 2 (Figure 4-4 B), and Experiments 3 and 4 (Figure 4-6 C). Thus, individual differences in aPFC volume are a candidate mechanism to explain the observed tradeoff effect between Δ*d′* and Δmeta-*d′*.

We further explored the relationship of aPFC volume to changes over time in *d′* and meta-*d′* by analyzing the patterns of correlation between d'_1, meta-d'_1, d'_2, meta-d'_2, and aPFC volume. As expected, d'_1 and d'_2 positively correlated (Pearson's r = 0.82, p < .001; Figure 4-5 A). Consistent with SDT expectation, meta-*d′* positively correlated with *d′* in each block half (*r*s = 0.57, 0.51; *p*s < .001; Figure 4-5 B).

aPFC volume did not correlate with either d'_1 (p = .8) or d'_2 (p = .2), but a partial correlation between aPFC volume and d'_2, controlling for d'_1, was significant (r = .33, p = .03; Figure 4-5 C). Thus, larger aPFC volume was associated with better perceptual vigilance (higher Δ*d′*).

aPFC volume was significantly correlated with meta-d'_1 (r = .43, p = .005), and this correlation remained significant when controlling for d'_1 (r = .50, p = .001; Figure 5C). Although aPFC volume also correlated with meta-d'_2 (r = .33, p = .04), this correlation did not remain significant when controlling for d'_2 (r = .26, p = .1) or meta-d'_1 (r = -.02, p = .9). Indeed, although aPFC regions were selected on the basis of their correlation with overall meta-*d′* / *d′* (r = .34, p = .03), aPFC volume correlated with meta-d'_1 / d'_1 (r = .51, p = .0006) but not meta-d'_2 / d'_2 (r = .1, p = .5). Thus, aPFC volume robustly correlated with metacognitive sensitivity only in the first block half. The significant correlation between aPFC volume and meta-d'_2 appears to be attributable to the fact that aPFC volume correlates with d'_2, which in turn correlates with meta-d'_2. Because larger aPFC volume was associated with higher *initial* metacognitive

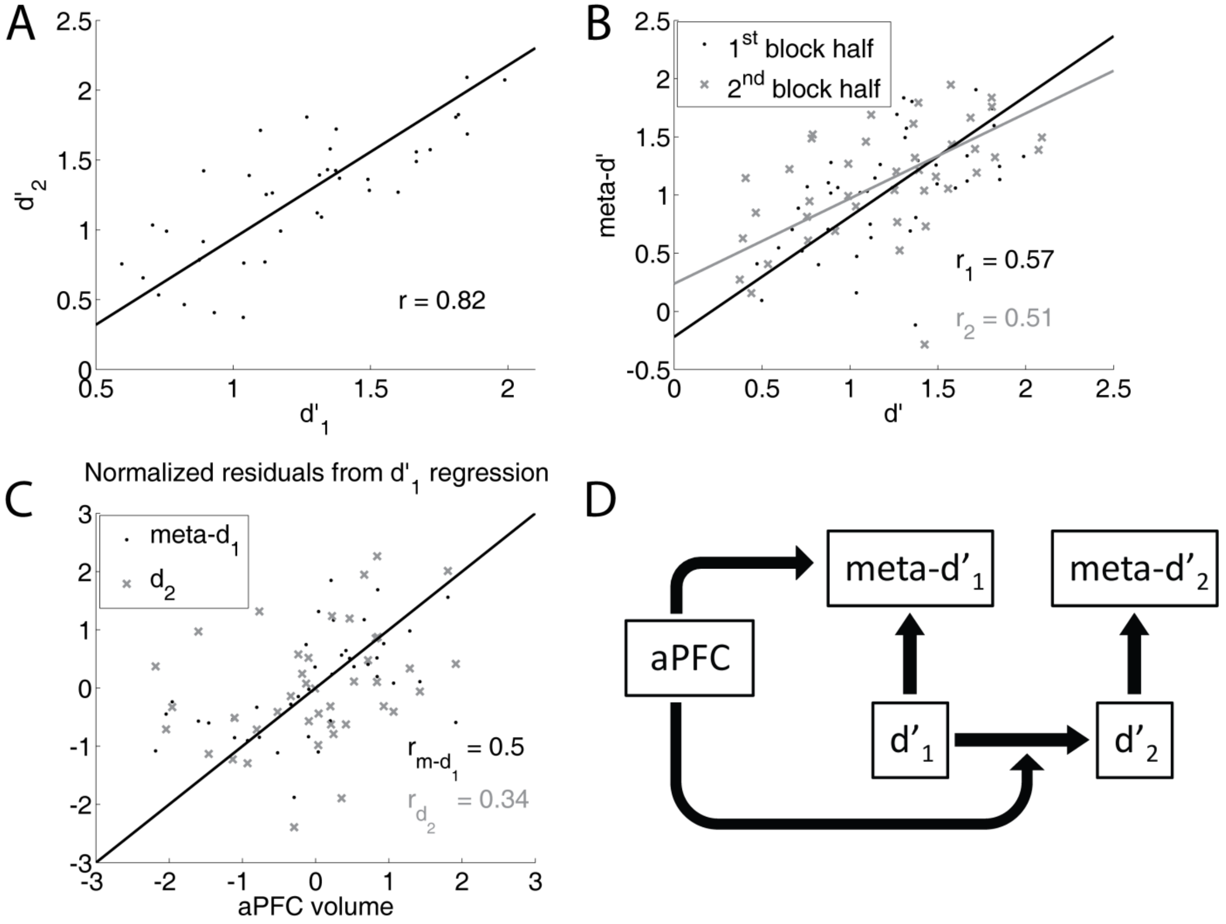

Figure 4-5. Model of the relationship between aPFC volume and changes in perceptual and metacognitive performance. Correlation analyses from Experiment 2 reveal significant positive correlations between **(A)** *d′* across block halves ($p < .001$); **(B)** meta-*d′* and *d′* within block halves (*p*s < .001); **(C)** aPFC volume with first-half meta-*d′* ($p = .001$) and second-half *d′* ($p = .03$), after removing variation due to first-half *d′*. (Lines of best fit for both correlations overlap.) **(D)** A schematic representation based on the correlations exhibited in panels A-C.

sensitivity only, the sign of the correlation between aPFC volume and Δmeta-*d′* was negative (though non-significant; $r = -.15$, $p = .3$).

In Figure 4-5 D, we present a simple schematic account to summarize these patterns of correlations. On this account, *d′* in the 2nd block half depends heavily on initial *d′*, and meta-*d′* in each block half is largely a consequence of *d′*. Without further components, this account would be consistent with SDT expectation. However, there is an additional component corresponding to aPFC volume, and this factor contributes both to better initial metacognition and to better maintenance of perceptual performance over time. Larger aPFC is associated with larger meta-d'_1 and therefore with smaller Δmeta-*d′*. Larger aPFC is also associated with larger Δ*d′*. Since larger aPFC is associated with positive values for Δ*d′* and negative values for Δmeta-*d′*, the contributions of aPFC appear to drive the deviation from SDT expectation encapsulated in the tradeoff relationship between Δ*d′* and Δmeta-*d′*. (See also "SDT simulations better characterize the data when taking into account the aPFC model" below.)

On this account, aPFC could be considered as a flexible cognitive resource that can contribute to both metacognitive monitoring and top-down control of perceptual task performance. To provide an additional test of this account, in Experiments 3 and 4 we included conditions where subjects did not have to provide metacognitive judgments in the first half of some experimental blocks. On this "resource" account, we might expect that when subjects do not have the initial cognitive burden of placing metacognitive judgments, the resources shared by perceptual and metacognitive processes can be better applied to the task of maintaining perceptual vigilance.

Experiments 3 and 4

In Experiment 3, we used a design similar to Experiment 1, with the primary difference that in even-numbered blocks, subjects were not asked to provide confidence ratings in the first half of each block (Figure 4-1 B). We shall call these blocks "partial type 2 blocks," as opposed to the blocks in which metacognitive judgments are required on every trial, which we shall call "whole type 2 blocks." According to a resource interpretation of the aPFC schematic (Figure 4-5 D), in the absence of the need to "boost" metacognitive performance, subjects should be better at maintaining perceptual performance over time in partial than in whole type 2 blocks (Figure 4-6 D). Experiment 4 was similar to

Experiment 3, but used a point-wagering system with feedback on perceptual and metacognitive after each block (Figure 4-1 C). In both experiments, trial length was a constant 2.533 s, yielding blocks of a 253.3 s duration.

We verified that, as in Experiments 1 and 2, Δ*d′* and Δmeta-*d′* in whole type 2 blocks exhibited a tradeoff in violation of SDT expectation (Figure 4-6 C). The Pearson's *r* correlation coefficients for Δ*d′* and Δmeta-*d′* were -.22 and -.08 in Experiment 3 and 4, respectively, whereas the SDT-expected value for *r*, according to computational simulations, was .41 (see Figure 4-7 C and "Monte Carlo SDT simulations" in Supplemental Information). Under the null hypothesis that changes in *d′* and meta-*d′* are generated by an SDT process with an expected *r* = .41, we estimate that the empirically observed values of *r* = -.22 and *r* = -.08 correspond to one-tailed *p*-values of 0.001 and .004 in Experiment 3 and 4, respectively (Figure 4-7 C). The Deming regression slope relating Δ*d′* and Δmeta-*d′* were -6.22 and -2.29, lower than the SDT-expected value of 1 (see "Regression of Δmeta-*d′* onto Δ*d′*" in Supplemental Information).

Next, we tested whether the manipulation on task demand yielded the expected effect on perceptual performance over time. A 2 (Block Type: partial type 2, whole type 2) x 2 (Time: 1st block half, 2nd block half) x 2 (Experiment: 3, 4) mixed design ANOVA on *d′* revealed a significant Block Type x Time interaction (*p* = .002). The interaction is driven by the fact that Δ*d′* is smaller for whole type 2 blocks (mean = -.20) than for partial type 2 blocks (mean = .04) (Figure 4-6 A, B).

The Block Type x Time x Experiment interaction was not significant (*p* = .4), suggesting that the difference in Δ*d′* for whole and partial type 2 blocks is robust across Experiment 3 (where participants made metacognitive judgments by rating confidence) and Experiment 4 (where participants made metacognitive judgments by wagering points, were instructed to maximize points earned, and received performance feedback after each block). Thus, the observed decrement in perceptual performance is

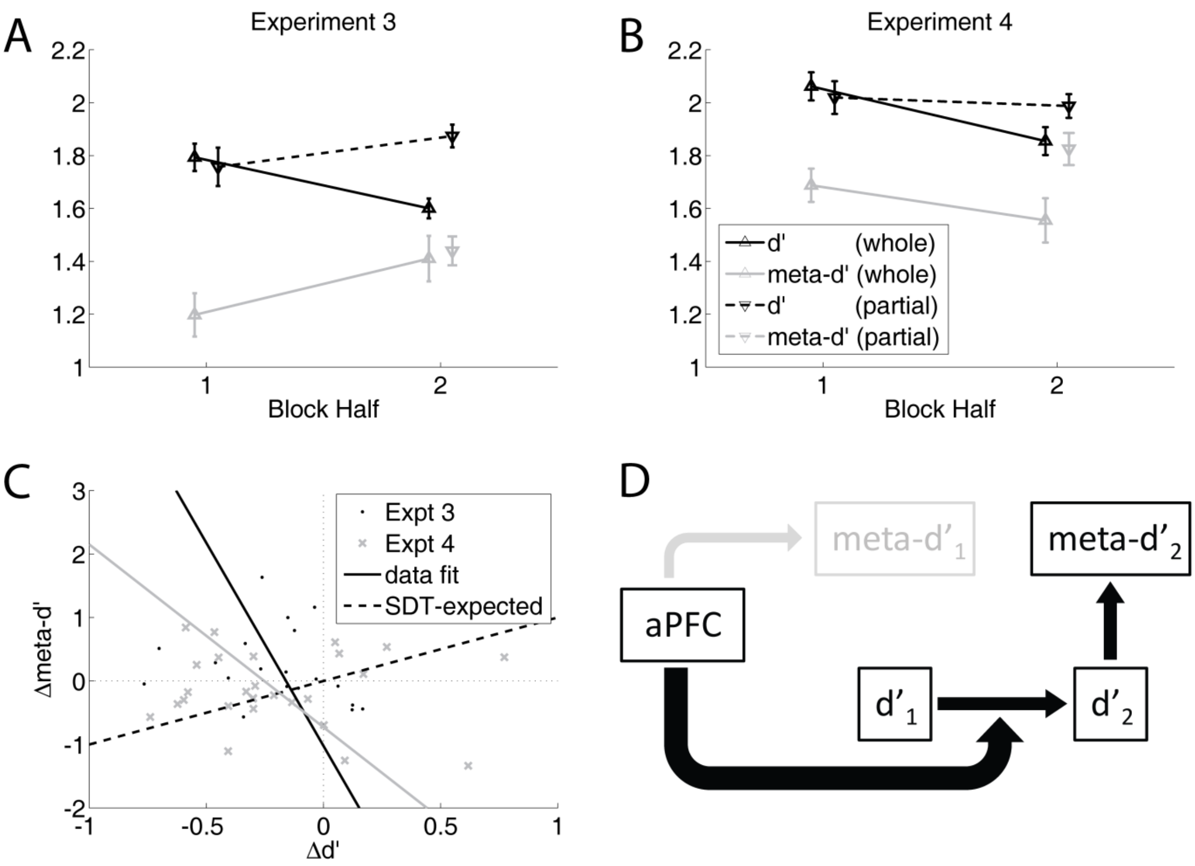

Figure 4-6. Results for Experiments 3 and 4. (A, B) Mean perceptual (*d'*) and metacognitive (meta-*d'*) performance over time. When subjects were not required to place metacognitive judgments in the first block half (partial type 2 blocks), perceptual vigilance increased (Block Type x Time interaction, *p* = .002) but metacognition in the second block half, as measured by meta-d'_2 / d'_2, was not affected (Block type, *p* > .4). Error bars represent within-subjects standard errors (Morey, 2008). (**C) Between-subject correlation of changes in perceptual and metacognitive performance.** As in Experiments 1 and 2, the between-subject relationships between changes in *d'* and meta-*d'* were substantially lower than SDT expectation. **(D) Resource account of findings.** The results of Experiments 3 and 4 can be understood in terms of the model derived from Experiment 2. By relieving subjects of the requirement to place metacognitive judgments in the first block half, aPFC resources normally dedicated to initial metacognitive performance may have been spared for the separate task of maintaining perceptual vigilance.

not attributable to lack of motivation or lack of a clear objective for how to perform the metacognitive task.

A 2 (Block Type: whole, partial) x 2 (Experiment: 3, 4) ANOVA yielded a nearly significant main effect of Block Type on meta-d'_2 (p = .053), such that meta-d'_2 was higher for partial type 2 blocks. However, the same ANOVA design shows that d'_2 was also higher for partial type 2 blocks (p < .001), and so the larger value for meta-d'_2 in partial type 2 blocks was likely mediated by the larger d'_2 value. Indeed, the same ANOVA analysis, when applied to the ratio meta-d'_2 / d'_2, did not reveal a main effect of Block Type (p > .4). Thus, the experimental manipulation on initial metacognitive demand did not influence metacognitive sensitivity in the second block half.

SDT simulations better characterize the data when taking into account the influence of aPFC

Finally, we performed Monte Carlo SDT simulations in order to computationally assess the empirical results in light of SDT expectation, and to investigate whether the SDT model could yield a closer fit to the empirical data when taking into account the relationship between aPFC volume and task performance (Figure 4-5 D). See “Monte Carlo SDT simulations” in Methods for full methods.

For each simulated subject, we defined the parameters of an SDT model specifying performance in the first and second block half of a binary decision task with confidence ratings. SDT model parameters were sampled from distributions closely reflecting the statistical patterns in Experiments 1 – 4. Random samples were then drawn from the SDT models in order to generate a simulated value for Δd' and Δmeta-d'. In all, we simulated 2000 experiments, each containing 30 simulated subjects. Consistent with SDT expectation, these simulations yielded a strong positive correlation between Δd' and Δmeta-d' (Figure 4-7 A). Next, we adjusted the initial simulation values for meta-d'_1 on the basis of regression-estimated relationships between Δd', meta-d'_1, and aPFC volume in Experiment 2. This

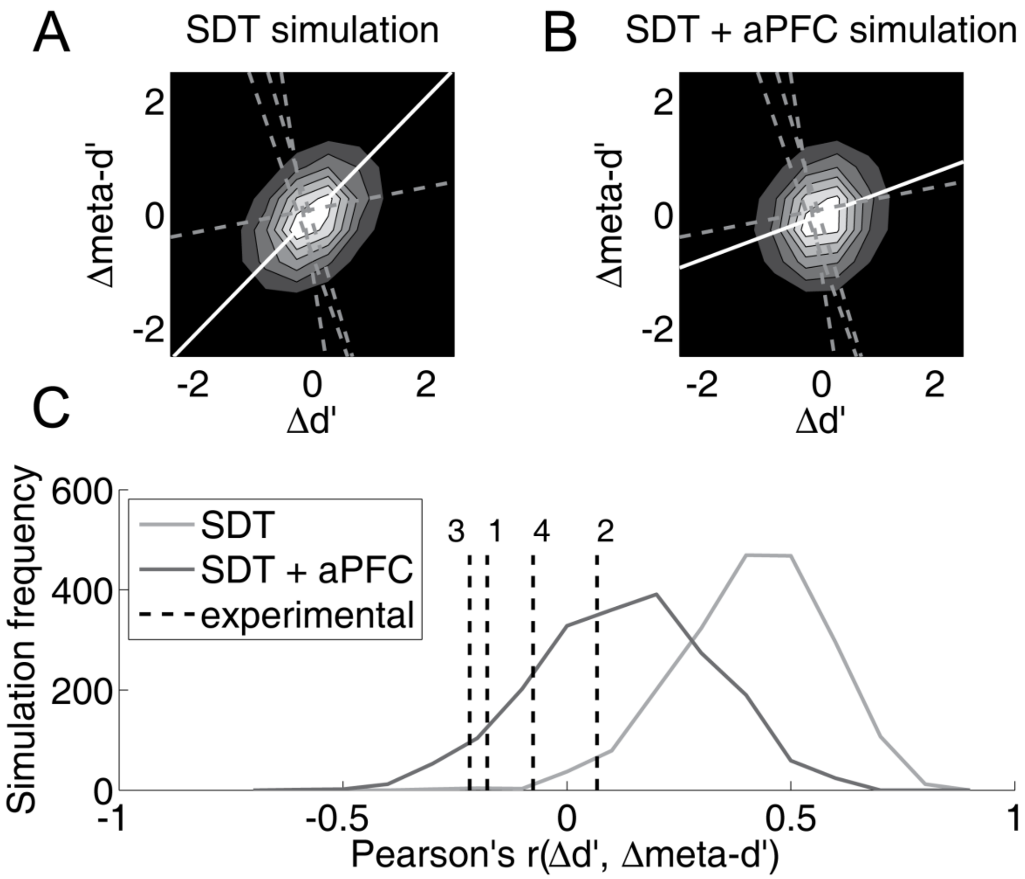

Figure 4-7. Signal detection theory simulations of the relationship between changes in perceptual and metacognitive sensitivity. (A) Basic SDT model. In a series of SDT simulations closely matching the properties of Experiments 1 – 4, changes in *d′* and meta-*d′* across block half are strongly positively related. Displayed is a contour plot based on the two-dimensional histogram of Δmeta-*d′* vs Δ*d′* for all simulated subjects in all simulated experiments. White line is line of best fit to simulated data; gray dashed lines are lines of best fit from data in Experiments 1 – 4. **(B) SDT model with aPFC adjustment.** We adjusted the outcomes of the initial SDT simulation so as to conform to the empirically observed relationships between aPFC volume, Δ*d′*, and meta-d'_1 / d'_1 in Experiment 2 (see Methods for details). This substantially weakened the relationship between Δ*d′* and Δmeta-*d′* in the simulated data, as demonstrated by a more circular contour plot and smaller slope for the line of best fit. **(C) Distributions of correlation coefficients for Δ*d′* and Δmeta-*d′*.** Across 2000 simulated experiments, the basic SDT model yielded correlation values consistently higher than those observed in Experiments 1 – 4 (one-tailed *p*-values: 0.002, 0.026, 0.001, 0.004). The adjusted model incorporating the aPFC findings from Experiment 2 yielded a distribution of correlations more closely in line with the data (one-tailed *p*-values: 0.067, 0.383, 0.043, 0.161).

adjustment significantly attenuated the positive correlation between simulated values for Δ*d'* and Δmeta-*d'* (Figure 4-7 B).

For each simulated experiment, we computed the Pearson's *r* correlation for Δ*d'* and Δmeta-*d'*, yielding 2000 *r* values. The distribution of simulated *r* values under the "SDT" and "SDT + aPFC" models is displayed in Figure 4-7 C alongside the empirically observed *r* values from Experiments 1 – 4. Under the "SDT" model, the distribution's mean value is 0.412 and only 0.9% of all values are lower than zero. Under the "SDT + aPFC" model, the mean shifts to 0.127 and 25.7% of all values are lower than zero, which is in better agreement with the data. For each empirical *r* value, we can compute a corresponding one-tailed *p*-value using the *r* distribution for the "SDT" and "SDT + aPFC" models. The empirical *r* values from Experiments 1 – 4 are -0.18, 0.07, -0.22, and -.08. Under the "SDT" model, these correspond to *p*-values of 0.002, 0.026, 0.001, and 0.004. Under the "SDT + aPFC" model, these *p*-values increase on average by a factor of about 30, to 0.067, 0.383, 0.043, and 0.161. Thus, the "SDT + aPFC" model is considerably better in accommodating the observed patterns of correlation between Δ*d'* and Δmeta-*d'* than is the standard SDT model.

Discussion

In summary, across four experiments, we find a robust tradeoff effect whereby changes in perceptual and metacognitive sensitivity within a block of trials are negatively or weakly correlated, contradicting the strong positive relationship predicted by single-process signal detection theory (SDT). Voxel-based morphometry analysis suggests that this tradeoff effect may be explained by the contribution of neural resources in anterior prefrontal cortex (aPFC). Consistent with this account, perceptual vigilance decrements are alleviated when subjects are not required to provide metacognitive judgments in the first half of a block of trials.

Tradeoff relationship between perceptual and metacognitive vigilance

The classical SDT model, which has enjoyed considerable success in modeling two-choice decision paradigms with confidence ratings (Swets, 1986a; Macmillan & Creelman, 2005), predicts a strong, positive relationship between primary task performance and metacognitive performance (Galvin et al., 2003; Maniscalco & Lau, 2012). In agreement with this prediction, we found that overall *d′* correlated positively with overall meta-*d′* in all four experiments, and also that *d′* correlated positively with meta-*d′* in seven out of eight block halves in the four experiments. Thus, when considering the Pearson's correlation between Δ*d′* and Δmeta-*d′*, we used SDT as the null hypothesis describing the expected distribution of correlation coefficients. We used Monte Carlo SDT simulations to construct the SDT-expected distribution of *r* values for Δ*d′* and Δmeta-*d′*, which yielded a distribution with a mean *r* value of .41 (see Figure 4-7 C and "Monte Carlo SDT simulations" in Methods).

We found that, although *d′* and meta-*d′* were robustly positively correlated overall and *within* block halves, nonetheless, *changes* in these measures *across* block half failed to exhibit positive correlations, contradicting SDT expectation. Importantly, although the correlation coefficients for Δ*d′* and Δmeta-*d′* were small in magnitude, the relevant point of comparison is not with a distribution whose mean $r = 0$, but rather with the SDT distribution whose mean $r > 0$. The correlations in Experiments 1 – 4 significantly deviated from this SDT expectation. Thus, relative to the SDT-expected positive relationship, perceptual and metacognitive vigilance appeared to "trade off," such that improvement in one precluded comparable improvement in the other.

Interpreting the tradeoff relationship

Research on the perceptual vigilance decrement has suggested that the decrement is caused by the depletion of limited cognitive resources (e.g. Grier et al., 2003; Helton & Warm, 2008; Helton et al., 2005; Warm et al., 2008). Experiment 2 of the present study suggests that regions of aPFC whose anatomical structure has previously been associated with metacognitive sensitivity in visual tasks (Fleming et al., 2010; McCurdy et al., 2013) may partially instantiate the resources supporting perceptual vigilance, since larger gray matter volume in these regions is associated with smaller declines in perceptual sensitivity.

The gray matter volume of aPFC was also associated with better metacognitive sensitivity during the first, but not second, half of each block (Figure 4-4 C-D; Figure 4-5 C). This may have driven a negative relationship between aPFC volume and Δmeta-*d′* in two ways. First, higher values for meta-d'_1 would directly lead to lower values for Δmeta-*d′*. Second, according to SDT, meta-*d′* is theoretically constrained to be less than or equal to *d′* (Maniscalco & Lau, 2012). Therefore, all else being equal, better meta-d'_1 leaves less room for meta-d'_2 to improve, entailing a smaller maximum possible value for Δmeta-*d′*.

Thus, aPFC simultaneously exhibited a positive association with Δ*d′* and a negative association with Δmeta-*d′*. Subjects with larger aPFC exhibited strong perceptual vigilance (higher Δ*d′*) as well as SDT-ideal metacognitive performance (meta-*d′* = *d′*; Figure 4-4 D). Conversely, subjects with smaller aPFC exhibited poorer perceptual vigilance (lower Δ*d′*) and poorer initial metacognition (contributing to higher Δmeta-*d′*; Figure 4-4 C). In this way, individual differences in aPFC volume could produce the tradeoff effect whereby Δ*d′* and Δmeta-*d′* failed to positively correlate (Figure 4-2 B, 4-4 B, 4-6 C).

One way of interpreting these findings is that perception and metacognition are subserved by separate processes that can independently tap into a common cognitive resource housed in aPFC. Presumably, as a block of trials wears on, resources would be increasingly allocated to the perceptual process (and thus away from the metacognitive process) in order to counteract the perceptual vigilance decrement (Figure 4-5 D).

This account views perception and metacognition as separate processes that can draw upon a common set of limited cognitive resources in a flexible manner, creating the potential for interference and competition for resources when both tasks are performed concurrently (Kahneman, 1973; Matthews et al., 2000; Wickens, 2002). More generally, this interpretation is consistent with accounts ascribing a broadly domain-general functionality to prefrontal cortex in guiding behavior (e.g. Koechlin & Summerfield, 2007; Badre, 2008; Passingham & Wise, 2012).

An alternative account is that since larger aPFC is associated with superior visual metacognition, the positive association between aPFC volume and perceptual vigilance could be mediated by superior metacognitive monitoring. Higher metacognitive sensitivity entails better ability to gauge ongoing perceptual performance, which could enable better ongoing regulation of task performance. On this account, aPFC is not a domain-general resource, but rather serves a specifically metacognitive function.

However, if better metacognitive monitoring directly contributes to superior perceptual vigilance, we might expect that perceptual vigilance should decrease when subjects are not required to engage in metacognitive monitoring. The resource account makes the opposite prediction; relieving the burden of placing confidence ratings should free up resources to support perceptual vigilance. In Experiments 3 and 4, we found that subjects were indeed more perceptually vigilant when not required to place confidence ratings in the first half of a block, more in line with the resource account than the metacognitive monitoring account. However, we take this result to be suggestive rather than decisive. Ultimately, these hypotheses will need to be further explored in future research.

Implications for models of metacognition

An active area of research concerns the relationship between perceptual and metacognitive processing. According to some accounts, seemingly complex and high-level functions such as metacognition and awareness actually bear simple and direct relationships to basic perceptual processing (Kepecs et al., 2008; Kiani and Shadlen, 2009; Kepecs and Mainen, 2012). The intuition behind these models is captured well by the conventional SDT model of confidence ratings, which characterizes perceptual judgments and confidence ratings as originating from the comparison of the same sensory information to different decision criteria (Macmillan & Creelman, 2005; Appendix A). Crucially, if perceptual and metacognitive judgments are different evaluations of the same underlying sensory information, then they should have similar informational content (formally, *d′* = meta-*d′;* Galvin et al., 2003; Maniscalco & Lau, 2012).

However, whereas the SDT model predicts a strong positive relationship between perceptual and metacognitive vigilance, we consistently observed this relationship to be neutral or negative. In our SDT-based simulations, we found that the empirical correlations between Δ*d′* and Δmeta-*d′* could not plausibly be accounted for by sampling variation under the SDT model (Figure 4-7 A, C). However, adjusting the simulation outcomes to reflect the mediating effect of aPFC volume on the behavioral measures entailed a theoretical outcome more in line with the data (Figure 4-7 B, C). In turn, the fact that aPFC volume had an opposite direction of association with Δ*d′* and Δmeta-*d′* suggests that perception and metacognition are separate processes with dissociable levels of sensitivity.

Why do we give subjects short breaks in perceptual experiments?

Though originally found in the context of long task durations (30+ min), the vigilance decrement has been shown to arise as early as the first 5 – 10 minutes of task performance (e.g. Nuechterlein, Parasuraman, & Jiang, 1983; Temple et al., 2000), dependent on factors such as overall perceptual sensitivity, rate of stimulus presentation, type of stimuli used, and memory load (See et al., 1995). Vigilance decrements are further associated with subjective effects such as reduced arousal and elevated feelings of stress (Helton & Warm, 2008; Warm et al., 2008). Thus, a wide range of experimental tasks may be subjectively fatiguing and induce relatively rapid decrements in task performance.

In the current work, we found that perceptual (Experiments 3 – 4) and metacognitive (Experiment 1) vigilance decrements can occur even in experimental blocks of approximately 4 – 5 minutes in a fairly simple and standard visual discrimination task. Because we analyzed performance as a function of time across repeated blocks of trials, rather than analyzing the dynamics of task performance across a single prolonged block of trials, these results suggest a systematic pattern of performance decrements occurring within repeated blocks of trials that are nonetheless alleviated by regular intervals of rest.

What cognitive mechanisms benefit from the regular intervals of rest commonly used in perceptual experiments? The tradeoff between perceptual and metacognitive vigilance found in Experiments 1 – 4, and the elevation of perceptual vigilance solely by relieving metacognitive task demand in Experiments 3 – 4, suggest the workings of a higher-level cognitive resource. The results of Experiment 2 identify aPFC as a contributor to this resource. Thus, our results suggest that rest primarily refreshes high-level cognitive resources, located at least partially in aPFC, rather than lower-level sensory mechanisms.

General Discussion

Summary of findings

The main thrust of this dissertation is to probe the relationship between objective and subjective performance in visual tasks in order to make inferences about the overall structure of the processes that underlie objective and subjective vision. We have demonstrated dissociations between objective and subjective aspects of visual processing caused by a variety of factors: stimulus properties (Chapter 1), dual task demands (Chapter 2), direct interference with brain activity (Chapter 3), and naturally occurring changes in performance over time (Chapter 4). In each case, the existence of the observed dissociation appears to be attributable to properties of the prefrontal cortex. Taken together, these findings are difficult to reconcile for some currently popular theories of subjective visual experience and metacognition, but are readily accounted for by higher-order, hierarchical models of the nature of subjective visual perception.

In Chapter 1, we replicated the finding of Lau and Passingham (2006) that objective stimulus discrimination performance and subjective reports of perceptual clarity in the metacontrast masking paradigm can dissociate. Although both objective and subjective measures are U-shaped functions of the target-mask stimulus onset asynchrony (SOA), these functions are asymmetric, such that there exist certain SOAs for which objective performance is equivalent but subjective reports of perceptual clarity differ. A formal comparison of a wide array of signal detection (SDT) models implementing Single Channel, Dual Channel, and Hierarchical structures suggests that the data are best accounted for by a Hierarchical model in which there exists an early, objective processing stage and a late, subjective processing stage. Processing in the early stage determines objective stimulus discrimination performance. Processing in the late stage, which generates a subjective report of perceptual clarity, inherits a somewhat weaker and noisier version of the information used to make the objective stimulus judgment. According to the parameter values of the best-fitting Hierarchical model, short and long SOAs can elicit the same objective discrimination performance and different reports of subjective visibility because although early, objective perceptual processing is similar at both SOAs, the sensory signal is better transmitted to the later, subjective processing stage at the long SOA than at the short SOA. This characterization is consistent with the finding of Lau and Passingham (2006) that higher visibility at the long SOA is associated with heightened activation of dorsolateral prefrontal cortex (dlPFC).

Comparison of the Single Channel, Dual Channel, and Hierarchical model structures in Figure 1-1 suggests a general way in which the models can be distinguished. According to the Hierarchical model, interference with sensory representations at late-stage subjective processing should impair visual metacognition while leaving objective processing intact. By way of contrast, the Single Channel and Dual Channel models both suppose that differences in processing that manifest as differences in metacognitive sensitivity should also manifest as differences in perceptual sensitivity. Thus, further support for the Hierarchical model can be provided by demonstrating interventions that selectively impair metacognitive sensitivity by targeting late stage neural processing. We provided two such demonstrations in Chapters 2 and 3.

In Chapter 2, we investigated objective and subjective visual processing in the context of a concurrent working memory (WM) task. Specifically, we assessed the impact of maintenance and manipulation of WM contents on perceptual and metacognitive sensitivity. We found that overall, increasing the burden of WM maintenance impairs both objective and subjective visual performance. However, increasing the burden of active manipulation of WM contents selectively impaired metacognitive sensitivity. Since active manipulation of WM contents has been previously associated with dlPFC function, and since dlPFC and nearby regions of PFC have been previously associated with visual metacogniton, these findings suggest that a common mechanism in dlPFC may contribute to both manipulation of WM contents and metacognitive evaluation of objective perceptual performance.

In Chapter 3, we investigated the effect of transcranial magnetic stimulation (TMS) to bilateral dlPFC. We found that this intervention selectively impairs metacognitive sensitivity while leaving basic perceptual performance intact. This result more directly demonstrates the connection between dlPFC function and metacognitive sensitivity that was suggested by the findings in Chapter 2.

We note also that Chapter 2 and 3 showed similar results even though the study in Chapter 2 made use of confidence ratings entered after the objective stimulus discrimination response, whereas the study in Chapter 3 made use of perceptual clarity ratings entered simultaneously with the stimulus discrimination response. These similarities further corroborate the close relationship between clarity and confidence ratings discussed in the General introduction, and to some extent help mitigate concerns over the potential impact of simultaneous or sequential entry of objective and subjective perceptual responses on behavioral outcomes.

In Chapter 4, we find a naturally occurring dissociation between perceptual and metacognitive sensitivity. As time progresses within a continuous block of trials, perceptual and metacognitive sensitivity change, but these changes do not correlate with each other across subjects. This failure to correlate is surprising, given the robust theoretical expectation (Galvin et al., 2003) and empirical evidence (Maniscalco & Lau, 2012) that perceptual and metacognitive sensitivity should positively correlate. Indeed, perceptual and metacognitive sensitivity *do* positively correlate, as expected, *within* each block half. It is only the *change* between the two across block halves that surprisingly fails to correlate. We found that between-subject variation in the relationship between perceptual and metacognitive vigilance was associated with gray matter volume in regions of the anterior prefrontal cortex (aPFC) that correlate with overall visual metacognition (Fleming et al, 2010; McCurdy et al., 2013). We took these findings as suggestive of a *tradeoff* between perceptual and metacognitive vigilance that is mediated by a common resource in aPFC. This resource account was corroborated by findings that, when subjects did not have to rate confidence in the first half of a block of trials, perceptual vigilance improved.

Relations and contributions to ongoing research

As reviewed in the General Introduction, there currently exist a rather wide range of views on the cognitive and neural bases of visual awareness and metacognition. Lau & Rosenthal (2011) categorized views of visual awareness into first order-views, neuronal global workspace theory, information integration theory, and higher-order views. With respect to characterizing the relationship between objective and subjective visual processing, first-order views and neuronal global workspace theory are alike in that they assume a direct and intimate relationship between objective and subjective vision. Thus, in this respect, both first-order theories and neuronal global workspace theory can be considered to have a Single Channel model structure (Figure 1-1). A revision to neuronal global workspace theory that includes a separate processing channel for unconscious processing has recently been proposed (Del Cul et al., 2009), thus taking on a Dual Channel structure (Figure 1-1). Higher-order theories of visual awareness closely map onto the Hierarchical model structure supported in Chapter 1.

In the domain of studying perceptual confidence judgments, there is a similar bifurcation between views that suppose a direct and intimate relationship between objective stimulus processing and ratings of confidence (Kepecs et al., 2008; Kiani & Shadlen, 2009; Kepecs & Mainen, 2012), and views that suppose that confidence ratings are driven by higher-order mechanisms evaluating lower-level perceptual processing (Fleming et al., 2010; Pleskac & Busemeyer, 2010; McCurdy et al., 2013). Here, too, the former kind of view resembles a Single Channel structure and the latter, a Hierarchical structure.

We have presented evidence favoring the Hierarchical view, but this is not the first demonstration of evidence consistent with such a view. So what novel perspectives, methods, points of emphasis, and empirical findings have we introduced in order to advance the discussion?

A signal detection theoretic framework for analyzing and interpreting metacognitive sensitivity

How should we analyze metacognitive performance? A straightforward approach is to simply compute averages for subjective ratings of perceptual clarity or confidence in different experimental conditions. However, raw subjective rating data is determined in large part by an observer's idiosyncratic response biases, and it cannot tell us *how well* the observer is able to distinguish between his own correct and incorrect perceptual responses. Simple measures of metacognitive accuracy, such as a correlation between accuracy and confidence (Kornell et al., 2007), can be computed, but such measures are in danger of confounding sensitivity and response bias—in general, a measure does not purely capture sensitivity unless it can provide a satisfactory fit to Receiver Operating Characteristic (ROC) curves (Swets, 1986a). Previous attempts to use SDT to measure metacognitive sensitivity (Kunimoto et al., 2001) have been empirically shown to be inadequate (Evans & Azzopardi, 2007) due to not properly contextualizing the analysis of subjective ratings into the structure of traditional SDT (Galvin et al., 2003).

Our methodology (Maniscalco & Lau, 2012; Appendix A) builds on the theoretical analysis of Galvin et al and provides a straightforward way to measure metacognitive sensitivity, independently from response bias. Furthermore, this SDT framework allows us to meaningfully interpret the *actual* level of metacognitive sensitivity (meta-*d'*) by way of comparison to the SDT-*expected* level of metacognitive sensitivity (equivalent to *d'*). Thus, the framework is very useful for measuring and comparing perceptual and metacognitive sensitivity, and the comparisons between the two can be quite informative and useful for hypothesis testing. This methodological advance has formed the backbone of the current work and can be fruitfully put to use in future studies.

Isolating subjective measures from performance confounds

Most studies on objective and subjective visual processing to date have not properly controlled for aspects of objective perceptual processing when studying awareness and metacognition. For instance, a common approach is to create experimental conditions where the stimuli are similar or identical and yet subjective reports differ, and investigate differences in behavior, cognition, and brain activity (e.g. Dehaene et al, 2001). A related approach is to compare trials where a subject successfully reports the presence of a stimulus ('hits') to trials where the stimulus is presented but the subject reports not having seen it ('misses') (e.g. Lamy, Salti, & Bar-Haim, 2009), or to compare trials where reports of confidence are high against trials where reports of confidence are low (e.g. Kiani and Shadlen, 2009).

The problem with approaches such as these is that the comparison groups typically do not differ only with respect to subjective reports, but also with respect to objective perceptual processing capacity. Experimental conditions where a stimulus is clearly visible, all else being equal, tend to also be associated with better, more accurate objective processing of the stimulus. Even the approach of comparing aware vs unaware trials, or high vs low confidence trials, suffers a similar problem. Such an approach implicitly assume a so-called threshold model of perception, according to which all 'aware' trials are essentially identical manifestations of the same underlying perceptual state, but this model contradicts signal detection theory and provides a poor fit to empirical ROC curves (Swets 1986b). On the SDT model, the objective perceptual signal associated with hits and high confidence responses is variable, but is higher on average than it is for misses and low confidence responses. Thus, these single trial comparisons do not succeed in avoiding performance confounds (Lau, 2008b).

Our approach is careful to isolate subjective measures from associated performance confounds. In Chapter 1, we capitalized on an experimental paradigm which dissociates objective and subjective measures, including conditions in which the ability to objectively discriminate the target is equivalent, but average reports of perceptual clarity differ. In Chapters 3 – 4, we use the theoretical SDT framework to measure relative metacognitive sensitivity, i.e. the observed value of metacognitive sensitivity in comparison with the SDT-expected value based on objective task performance. This computational approach allows us to theoretically isolate aspects of metacognitive processing in and of themselves, even when perceptual sensitivity is not equated by experimental means (Maniscalco & Lau, 2012; Appendix A).

Thus, our experimental and methodological approach has allowed us to handle the problem of performance confounds and make inferences relating specifically to the subjective aspects of visual processing.

Evidence for a causal role of dlPFC in visual metacognitive sensitivity

While prior studies have suggested a link between PFC and ratings of visual clarity and confidence (e.g. Lau & Passingham, 2006; Fleming et al., 2010; Fleming et al, 2012), it has not been as well established in the literature that PFC plays a causal role in visual metacognition. A notable recent study in this regard is Del Cul et al. (2009), which found that subjects with prefrontal lesions have similar objective visual performance as healthy controls for trials in which they report seeing or not seeing the stimulus, and yet have lower mean levels of reported visibility for correct and incorrect trials. However, patients also performed worse on the task overall (posing a potential performance confound), and Del Cul et al. did not explicitly analyze metacognitive sensitivity.

In Chapters 2 and 3, we provided converging evidence that dlPFC plays a causal role in relative metacognitive sensitivity. We showed that cognitive and neural interference with this region can selectively impair metacognitive performance, even when variability in objective task performance is taken into account mathematically in the SDT framework. This suggests the relationship between the anatomy and activity of PFC and subjective visual processing is not epiphenomenal, but rather that PFC plays a causal role in determining reports of visual clarity and confidence. Importantly, the effects of the experimental manipulations reported in Chapters 2 and 3 cannot be attributed to mere changes in response bias for the subjective rating task, since we demonstrated that these manipulations affect metacognitive sensitivity—the informational capacity with which the subjective ratings distinguish between correct and incorrect responses, regardless of response biases in the subjective rating.

Competition between perceptual and metacognitive processes

The evidence presented in Chapter 4 suggests that perceptual and metacognitive processes are not only separate, but may even compete for limited resources. In turn, this suggests that the common experimental scenario of performing a perceptual task and then providing a subjective rating may constitute a kind of dual-task scenario, with the potential for the two processes to interfe or compete with each other. Along these lines, Petrusic and Baranski (2003) compared performance on a perceptual task in two conditions in which confidence ratings were and were not required. They found that when subjects had to provide confidence ratings, reaction times for the primary perceptual task became longer, although there was no apparent effect on accuracy in the perceptual task. An interesting question for future research is whether tradeoffs between perceptual and metacognitive processing can be controlled by experimental manipulations or task instruction in a way analogous to the well-known speed-accuracy tradeoff.

If perceptual and metacognitive processes can compete for prefrontal resources, this also raises the possibility that certain contributions of the PFC to metacognition may reflect domain general functions that happen to be commonly recruited for metacognitive evaluation, rather than being functions that are specifically dedicated to metacognition per se (Koechlin & Summerfield, 2007; Badre, 2008; Passingham & Wise, 2012).

Addressing a common critique from first-order theorists

Proponents of first-order views hold that visual awareness and/or metacognition occurs in early sensory processing regions, rather than at later, higher-order stages. A common critique of first-order theorists is that higher-order processing, as expressed e.g. in prefrontal and frontoparietal networks, may reflect cognitive functions such as attention, memory, language, mechanisms for generating perceptual reports, and so on (Lamme, 2006). In terms of neural structure and function, such mechanisms could conceivably be downstream, higher-level functions that are reliably co-activated with aspects of subjective visual processing and yet do not constitute the core features of visual awareness, clarity, or confidence per se (Pins & Ffytche, 2003). In terms of cognition and behavior, it may be difficult to disentangle such secondary functions from subjective visual processing since these functions, especially the function of generating introspective reports, are typically the very means by which we access and measure subjective vision. This is particularly so in the case of visual awareness and ratings of visual clarity, as subjective reports on these processes constitute direct commentaries on the nature of the observer’s private and otherwise inaccessible subjective experience.

However, all of our findings have pertained to selective effects on ratings of perceptual clarity (Chapter 1) and metacognitive sensitivity (Chapters 1 – 4), controlling for differences in objective perceptual performance experimentally (Chapter 1) or mathematically (Chapters 2 – 4). Thus, to the degree that the current findings implicate PFC in subjective visual processing, the specific role played by PFC cannot be attributed to broad mechanisms such as attention or perceptual decision making. If PFC’s effect on visual metacognition were mediated by general purpose attentional or perceptual decision making mechanisms, such effects would presumably also be made apparent in objective perceptual performance. Thus, the fact that PFC's involvement was specific to metacognition at the very least implies the existence of separate underlying mechanisms for objective and subjective visual processing.

Furthermore, the findings that disruption of PFC impairs metacognitive, but not perceptual sensitivity, suggests that PFC not only has specific associations with subjective vision, but that this relationship is causal in nature. These findings are inconsistent with the interpretation that PFC reflects a downstream process that follows as a typical secondary result of subjective visual processing without directly contributing to such processing. Instead, PFC function appears to be necessary for the optimal functioning of metacognitive evaluation.

A third observation is that in the TMS study reported in Chapter 3, the subjective rating was a rating of perceptual clarity rather than a confidence rating. This suggests that the role of PFC in subjective visual processing is not purely decisional, in the sense of making an abstract overall evaluation of objective visual processing, but rather that PFC contributes to the informational integrity carried by direct reports on the nature of ongoing visual experience. Of course, it is possible to object that PFC may only be contributing to the *report* of perceptual clarity, rather than to the phenomenological clarity of the visual experience itself. We acknowledge that this is a possibility, and take this observation regarding the visual phenomenology only to be suggestive. At the very least, however, the current results support the claim that the *report* on perceptual clarity, whatever its ultimate nature, is specific to subjective visual processing, rather than being subsumed into a more general process that includes the reporting of both objective and subjective perceptual decisions.

Relation of perceptual metacognition to PFC function

In Chapter 2, we showed that metacognitive sensitivity is impaired when subjects have to perform extensive operations on the contents of working memory. In turn, this suggests the possibility that a common cognitive mechanism, presumably housed in dlPFC, contributes to both activities.

However, working memory and metacognition are not typically thought of as being interrelated processes, to the extent that they draw upon a common, crucial, limited set of functions or resources.

In fact, recent proposals suggest that there may be no part of PFC that is specialized for the cognitive function of “working memory” as such. Instead, working memory may be an emergent phenomenon that arises from the flexible coordination of multiple cognitive/neural mechanisms that themselves are specialized for separate functions (Postle, 2006; D’Esposito, 2007). For instance, working memory could be characterized as the top-down attentional maintenance of sensory-, representation, or action-related functions (Postle, 2006).

In turn, it may similarly be the case that perceptual metacognition is not a unitary cognitive function carried out by a specialized part of the brain, but rather emerges as the result of the coordination of a set of more basic component perceptual/cognitive/neural processes. For instance, as discussed in the Discussion section of Chapter 2, if dlPFC houses a common mechanism that contributes to both manipulation of WM contents and perceptual metacognition, at least two possibilities for such a mechanism present themselves. One is that dlPFC may contribute to the re-organization and re-representation of sensory information, as captured by the phenomenon of “chunking” large amounts of information into more parsimonious, higher-level units of organization in memory tasks (Bor et al., 2003), and the idea that perceptual metacognition may involve the construction of meta-representations or “representational re-descriptions” of sensory information (Nelson & Narens, 1990; Schooler, 2002; Timmermans, Schilbach, Pasquali, & Cleeremans, 2012). Another possible common mechanism concerns the phenomenon of response selection in working memory (Curtis & D’Esposito, 2003), whereby a mapping from the current context to an appropriate motor response is computed, and the phenomenon of criterion setting in metacognition, whereby a mapping from perceptual information onto a categorical confidence response is computed (Green & Swets, 1966; Appendix A).

A similar consideration is suggested by the findings of Chapter 4, in which aPFC appears to contribute to the dynamics of both perceptual and metacognitive vigilance. The pattern of results in that series of studies suggested that while aPFC supports metacognitive sensitivity at the outset of a block of trials, as time wears on and perceptual sensitivity potentially declines due to vigilance decrement effects, aPFC shifts to supporting perceptual sensitivity and no longer supports metacognition. Here again is a case where a metacognitive process seems to share crucial yet limited resources with a seemingly unrelated cognitive function (in this case, perceptual vigilance). Perhaps, then, perceptual metacognition arises as an emergent phenomenon from the right coordination of a set of component neural processes, processes that are themselves not intrinsically "metacognitive" but rather can serve as essential components to a wide array of other complex functions that are similarly dynamic and emergent, such as working memory and perceptual vigilance.

However, from a broader point of view, the evaluative function of subjective judgments of perceptual clarity and confidence do seem to dovetail well with the overall function or purpose of prefrontal cortex, which can be characterized as the sophisticated control of behavior, taking into account context, contingencies, and goals at various levels of conceptual abstraction and temporal separation (Fuster, 2001; Koechlin & Summerfield, 2007; Badre, 2008; Passingham & Wise, 2012). Metacognitive judgments presumably participate in this process by modulating stimulus-response mappings according to confidence in the perceptual identification of the stimulus. For instance, the same perceptual identification of a message written on a sign in the distance can be occasion for slightly different behaviors depending on the level of confidence with which the perceptual identification is made. If one reads the message clearly and fluently, one might resume the previously ongoing behavior; however, had one been less confident in the perceptual identification of the message, one might have looked at the sign longer or approached it in order to get a better look, until metacognitive evaluation confirmed that the perceptual identification was indeed trustworthy and behavior could now be engaged in a new pursuit. Such a role for perceptual metacognition in the control of behavior would be in keeping with theories postulating an anterior-posterior organization of abstract-concrete motor planning functions (Koechlin & Summerfield, 2007; Badre, 2008), to the extent that metacognition is generally associated with more anterior regions of lateral PFC such as anterior, rostrolateral, and dorsolateral PFC and plausibly serves the general function of supporting relatively sophisticated and abstract modulations of sensorimotor processing.

Functions of subjective visual processing

We have argued that subjective visual processing is a kind of secondary, late stage process that supports explicitly represented *knowledge* about first-order perceptual processing, but does not directly participate in that first-order perceptual processing by means of which perceptual decisions about the state of the objective world are derived. On the way to arriving at this view, we have emphasized the need to treat objective perceptual performance as a confound in the study of subjective visual processing, and that only when such objective processing is “factored out” can we make inferences specific to subjective vision. This may give one the impression that there is not much work left for subjective visual processing to do in the cognitive economy. Indeed, many cognitive functions that once were thought to require conscious awareness, such as relatively complex cognitive control functions involved with task cuing (Lau & Passingham, 2006) and response inhibition (Van Gaal, Lamme, & Ridderinkhof, 2010), have been shown to be functional even when the relevant stimuli are not consciously perceived (Lau, 2009). In the face of such considerations, it may seem as if subjective visual processing is essentially epiphenomenal, and that in principle any function could potentially be carried out unconsciously.

However, even if subjective visual processing does not directly participate in the objective perceptual process, and even if certain cognitive control functions can be performed in the absence of awareness of the stimuli triggering the relevant control mechanisms, there are still several candidate functions for subjective visual processing that are congruent with its role as a purely evaluative function that does not directly intervene in first-order objective visual processing.

Initiation of spontaneous behavior

It is possible that demonstrations of the influence of unconscious stimuli on ongoing behavior may be relatively limited to the modulation of a behavior that has been initiated for some other reason—e.g. as a response to a consciously perceived stimulus, or as a result of task instruction, rather than resulting from spontaneous initiation. That is to say, visual awareness and/or higher levels of visual clarity or confidence may support a unique role for the spontaneous initiation of behavior or adoption of task sets. For instance, although blindsight patients can perform visual discriminations above chance level when prompted to make forced choice responses (Weiskrantz, 1997), their absence of visual awareness entails that they will not knowingly initiate spontaneous action on the basis of visual stimulation, since in a sense they do not *know about* (are not *aware of*) whatever objective visual processing is taking place. By way of illustration, de Gelder et al. (2008) reported that a blindsight patient with complete cortical blindness across the whole visual field, TN, was successfully able to skillfully navigate a hallway filled with obstacles. In a news article about the study, de Gelder is quoted as saying, “At first he [TN] was nervous. He said he wouldn't be able to do it because he was blind.” Thus, although TN was able to navigate the hallway quite skillfully even without visual awareness, his residual visual processing was not sufficient to prompt the spontaneous execution of such behavior.

Information seeking, error monitoring, and error correction

Metacognitive evaluations of perceptual clarity and confidence are closely related to error monitoring and correction (Yeung & Summerfield, 2012). Subjects are capable of spontaneously detecting and correcting their own errors in simple tasks without external feedback (Rabbitt, 1966), particularly under conditions of high time pressure (Yeung & Summerfield, 2012). Subjects also tend to have longer reaction times on trials following errors (Rabbitt, 1966). Error correction and post-error slowing demonstrate a direct way in which metacognitive evaluation can influence behavior. (Although post-error slowing may also be attributed to objective performance confounds in the comparison of correct and incorrect trials.) A related concept is that low levels of confidence may prompt information seeking behavior in order to resolve the perceptual uncertainty.

Learning

A phenomenon known as the "hypercorrection effect" in the memory literature concerns the influence of confidence upon error correction (Butterfield & Metcalfe, 2001; Butterfield & Mangels, 2003; Fazio & Marsh, 2010). When subjects perform a memory test and receive performance feedback, on a subsequent re-test of the same items they are more likely to correct first-test errors that were endorsed with high, rather than low, confidence. Thus, confidence modulates learning following performance feedback, such that larger discrepancies between expectation and reality occasion stronger learning.

Communication

Accurate reporting of perceptual confidence can facilitate group decision making, and in fact two freely communicating observers who accurately calibrate their confidence ratings can perform better than a lone observer (Bahrami et al., 2010). Thus, metacognitive evaluation can facilitate the communication and integration of information coded in internal states across observers.

Conclusion

This dissertation demonstrates that objective and subjective aspects of visual processing are not only conceptually distinct, but are also functionally dissociable and dependent upon different brain structures. Converging evidence suggests that metacognitive judgments of perceptual clarity and confidence are subserved by a high-level processing stage housed in regions of anterior and lateral prefrontal cortex. Our methodological approach of controlling for performance confounds and situating analysis of metacognitive performance in a signal detection theory framework will hopefully continue to shed light in the future on the evolving field of research concerning the relationship between objective and subjective vision.

References

Adams, J. K. (1957). Laboratory studies of behavior without awareness. *Psychological bulletin*, *54*(5), 383–405.

Arfeller, C., Vonthein, R., Plontke, S. K., & Plewnia, C. (2009). Efficacy and safety of bilateral continuous theta burst stimulation (cTBS) for the treatment of chronic tinnitus: design of a three-armed randomized controlled trial. *Trials*, *10*, 74. doi:10.1186/1745-6215-10-74

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. doi:10.1016/j.neuroimage.2007.07.007

Baars, B. J. (1989). *A Cognitive Theory of Consciousness*. Cambridge, Mass: Cambridge University Press.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in cognitive sciences*, *12*(5), 193–200. doi:10.1016/j.tics.2008.02.004

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science (New York, N.Y.)*, *329*(5995), 1081–1085. doi:10.1126/science.1185718

Bargh, J. A., & Morsella, E. (2008). The Unconscious Mind. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *3*(1), 73–79.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: applications to recognition memory. *Psychological Review*, *116*(1), 84–115. doi:10.1037/a0014351

Berry, C. J., Shanks, D. R., & Henson, R. N. A. (2008). A unitary signal-detection model of implicit and explicit memory. *Trends in cognitive sciences*, *12*(10), 367–373. doi:10.1016/j.tics.2008.06.005

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: the underconfidence phenomenon. *Perception & psychophysics*, *54*(1), 75–81.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature reviews. Neuroscience*, *3*(1), 13–21. doi:10.1038/nrn701

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *The Behavioral and Brain Sciences*, *30*(5-6), 481–99; discussion 499–548. doi:S0140525X07002786

Bor, D., Duncan, J., Wiseman, R. J., & Owen, A. M. (2003). Encoding strategies dissociate prefrontal activity from working memory demand. *Neuron*, *37*(2), 361–367.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Breitmeyer, B. G. (1984). *Visual Masking: An Integrative Approach* (First Edition.). Oxford University Press, USA.

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *NeuroImage*, *16*, S497.

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of experimental psychology. Learning, memory, and cognition*, *31*(4), 587–599. doi:10.1037/0278-7393.31.4.587

Burnham, K. P., & Anderson, D. (2003). *Model Selection and Multi-Model Inference* (2nd ed.). Springer.

Butterfield, B, & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of experimental psychology. Learning, memory, and cognition*, *27*(6), 1491–1494.

Butterfield, Brady, & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. *Brain research. Cognitive brain research*, *17*(3), 793–817.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Cheesman, J., & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian journal of psychology*, *40*(4), 343–367.

Chun, M. M., & 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), 109–127.

Clarke, F. R., Birdsall, T. G., & Tanner, J. (1959). Two Types of ROC Curves and Definitions of Parameters. *The Journal of the Acoustical Society of America*, *31*(5), 629–630. doi:10.1121/1.1907764

Cleeremans, A. (2008). Consciousness: the radical plasticity thesis. *Progress in brain research*, *168*, 19–33. doi:10.1016/S0079-6123(07)68003-0

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Consciousness and metarepresentation: a computational sketch. *Neural Networks: The Official Journal of the International Neural Network Society*, *20*(9), 1032–1039. doi:10.1016/j.neunet.2007.09.011

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. doi:10.1016/S1364-6613(03)00197-9

D'Esposito, M, Postle, B. R., Ballard, D., & Lease, J. (1999). Maintenance versus manipulation of information held in working memory: an event-related fMRI study. *Brain and cognition*, *41*(1), 66–86. doi:10.1006/brcg.1999.1096

D'Esposito, Mark. (2007). From cognitive to neural models of working memory. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1481), 761–772. doi:10.1098/rstb.2007.2086

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press.

De Gelder, B., Tamietto, M., van Boxtel, G., Goebel, R., Sahraie, A., van den Stock, J., … Pegna, A. (2008). Intact navigation skills after bilateral loss of striate cortex. *Current biology: CB*, *18*(24), R1128–1129. doi:10.1016/j.cub.2008.11.002

Debner, J. A., & Jacoby, L. L. (1994). Unconscious perception: attention, awareness, and control. *Journal of experimental psychology. Learning, memory, and cognition*, *20*(2), 304–317.

Dehaene, S, Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature neuroscience*, *4*(7), 752–758. doi:10.1038/89551

Dehaene, Stanislas, & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227. doi:10.1016/j.neuron.2011.03.018

Dehaene, Stanislas, Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. doi:10.1016/j.tics.2006.03.007

Dehaene, Stanislas, Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8520–5. doi:12829797

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132*(9), 2531–2540. doi:10.1093/brain/awp111

Del Cul, Antoine, Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS biology*, *5*(10), e260. doi:10.1371/journal.pbio.0050260

Deming, W. E. (1943). Statistical adjustment of data. Retrieved from http://psycnet.apa.org/index.cfm?fa=search.displayrecord&uid=1944-00642-000

Dienes, Z. (2004). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies*, *11*, 25–45.

Dienes, Zoltán. (2008). Subjective measures of unconscious knowledge. *Progress in brain research*, *168*, 49–64. doi:10.1016/S0079-6123(07)68005-4

Dienes, Zoltán, M, T., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1322–1338. doi:10.1037/0278-7393.21.5.1322

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals--Rating-method data. *Journal of Mathematical Psychology*, *6*(3), 487–496. doi:10.1016/0022-2496(69)90019-4

Eriksen, C. W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. *Psychological review*, *67*, 279–300.

Evans, S., & Azzopardi, P. (2007). Evaluation of a “bias-free” measure of awareness. *Spatial Vision*, *20*(1-2), 61–77.

Fazio, L. K., & Marsh, E. J. (2010). Correcting false memories. *Psychological science*, *21*(6), 801–803. doi:10.1177/0956797610371341

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, *10*, 507–521.

Fleck, M. S., Daselaar, S. M., Dobbins, I. G., & Cabeza, R. (2006). Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. *Cerebral cortex (New York, N.Y.: 1991)*, *16*(11), 1623–1630. doi:10.1093/cercor/bhj097

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *32*(18), 6117–6125. doi:10.1523/JNEUROSCI.6489-11.2012

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science (New York, N.Y.)*, *329*(5998), 1541–1543. doi:10.1126/science.1191883

Fuster, J. M. (2001). The prefrontal cortex--an update: time is of the essence. *Neuron*, *30*(2), 319–333.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–76. doi:15000533

Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences*, *97*(22), 12380 –12384. doi:10.1073/pnas.97.22.12380

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., & Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human factors*, *45*(3), 349–359.

Grosbras, M.-H., & Paus, T. (2003). Transcranial magnetic stimulation of the human frontal eye field facilitates visual awareness. *The European Journal of Neuroscience*, *18*(11), 3121–6. doi:14656308

Grossheinrich, N., Rau, A., Pogarell, O., Hennig-Fast, K., Reinl, M., Karch, S., … Padberg, F. (2009). Theta burst stimulation of the prefrontal cortex: safety and impact on cognition, mood, and resting electroencephalogram. *Biological psychiatry*, *65*(9), 778–784. doi:10.1016/j.biopsych.2008.10.029

Han, S.-H., & Kim, M.-S. (2004). Visual Search Does Not Remain Efficient When Executive Working Memory Is Working. *Psychological Science*, *15*(9), 623–628. doi:10.1111/j.0956-7976.2004.00730.x

Helton, W. S., Hollander, T. D., Warm, J. S., Matthews, G., Dember, W. N., Wallaart, M., … Hancock, P. A. (2005). Signal regularity and the mindlessness model of vigilance. *British journal of psychology (London, England: 1953)*, *96*(Pt 2), 249–261. doi:10.1348/000712605X38369

Helton, W. S., & Warm, J. S. (2008). Signal salience and the mindlessness theory of vigilance. *Acta psychologica*, *129*(1), 18–25. doi:10.1016/j.actpsy.2008.04.002

Henson, R. N., Rugg, M. D., Shallice, T., & Dolan, R. J. (2000). Confidence in recognition memory for words: dissociating right prefrontal roles in episodic retrieval. *Journal of cognitive neuroscience*, *12*(6), 913–923.

Hintzman, D. L., & Curran, T. (1994). Retrieval Dynamics of Recognition and Frequency Judgments: Evidence for Separate Processes of Familiarity and Recall. *Journal of Memory and Language*, *33*(1), 1–18. doi:10.1006/jmla.1994.1001

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences*, *9*(01), 1–23. doi:10.1017/S0140525X00021269

Holender, D., & Duscherer, K. (2004). Unconscious perception: the need for a paradigm shift. *Perception & psychophysics*, *66*(5), 872–881; discussion 888–895.

Huang, Y.-Z., Edwards, M. J., Rounis, E., Bhatia, K. P., & Rothwell, J. C. (2005). Theta burst stimulation of the human motor cortex. *Neuron*, *45*(2), 201–6. doi:S0896627304008463

Huang, Y.-Z., Rothwell, J. C., Edwards, M. J., & Chen, R.-S. (2008). Effect of physiological activity on an NMDA-dependent form of cortical plasticity in human. *Cerebral Cortex (New York, N.Y.: 1991)*, *18*(3), 563–70. doi:bhm087

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. doi:10.1016/0749-596X(91)90025-F

James, W. (1890). *The principles of psychology*. New York: Holt.

Jannati, A., & Di Lollo, V. (2012). Relative blindsight arises from a criterion confound in metacontrast masking: implications for theories of consciousness. *Consciousness and cognition*, *21*(1), 307–314. doi:10.1016/j.concog.2011.10.003

Jolij, J., & Lamme, V. A. F. (2005). Repression of unconscious information by conscious processing: evidence from affective blindsight induced by transcranial magnetic stimulation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(30), 10747–51. doi:10.1073/pnas.0500834102

Kahneman, D. (1973). *Attention and Effort*. Englewood, NJ: Prentice-Hall.

Kanai, R., Walsh, V., & Tseng, C.-H. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*. doi:10.1016/j.concog.2010.06.003

Karmiloff-Smith, A. (1992). *Beyond modularity: a develop- mental perspective on cognitive science*. Cambridge, MA: MIT Press.

Kelley, C. M. ., & Lindsay, D. S. (1993). Remembering Mistaken for Knowing: Ease of Retrieval as a Basis for Confidence in Answers to General Knowledge Questions. *Journal of Memory and Language*, *32*(1), 1–24. doi:10.1006/jmla.1993.1001

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1322–1337. doi:10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–31. doi:nature07200

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science (New York, N.Y.)*, *324*(5928), 759–764. doi:10.1126/science.1169405

Kirkpatrick, S., Gelatt, C. D., Jr, & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science (New York, N.Y.)*, *220*(4598), 671–680. doi:10.1126/science.220.4598.671

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in cognitive sciences*, *11*(6), 229–235. doi:10.1016/j.tics.2007.04.005

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature*, *377*(6547), 336–8. doi:7566086

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological review*, *100*(4), 609–639.

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science: A Journal of the American Psychological Society / APS*, *18*(1), 64–71. doi:PSCI1850

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1481), 857–75.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340. doi:10.1006/ccog.2000.0494

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. London: Academic Press.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in cognitive sciences*, *10*(11), 494–501. doi:10.1016/j.tics.2006.09.001

Lamy, D., Salti, M., & Bar-Haim, Y. (2009). Neural correlates of subjective awareness and unconscious processing: an ERP study. *Journal of cognitive neuroscience*, *21*(7), 1435–1446. doi:10.1162/jocn.2009.21064

Latto, R., & Cowey, A. (1971). Visual field defects after frontal eye-field lesions in monkeys. *Brain Research*, *30*(1), 1–24. doi:4999140

Lau, H. (2008a). A higher order Bayesian decision theory of consciousness. *Progress in Brain Research*, *168*, 35–48. doi:10.1016/S0079-6123(07)68004-2

Lau, H. (2008b). Are we studying consciousness yet? In Lawrence Weiskrantz & M. Davies (Eds.), *Frontiers of Consciousness: Chichele Lectures*. Oxford University Press.

Lau, H. (2009). Volition and the functions of consciousness. In Michael Gazzaniga (Ed.), *The Cognitive Neurosciences IV* (pp. 1191–1200). MIT Press.

Lau, H. (2011). Theoretical motivations for investigating the neural correlates of consciousness. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 1–7. doi:10.1002/wcs.93

Lau, H., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18763–8.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. doi:10.1016/j.tics.2011.05.009

Link, S. W. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Macmillan, N A, & Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological bulletin*, *98*(1), 185–199.

Macmillan, Neil A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). Lawrence Erlbaum.

Mangan, B. (2001). Sensation's Ghost. The Non-Sensory "Fringe" of Consciousness. *PSYCHE*, *718 (October)*.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. doi:10.1016/j.concog.2011.09.021

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in cognitive sciences*, *9*(6), 296–305. doi:10.1016/j.tics.2005.04.010

Matthews, G., Davies, R. D., Westerman, S. J., & Stammers, R. B. (2000). *Human Performance: Cognition, Stress, and Individual Differences*. Psychology Press.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *33*(5), 1897–1906. doi:10.1523/JNEUROSCI.1890-12.2013

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*, 167–202. doi:10.1146/annurev.neuro.24.1.167

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: a nonparametric approach to statistical inference*. Newbury Park, Calif.: Sage Publications.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435–440. doi:10.1016/j.tics.2010.07.004

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64.

Mottaghy, F. M., Gangitano, M., Sparing, R., Krause, B. J., & Pascual-Leone, A. (2002). Segregation of areas related to visual working memory in the prefrontal cortex revealed by rTMS. *Cerebral Cortex (New York, N.Y.: 1991)*, *12*(4), 369–75. doi:11884352

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: an explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*(3), 465–94. doi:18567246

Münchau, A., Bloem, B. R., Irlbacher, K., Trimble, M. R., & Rothwell, J. C. (2002). Functional connectivity of human premotor and motor cortex explored with repetitive transcranial magnetic stimulation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *22*(2), 554–61. doi:11784802

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *Psychology of Learning and Motivation-Advances in Research and Theory*, *26*, 125–173.

Nuechterlein, K. H., Parasuraman, R., & Jiang, Q. (1983). Visual sustained attention: image degradation produces rapid sensitivity decrement over time. *Science (New York, N.Y.)*, *220*(4594), 327–329.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology*, *5*(3), 377–391. doi:10.1016/0022-2496(68)90083-7

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in cognitive sciences*, *12*(6), 237–241. doi:10.1016/j.tics.2008.02.014

Owen, A. M., Morris, R. G., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1996). Double dissociations of memory and executive functions in working memory tasks following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Brain: a journal of neurology*, *119 (Pt 5)*, 1597–1615.

Pascual-Leone, A., Valls-Solé, J., Wassermann, E. M., & Hallett, M. (1994). Responses to rapid-rate transcranial magnetic stimulation of the human motor cortex. *Brain: A Journal of Neurology*, *117 (Pt 4)*, 847–58. doi:7922470

Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition*, *117*(2), 182–190. doi:10.1016/j.cognition.2010.08.010

Passingham, R. E., & Wise, S. P. (2012). *The Neurobiology of the Prefrontal Cortex: Anatomy, Evolution, and the Origin of Insight*. Oxford University Press.

Peirce, C. S., & Jastrow,, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*, 73–83.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A., & Lau, H. (2011). Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *NeuroImage*, *58*(2), 605–611. doi:10.1016/j.neuroimage.2011.06.081

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257–61. doi:nn1840

Petrides, M. (1989). Frontal lobes and memory. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 3, pp. 75–90). Amsterdam: Elsevier.

Petrides, M. (1995). Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *15*(1 Pt 1), 359–375.

Petrides, M, & Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, *20*(3), 249–262.

Petrides, Michael. (2000). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research*, *133*(1), 44–54. doi:10.1007/s002210000399

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic bulletin & review*, *10*(1), 177–183.

Pins, D., & Ffytche, D. (2003). The neural correlates of conscious vision. *Cerebral cortex (New York, N.Y.: 1991)*, *13*(5), 461–474.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi:10.1037/a0019737

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'e. *Psychological Bulletin*, *71*(3), 161–173. doi:10.1037/h0026862

Pollen, D. A. (1995). Cortical areas in visual awareness. *Nature*, *377*(6547), 293–5. doi:7566083

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38. doi:10.1016/j.neuroscience.2005.06.005

Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of experimental psychology*, *71*(2), 264–272.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*(12), 1513–1515. doi:10.1038/nn.2948

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, *3*(1), 1–23. doi:10.1023/B:PHEN.0000041900.30172.e8

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. doi:10.1111/1467-9280.00067

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–175. doi:10.1080/17588921003632529

Rounis, E., Stephan, K. E., Lee, L., Siebner, H. R., Pesenti, A., Friston, K. J., … Frackowiak, R. S. J. (2006). Acute changes in frontoparietal activity after repetitive transcranial magnetic stimulation over the dorsolateral prefrontal cortex in a cued reaction time task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *26*(38), 9629–38. doi:26/38/9629

Rowe, J B, & Passingham, R. E. (2001). Working memory for location and time: activity in prefrontal area 46 relates to selection rather than maintenance in memory. *NeuroImage*, *14*(1 Pt 1), 77–86. doi:10.1006/nimg.2001.0784

Rowe, J., Friston, K., Frackowiak, R., & Passingham, R. (2002). Attention to action: specific modulation of corticocortical interactions in humans. *NeuroImage*, *17*(2), 988–998.

Rowe, James B., Toni, I., Josephs, O., Frackowiak, R. S. J., & Passingham, R. E. (2000). The Prefrontal Cortex: Response Selection or Maintenance Within Working Memory? *Science*, *288*(5471), 1656–1660. doi:10.1126/science.288.5471.1656

Ruff, C. C., Blankenburg, F., Bjoertomt, O., Bestmann, S., Freeman, E., Haynes, J.-D., … Driver, J. (2006). Concurrent TMS-fMRI and psychophysics reveal frontal influences on human retinotopic visual cortex. *Current Biology: CB*, *16*(15), 1479–88. doi:S0960-9822(06)01818-5

Sackur, J. (2013). Two dimensions of visibility revealed by multidimensional scaling of metacontrast. *Cognition*, *126*(2), 173–180. doi:10.1016/j.cognition.2012.09.013

Sahraie, A., Weiskrantz, L., Barbur, J. L., Simmons, A., Williams, S. C., & Brammer, M. J. (1997). Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(17), 9406–11. doi:9256495

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: is one measure better than the other? *Consciousness and cognition*, *19*(4), 1069–1078. doi:10.1016/j.concog.2009.12.013

Schmidt, T., & Vorberg, D. (2006). Criteria for unconscious cognition: three types of dissociation. *Perception & Psychophysics*, *68*(3), 489–504.

Schooler, J. W. (2002). Re-representing consciousness: dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, *6*(8), 339–344. doi:10.1016/S1364-6613(02)01949-6

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, *117*(2), 230–249. doi:10.1037/0033-2909.117.2.230

Sidis, B. (1898). *The psychology of suggestion*. New York: Appleton.

Slachevsky, A., Pillon, B., Fourneret, P., Pradat-Diehl, P., Jeannerod, M., & Dubois, B. (2001). Preserved adjustment but impaired awareness in a sensory-motor conflict following prefrontal lesions. *Journal of Cognitive Neuroscience*, *13*(3), 332–40. doi:11371311

Slachevsky, A., Pillon, B., Fourneret, P., Renié, L., Levy, R., Jeannerod, M., & Dubois, B. (2003). The prefrontal cortex and conscious monitoring of action: an experimental study. *Neuropsychologia*, *41*(6), 655–65. doi:12591023

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature reviews. Neuroscience*, *8*(11), 872–883. doi:10.1038/nrn2154

Stroh, M., Shaw, A. M., & Washburn, M. A. (1908). A study in guessing. *American Psychologist*, *19*, 243–245.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, *99*(1), 100–17. doi:3704032

Talelli, P., Greenwood, R. J., & Rothwell, J. C. (2007). Exploring Theta Burst Stimulation as an intervention to improve motor recovery in chronic stroke. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *118*(2), 333–42. doi:S1388-2457(06)01507-0

Tanner, W. P., Jr, & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological review*, *61*(6), 401–409.

Temple, J. G., Warm, J. S., Dember, W. N., Jones, K. S., LaGrange, C. M., & Matthews, G. (2000). The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task. *Human factors*, *42*(2), 183–194.

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: consciousness as an unconscious re-description process. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1412–1423. doi:10.1098/rstb.2011.0421

Tong, F. (2003). Primary visual cortex and visual awareness. *Nature reviews. Neuroscience*, *4*(3), 219–229. doi:10.1038/nrn1055

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *The Biological bulletin*, *215*(3), 216–242.

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: an update. *Annals of the New York Academy of Sciences*, *1124*, 239–61.

Tse, P. U., Martinez-Conde, S., Schlegel, A. A., & Macknik, S. L. (2005). Visibility, visual awareness, and visual masking of simple unattended targets are confined to areas in the occipital cortex beyond human V1/V2. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(47), 17178–17183. doi:10.1073/pnas.0508010102

Tsujimoto, S., Genovesio, A., & Wise, S. P. (2010). Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nature neuroscience*, *13*(1), 120–126. doi:10.1038/nn.2453

Turatto, M., Sandrini, M., & Miniussi, C. (2004). The role of the right dorsolateral prefrontal cortex in visual change awareness. *Neuroreport*, *15*(16), 2549–52. doi:00001756-200411150-00024

Van Gaal, S., Lamme, V. A. F., & Ridderinkhof, K. R. (2010). Unconsciously triggered conflict adaptation. *PloS One*, *5*(7), e11508. doi:10.1371/journal.pone.0011508

Warm, J S. (1984). An introduction to vigilance. In J S Warm (Ed.), *Sustained attention in human performance* (pp. 1–14). Chichester, England: Wiley.

Warm, Joel S, Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human factors*, *50*(3), 433–441.

Watson, A. B., & Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120.

Weiskrantz, L, Barbur, J. L., & Sahraie, A. (1995). Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). *Proceedings of the National Academy of Sciences of the United States of America*, *92*(13), 6122–6126.

Weiskrantz, Lawrence. (1986). *Blindsight: A Case Study and Implications*. Oxford University Press.

Weiskrantz, Lawrence. (1997). *Consciousness Lost and Found: A Neuropsychological Exploration* (1st ed.). Oxford University Press, USA.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. doi:10.1080/14639220210123806

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–76. doi:2006-23341-006

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1310–1321. doi:10.1098/rstb.2011.0416

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., … Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neuroscience research*, *68*(3), 199–206. doi:10.1016/j.neures.2010.07.2041

Zeki, S. (2008). The disunity of consciousness. *Progress in brain research*, *168*, 11–18. doi:10.1016/S0079-6123(07)68002-9

Appendix A

Signal detection theory analysis of type 1 and type 2 data: *d'* and meta-*d'*

Introduction

Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005) has provided a simple yet powerful methodology for distinguishing between *sensitivity* (an observer's ability to discriminate stimuli) and *response bias* (an observer's standards for producing different behavioral responses) in stimulus discrimination tasks. In tasks where an observer rates his confidence that his stimulus classification was correct, it may also be of interest to characterize how well the observer performs in placing these confidence ratings. For convenience, we can refer to the task of classifying stimuli as the type 1 task, and the task of rating confidence in classification accuracy as the type 2 task (Clarke, Birdsall, & Tanner, 1959). As with the type 1 task, SDT treatments of the type 2 task are concerned with independently characterizing an observer's type 2 sensitivity (how well confidence ratings discriminate between an observer's own correct and incorrect stimulus classifications) and type 2 response bias (the observer's standards for reporting different levels of confidence).

In this Appendix, we present an overview of the SDT analysis of type 1 and type 2 performance. We first provide a brief overview of type 1 SDT. We then demonstrate how the analysis of type 1 data can be extended to the type 2 task, with a discussion of how our approach compares to that of Galvin, Podd, Drga, & Whitmore (2003). We provide a more comprehensive methodological treatment of our SDT measure of type 2 sensitivity, meta-*d'* (Maniscalco & Lau, 2012), than has previously been published.

The SDT model and type 1 and type 2 ROC curves

Type 1 SDT

Suppose an observer is performing a task in which one of two possible stimulus classes (S1 or S2)[1] is presented on each trial, and that following each stimulus presentation, the observer must classify that stimulus as "S1" or "S2."[2] We may define 4 possible outcomes for each trial depending on the stimulus and the observer's response: hits, misses, false alarms, and correct rejections (Table A-1). When an S2 stimulus is shown, the observer's response can be either a hit (a correct classification as "S2") or a miss (an incorrect classification as "S1"). Similarly, when S1 is shown, the observer's response can be either a correct rejection (correct classification as "S1") or a false alarm (incorrect classification as "S2").[3]

A summary of the observer's performance is provided by hit rate and false alarm rate[4]:

$$Hit\ Rate = HR = p(resp = \text{"}S2\text{"} \mid stim = S2) = \frac{n(resp = \text{"S2"}\ , stim = S2)}{n(stim = S2)}$$

$$False\ Alarm\ Rate = FAR = p(resp = \text{"}S2\text{"} \mid stim = S1) = \frac{n(resp = \text{"S2"}\ , stim = S1)}{n(stim = S1)}$$

where *n*(*C*) denotes a count of the total number of trials satisfying the condition *C*.

Relative operating characteristic (ROC) curves define how changes in hit rate and false alarm rate are related. For instance, an observer may become more reluctant to produce "S2" responses if he

[1] Traditionally, S1 is taken to be the "signal absent" stimulus and S2 the "signal present" stimulus. Here we follow Macmillan & Creelman (2005) in using the more neutral terms S1 and S2 for the sake of generality.

[2] We will adopt the convention of placing "S1" and "S2" in quotation marks whenever they denote an observer's classification of a stimulus, and omitting quotation marks when these denote the objective stimulus identity.

[3] These category names are more intuitive when thinking of S1 and S2 as "signal absent" and "signal present." Then a hit is a successful detection of the signal, a miss is a failure to detect the signal, a correct rejection is an accurate assessment that no signal was presented, and a false alarm is a detection of a signal where none existed.

[4] Since hit rate and miss rate sum to 1, miss rate does not provide any extra information beyond that provided by hit rate and can be ignored; similarly for false alarm rate and correct rejection rate.

		Response	
		“S1”	“S2”
Stimulus	S1	Correct Rejection (CR)	False Alarm (FA)
	S2	Miss	Hit

Table A-1. Possible outcomes for the type 1 task.

is informed that S2 stimuli will rarely be presented, or if he is instructed that incorrect “S2” responses will be penalized more heavily than incorrect “S1” responses (e.g. Tanner & Swets, 1954; Macmillan & Creelman, 2005); such manipulations would tend to lower the observer’s probability of responding “S2,” and thus reduce false alarm rate and hit rate. By producing multiple such manipulations that alter the observer’s propensity to respond “S2,” multiple (FAR, HR) pairs can be collected and used to construct the ROC curve, which plots hit rate against false alarm rate (Figure A-1 B[5]).

On the presumption that such manipulations affect only the observer’s *standards* for responding “S2,” and not his underlying ability to discriminate S1 stimuli from S2 stimuli, the properties of the ROC curve as a whole should be informative regarding the observer’s *sensitivity* in discriminating S1 from S2, independent of the observer’s overall *response bias* for producing “S2” responses. The observer’s sensitivity thus determines the set of possible (FAR, HR) pairs the observer can produce (i.e. the ROC curve), whereas the observer’s response bias determines which amongst those possible pairs is actually exhibited, depending on whether the observer is conservative or liberal in responding “S2.” Higher sensitivity is associated with greater area underneath the ROC curve, whereas more conservative

[5] Note that the example ROC curve in Figure A-1 B is depicted as having been constructed from confidence data (Figure A-1 A), rather than from direct experimental manipulations on the observer’s criterion for responding “S2”. See the section titled “*Constructing pseudo-type 1 ROC curves from type 2 data*” below.

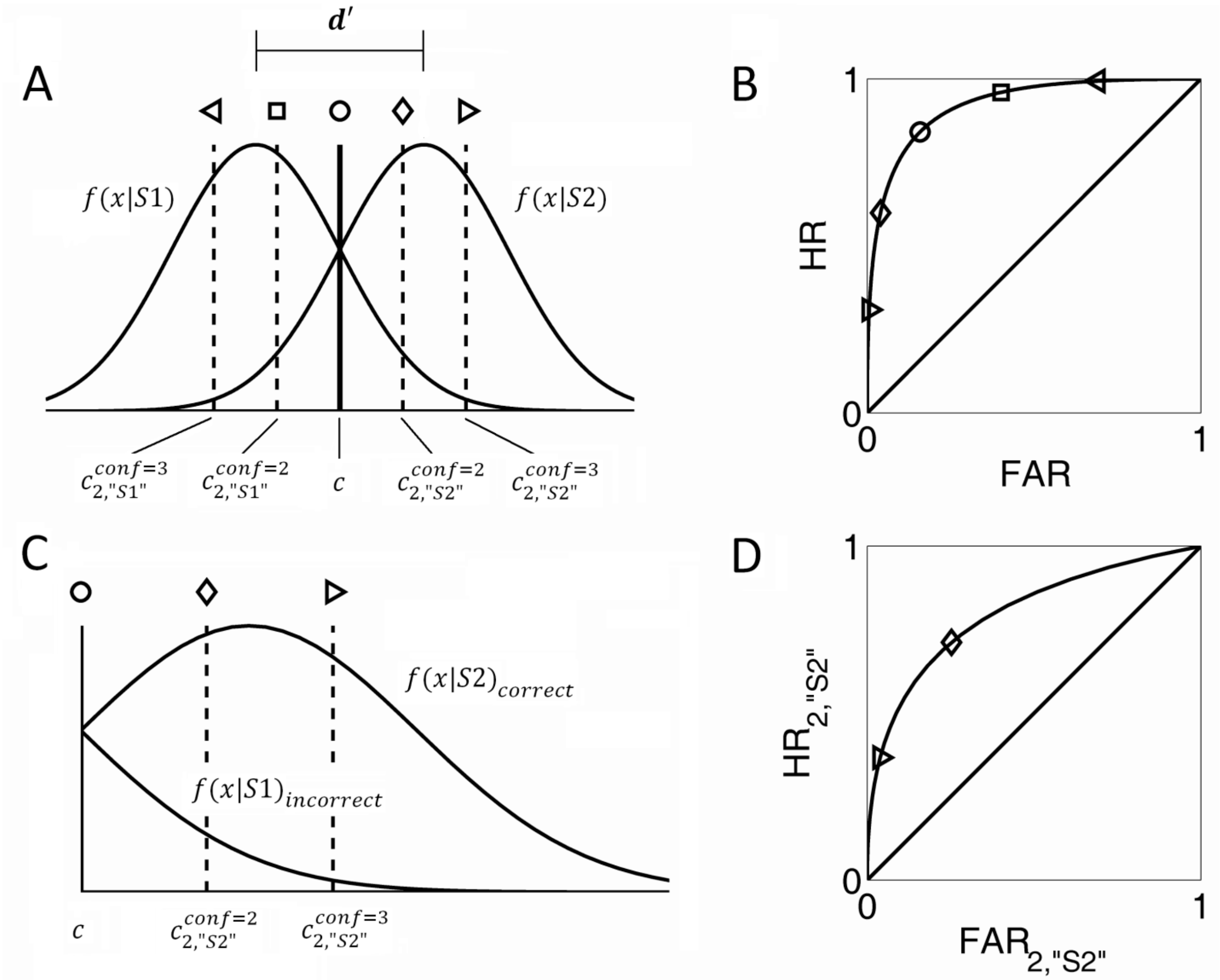

Figure A-1. Signal detection theory models of type 1 and type 2 ROC curves. (A) Type 1 SDT model. On each trial, a stimulus generates an internal response *x* within an observer, who must use *x* to decide whether the stimulus was S1 or S2. For each stimulus type, *x* is drawn from a normal distribution. The distance between these distributions is *d'*, which measures the observer's ability to discriminate S1 from S2. The stimulus is classified as "S2" if *x* exceeds a decision criterion *c*, and "S1" otherwise. In this example, the observer also rates decision confidence on a scale of 1 – 3 by comparing *x* to the additional response-specific type 2 criteria (dashed vertical lines). **(B)Type 1 ROC curve.** *d'* and *c* determine false alarm rate (FAR) and hit rate (HR). By holding *d'* constant and changing *c*, a characteristic set of (FAR, HR) points—the ROC curve—can be generated. In this example, shapes on the ROC curve mark the (FAR, HR) generated when using the corresponding criterion in panel A to classify the stimulus. (Note that, because this type 1 ROC curve is generated in part by the type 2 criteria in panel 1A, it is actually a pseudo-type 1 ROC curve, as discussed later in this paper.) **(C) Type 2 task for "S2" responses.** Consider only the trials where the observer classifies the stimulus as "S2," i.e. only the portion of the graph in panel A exceeding *c*. Then the S2 stimulus distribution corresponds to correct trials, and the S1 distribution to incorrect trials. The placement of the type 2 criteria determines the probability of high confidence for correct and incorrect trials—type 2 HR and type 2 FAR. *d′* and *c* jointly determine to what extent correct and incorrect trials for each response type are distinguishable. **(D) Type 2 ROC curve for "S2" responses.** The distributions in panel C can be used to derive type 2 FAR and HR for "S2" responses. By holding *d′* and *c* constant and changing $c_{2,\text{"S2"}}$, a set of type 2 (FAR, HR) points for "S2" responses—a response-specific type 2 ROC curve—can be generated. In this example, shapes on the ROC curve mark the ($\text{FAR}_{2,\text{"S2"}}$, $\text{HR}_{2,\text{"S2"}}$) generated when using the corresponding criterion in panel C to rate confidence.

response bias is associated with (FAR, HR) points falling more towards the lower-left portion of the ROC curve.

Measures of task performance have implied ROC curves (Swets, 1986a; Macmillan & Creelman, 2005). An implied ROC curve for a given measure of performance is a set of (FAR, HR) pairs that yield the same value for the measure. Thus, to the extent that empirical ROC curves dissociate sensitivity from bias, they provide an empirical target for theoretical measures of performance to emulate. If a proposed measure of sensitivity does not have implied ROC curves that match the properties of empirical ROC curves, then this measure cannot be said to provide a bias-free measure of sensitivity.

A core empirical strength of signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005; Figure A-1 A) is that it provides a simple computational model that provides close fits to empirical ROC curves (Green & Swets, 1966; Swets, 1986b). According to SDT, the observer performs the task of discriminating S1 from S2 by evaluating internal responses along a decision axis. Every time an S1 stimulus is shown, it produces in the mind of the observer an internal response drawn from a Gaussian probability density function. S2 stimulus presentations also generate such normally distributed internal responses. For the sake of simplicity, in the following we will assume that the probability density functions for S1 and S2 have an equal standard deviation σ.

The observer is able to discriminate S1 from S2 just to the extent that the internal responses produced by these stimuli are distinguishable, such that better sensitivity for discriminating S1 from S2 is associated with larger separation between the S1 and S2 internal response distributions. The SDT measure of sensitivity, *d'*, is thus the distance between the means of the S1 and S2 distributions, measured in units of their common standard deviation:

$$d' = \frac{\mu_{S2} - \mu_{S1}}{\sigma}$$

By convention, the internal response where the S1 and S2 distributions intersect is defined to have the value of zero, so that $\mu_{S2} = \sigma\, d' / 2$ and $\mu_{S1} = -\sigma\, d' / 2$. For simplicity, and without loss of generality, we can set $\sigma = 1$.

In order to classify an internal response *x* on a given trial as originating from an S1 or S2 stimulus, the observer compares the internal response to a *decision criterion*, *c*, and only produces “S2” classifications for internal responses that surpass the criterion.

$$response = \begin{cases} \text{"S1"}, & x \leq c \\ \text{"S2"}, & x > c \end{cases}$$

Since hit rate is the probability of responding “S2” when an S2 stimulus is shown, it can be calculated on the SDT model as the area underneath the portion of the S2 probability density function that exceeds *c*. Since the cumulative distribution function for the normal distribution with mean μ and standard deviation σ evaluated at x is

$$\Phi(x, \mu, \sigma) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

then hit rate can be derived from the parameters of the SDT model as

$$HR = 1 - \Phi(c, \mu_{S2}) = \ 1 - \Phi\left(c, \frac{d'}{2}\right)$$

And similarly,

$$FAR = 1 - \Phi(c, \mu_{S1}) = \ 1 - \Phi\left(c, -\frac{d'}{2}\right)$$

where omitting the σ parameter in ϕ is understood to be equivalent to setting σ = 1.

By systematically altering the value of *c* while holding *d′* constant, a set of (FAR, HR) pairs ranging between (0, 0) and (1, 1) can be generated, tracing out the shape of the ROC curve (Figure A-1 B). The family of ROC curves predicted by SDT matches well with empirical ROC curves across a range of experimental tasks and conditions (Green & Swets, 1966; Swets, 1986a; Swets, 1986b).

The parameters of the SDT model can be recovered from a given (FAR, HR) pair as

$$d' = \ z(HR) - \ z(FAR)$$

$$c = -.5 * [\ z(HR) + z(FAR)\ ]$$

where *z* is the inverse of the normal cumulative distribution function. Thus, SDT analysis allows us to separately characterize an observer's sensitivity (*d'*) and response bias (*c*) on the basis of a single (FAR, HR) pair, obviating the need to collect an entire empirical ROC curve in order to separately characterize sensitivity and bias—provided that the assumptions of the SDT model hold.

Type 2 SDT

Suppose we extend the empirical task described above, such that after classifying the stimulus as "S1" or "S2," the observer must provide a confidence rating that characterizes the likelihood of the stimulus classification being correct. This confidence rating task can be viewed as a secondary discrimination task. Just as the observer first had to discriminate whether the stimulus was S1 or S2 by means of providing a stimulus classification response, the observer now must discriminate whether that stimulus classification response itself was correct or incorrect by means of providing a confidence rating.[6] Following convention, we will refer to the task of classifying the stimulus as the "type 1" task, and the task of classifying the accuracy of the stimulus classification as the "type 2" task (Clarke et al., 1959; Galvin et al., 2003).

Type 2 hit rates and false alarm rates

A similar set of principles for the analysis of the type 1 task may be applied to the type 2 task. Consider the simple case where the observer rates confidence as either "high" or "low." We can then distinguish 4 possible outcomes in the type 2 task: high confidence correct trials, low confidence correct

[6] In principle, since the observer should always choose the stimulus classification response that is deemed most likely to be correct, then in a two-alternative task he should always judge that the chosen response is more likely to be correct than it is to be incorrect. Intuitively, then, the type 2 decision actually consists in deciding whether the type 1 response is *likely* to be correct or not, where the standard for what level of confidence merits being labeled as "likely to be correct" is determined by a subjective criterion than can be either conservative or liberal. Nonetheless, viewing the type 2 task as a discrimination between correct and incorrect stimulus classifications facilitates comparison with the type 1 task.

trials, low confidence incorrect trials, and high confidence incorrect trials. By direct analogy with the type 1 analysis, we may refer to these outcomes as type 2 hits, type 2 misses, type 2 correct rejections, and type 2 false alarms, respectively (Table A-2).[7]

Type 2 hit rate and type 2 false alarm rate summarize an observer's type 2 performance and may be calculated as

$$type\ 2\ HR = HR_2 = p(high\ conf \mid stim = resp) = \frac{n(high\ conf\ correct)}{n(correct)}$$

$$type\ 2\ FAR = FAR_2 = p(high\ conf \mid stim\ \neq resp) = \frac{n(high\ conf\ incorrect)}{n(incorrect)}$$

Since the binary classification task we have been discussing has two kinds of correct trials (hits and correct rejections) and two kinds of incorrect trials (misses and false alarms), the classification of type 2 performance can be further subdivided into a *response-specific* analysis, where we consider type 2 performance only for trials where the type 1 stimulus classification response was "S1" or "S2" (Table A-3).[8]

[7] The analogy is more intuitive when thinking of S1 as "signal absent" and S2 as "signal present". Then the type 2 analogue of "signal absent" is an incorrect stimulus classification, whereas the analogue of "signal present" is a correct stimulus classification. The type 2 task can then be thought of as involving the detection of this type 2 "signal."

[8] It is also possible to conduct a stimulus-specific analysis and construct stimulus-specific type 2 ROC curves. For S1 stimuli, this would consist in a plot of p(high conf|correct rejection) vs p(high conf|false alarm). Likewise for S2 stimuli—p(high conf|hit) vs p(high conf|miss). However, as will be made clear later in the text, the present approach to analyzing type 2 ROC curves in terms of the type 1 SDT model requires each type 2 (FAR, HR) pair to be generated by the application of a type 2 criterion to two overlapping distributions. For stimulus-specific type 2 data, the corresponding type 1 model consists of only one stimulus distribution, with separate type 2 criteria for "S1" and "S2" responses generating the type 2 FAR and type 2 HR. (e.g. for the S2 stimulus, a type 2 criterion for "S1" responses rates confidence for type 1 misses, and a separate type 2 criterion for "S2" responses rates confidence for type 1 hits.) Thus there is no analogue of meta-*d'* for stimulus-specific type 2 data, since *d'* is only defined with respect to the relationship between two stimulus distributions, whereas stimulus-specific analysis is restricted to only one stimulus distribution. It is possible that an analysis of stimulus-specific type 2 ROC curves

		Confidence	
		Low	High
Accuracy	Incorrect	Type 2 Correct Rejection	Type 2 False Alarm
	Correct	Type 2 Miss	Type 2 Hit

Table A-2. Possible outcomes for the type 2 task.

				Confidence	
				Low	High
Response	“S1”	Accuracy	Incorrect (Type 1 Miss)	$\mathrm{CR}_{2,\text{”S1”}}$	$\mathrm{FA}_{2,\text{”S1”}}$
			Correct (Type 1 Correct Rejection)	$\mathrm{Miss}_{2,\text{”S1”}}$	$\mathrm{Hit}_{2,\text{”S1”}}$
	“S2”	Accuracy	Incorrect (Type 1 False Alarm)	$\mathrm{CR}_{2,\text{”S2”}}$	$\mathrm{FA}_{2,\text{”S2”}}$
			Correct (Type 1 Hit)	$\mathrm{Miss}_{2,\text{”S2”}}$	$\mathrm{Hit}_{2,\text{”S2”}}$

Table A-3. Possible outcomes for the type 2 task, contingent on type 1 response (i.e. response-specific type 2 outcomes)

Thus, when considering type 2 performance only for “S1” responses,

could be conducted by positing how the type 2 criteria on either side of the type 1 criterion are coordinated, or similarly by supposing that the observer rates confidence according to an overall type 2 decision variable. For more elaboration, see the section below titled “Comparison of the current approach to that of Galvin et al (2003).”

$$HR_{2,"S1"} = p(high\,conf \mid stim = S1\,, resp = "S1") = \frac{n(high\,conf\ correct\,rejection)}{n(correct\,rejection)}$$

$$FAR_{2,"S1"} = p(high\,conf \mid stim = S2\,, resp = "S1") = \frac{n(high\,conf\ miss)}{n(miss)}$$

where the subscript “S1” indicates that these are type 2 data for type 1 “S1” responses.

Similarly for “S2” responses,

$$HR_{2,"S2"} = p(high\,conf \mid stim = S2\,, resp = "S2") = \frac{n(high\,conf\ hit)}{n(hit)}$$

$$FAR_{2,"S2"} = p(high\,conf \mid stim = S1\,, resp = "S2") = \frac{n(high\,conf\ false\ alarm)}{n(false\ alarm)}$$

From the above definitions, it follows that overall type 2 FAR and HR are weighted averages of the response-specific type 2 FARs and HRs, where the weights are determined by the proportion of correct and incorrect trials originating from each response type:

$$HR_2 = \frac{n(high\,conf\ correct)}{n(correct)} = \frac{n(high\,conf hit) + n(high\,conf\ CR)}{n(hit) + n(CR)}$$

$$= \frac{n(hit)\,HR_{2,"S2"} + n(CR)\,HR_{2,"S1"}}{n(hit) + n(CR)}$$

$$= p(hit|correct)\,HR_{2,"S2"} + [1 - p(hit|correct)]\,HR_{2,"S1"}$$

And similarly,

$$FAR_2 \ = p(FA|incorrect)\ FAR_{2,"S2"} + \ [1 - p(FA|incorrect)]\ FAR_{2,"S1"}$$

Confidence rating data may be richer than a mere binary classification. In the general case, the observer may rate confidence on either a discrete or continuous scale ranging from 1 to *H*. In this case, we can arbitrarily select a value *h*, 1 < *h* ≤ *H*, such that all confidence ratings greater than or equal to *h* are classified as "high confidence" and all others, "low confidence." We can denote this choice of imposing a binary classification upon the confidence data by writing e.g. $H_2^{conf=h}$, where the superscript *conf=h* indicates that this type 2 hit rate was calculated using a classification scheme where *h* was the smallest confidence rating considered to be "high." Thus, for instance,

$$HR_{2,"S2"}^{conf=h} = p(high\ conf\ |\ stim = S2\ , resp = "S2") = p(conf\ \ \geq h\ |\ hit)$$

Each choice of *h* generates a type 2 (FAR, HR) pair, and so calculating these for multiple values of *h* allows for the construction of a type 2 ROC curve with multiple points. When using a discrete confidence rating scale ranging from 1 to *H*, there are *H* – 1 ways of selecting *h*, allowing for the construction of a type 2 ROC curve with *H* – 1 points.

Adding response-specific type 2 criteria to the type 1 SDT model to capture type 2 data

As with the type 1 task, type 2 ROC curves allow us to separately assess an observer's sensitivity (how well confidence ratings discriminate correct from incorrect trials) and response bias (the overall propensity for reporting high confidence) in the type 2 task. However, fitting a computational model to type 2 ROC curves is somewhat more complicated than in the type 1 case. It is not sufficient to assume that correct and incorrect trials are associated with normal probability density functions in a direct analogy to the S1 and S2 distributions of type 1 SDT. The reason for this is that specifying the parameters of the type 1 SDT model—*d′* and *c*—places strong constraints on the probability density functions for correct and incorrect trials, and these derived distributions are not normally distributed (Galvin et al., 2003). In addition to this theoretical consideration, it has also been empirically demonstrated that conducting a type 2 SDT analysis that assumes normal distributions for correct and incorrect trials does not give a good fit to data (Evans & Azzopardi, 2007).

Thus, the structure of the SDT model for type 2 performance must take into account the structure of the SDT model for type 1 performance. Galvin et al. (2003) presented an approach for the SDT analysis of type 2 data based on analytically deriving formulae for the type 2 probability density functions under a suitable transformation of the type 1 decision axis. Here we present a simpler alternative approach on the basis of which response-specific type 2 ROC curves can be derived directly from the type 1 model.

In order for the type 1 SDT model to characterize type 2 data, we first need an added mechanism whereby confidence ratings can be generated. This can be accomplished by supposing that the observer simply uses additional decision criteria, analogous to the type 1 criterion *c*, to generate a confidence rating on the basis of the internal response on a given trial, *x*. In the simplest case, the observer makes a binary confidence rating—high or low—and thus needs to use two additional decision criteria to rate confidence for each kind of type 1 response. Call these response-specific type 2 criteria $c_{2, \text{"S1"}}$ and $c_{2,\text{"S2"}}$, where $c_{2, \text{"S1"}} < c$ and $c_{2, \text{"S2"}} > c$. Intuitively, confidence increases as the internal response *x* becomes more distant from *c*, i.e. as the internal response becomes more likely to have been generated by one stimulus distribution or the other[9]. More formally,

[9] See "Comparison of the current approach to that of Galvin et al (2003)" and footnote 11 for a more detailed consideration of the type 2 decision axis.

$$confidence_{resp\,=\,\text{"}S1\text{"}} = \begin{cases} \text{low}, & x \geq c_{2\,,\text{"}S1\text{"}} \\ \text{high}, & x < c_{2,\text{"}S1\text{"}} \end{cases}$$

$$confidence_{resp\,=\,\text{"}S2\text{"}} = \begin{cases} \text{low}, & x \leq c_{2,\text{"}S2\text{"}} \\ \text{high}, & x > c_{2,\text{"}S2\text{"}} \end{cases}$$

In the more general case of a discrete confidence scale ranging from 1 to *H*, then *H* – 1 type 2 criteria are required to rate confidence for each response type. (See e.g. Figure A-1 A, where two type 2 criteria on left/right of the type 1 criterion allow for confidence for “S1”/“S2” responses to be rated on a scale of 1 – 3.) We may define

$$\underline{\boldsymbol{c}}_{2,\text{"}S1\text{"}} = \left(c_{2,\text{"}S1\text{"}}^{conf=2}, c_{2,\text{"}S1\text{"}}^{conf=3}, \ldots, c_{2,\text{"}S1\text{"}}^{conf=H}\right)$$

$$\underline{\boldsymbol{c}}_{2,\text{"}S2\text{"}} = \left(c_{2,\text{"}S2\text{"}}^{conf=2}, c_{2,\text{"}S2\text{"}}^{conf=3}, \ldots, c_{2,\text{"}S2\text{"}}^{conf=H}\right)$$

where e.g. $\underline{\boldsymbol{c}}_{2,\text{"}S1\text{"}}$ is a tuple containing the *H* – 1 type 2 criteria for “S1” responses. Each $c_{2,\text{"}S1\text{"}}^{conf=y}$ denotes the type 2 criterion such that internal responses more extreme (i.e. more distant from the type 1 criterion) than $c_{2,\text{"}S1\text{"}}^{conf=y}$ are associated with confidence ratings of least *y*. More specifically,

$$confidence_{resp\,=\,\text{"}S1\text{"}} = \begin{cases} 1, & x \geq \ c_{2,\text{"}S1\text{"}}^{conf=2} \\ y, & c_{2,\text{"}S1\text{"}}^{conf=y+1} \leq x < \ c_{2,\text{"}S1\text{"}}^{conf=y}, \qquad 1 < y < H \\ H, & x < c_{2,\text{"}S1\text{"}}^{conf=H} \end{cases}$$

$$confidence_{resp\,=\,\text{"}S2\text{"}} = \begin{cases} 1, & x \leq \ c_{2,\text{"}S2\text{"}}^{conf=2} \\ y, & c_{2,\text{"}S2\text{"}}^{conf=y} < x \leq \ c_{2,\text{"}S2\text{"}}^{conf=y+1}, \qquad 1 < y < H \\ H, & x > c_{2,\text{"}S2\text{"}}^{conf=H} \end{cases}$$

The type 1 and type 2 decision criteria must have a certain ordering in order for the SDT model to be meaningful. Response-specific type 2 criteria corresponding to higher confidence ratings must be more distant from *c* than type 2 criteria corresponding to lower confidence ratings. Additionally, *c* must be larger than all type 2 criteria for “S1” responses but smaller than all type 2 criteria for “S2” responses. For convenience, we may define

$$\boldsymbol{c}_{ascending} = \left(c_{2,"S1"}^{conf=H}, c_{2,"S1"}^{conf=H-1}, \dots, c_{2,"S1"}^{conf=1}, c, c_{2,"S2"}^{conf=1}, c_{2,"S2"}^{conf=2}, \dots, c_{2,"S2"}^{conf=H}\right)$$

The ordering of decision criteria in $c_{ascending}$ from first to last is the same as the ordering of the criteria from left to right when displayed on an SDT graph (e.g. Figure A-1 A). These decision criteria are properly ordered only if each element of $c_{ascending}$ is at least as large as the previous element, i.e. only if the Boolean function $\gamma(c_{ascending})$ defined below is true:

$$\gamma(c_{ascending}) = \bigwedge_{i=1}^{2H-2} c_{ascending}(i+1) \geq c_{ascending}(i)$$

It will be necessary to use this function later on when discussing how to fit SDT models to type 2 data.

Calculating response-specific type 2 (FAR, HR) from the type 1 SDT model with response-specific type 2 criteria

Now let us consider how to calculate response-specific type 2 HR and type 2 FAR from the type 1 SDT model. Recall that

$$HR_{2,"S2"}^{conf=h} = p(conf\ \geq h \mid stim = S2\ , resp = "S2") = \frac{p(conf\ \geq h\ , hit)}{p(hit)}$$

As discussed above, *p*(hit), the hit rate, is the probability that an S2 stimulus generates an internal response that exceeds the type 1 criterion *c*. Similarly, *p*(conf ≥ *h*, hit), the probability of a hit endorsed with high confidence, is just the probability that an S2 stimulus generates an internal response that exceeds the high-confidence type 2 criterion for “S2” responses, $c_{2,"S2"}^{conf=h}$. Thus, we can straightforwardly characterize the probabilities in the numerator and denominator of $HR_{2,"S2"}^{conf=h}$ in terms of the type 1 SDT parameters, as follows:

$$HR_{2,"S2"}^{conf=h} = \frac{p(conf\ \geq h\ , hit)}{p(hit)} = \frac{1 - \Phi\left(c_{2,"S2"}^{conf=h}, \frac{d'}{2}\right)}{1 - \Phi\left(c, \frac{d'}{2}\right)}$$

By similar reasoning,

$$FAR_{2,"S2"}^{conf=h} = \frac{1 - \Phi\left(c_{2,"S2"}^{conf=h}, -\frac{d'}{2}\right)}{1 - \Phi\left(c, -\frac{d'}{2}\right)}$$

And likewise for “S1” responses,

$$HR_{2,"S1"}^{conf=h} = \frac{\Phi\left(c_{2,"S1"}^{conf=h}, -\frac{d'}{2}\right)}{\Phi\left(c, -\frac{d'}{2}\right)}$$

$$FAR_{2,"S1"}^{conf=h} = \frac{\Phi\left(c_{2,"S1"}^{conf=h}, \frac{d'}{2}\right)}{\Phi\left(c, \frac{d'}{2}\right)}$$

Figure A-1 C illustrates how type 2 (FAR, HR) arise from type 1 *d'* and *c* along with a type 2 criterion. For instance, suppose *h* = 3. Then the type 2 hit rate for "S2" responses, $HR_{2,"S2"}^{conf=3}$, is the probability of a high confidence hit (the area in the S2 distribution beyond $c_{2,"S2"}^{conf=3}$) divided by the probability of a hit (the area in the S2 distribution beyond *c*).

By systematically altering the value of the type 2 criteria while holding *d'* and *c* constant, a set of (FAR_2, HR_2) pairs ranging between (0, 0) and (1, 1) can be generated, tracing out a curvilinear prediction for the shape of the type 2 ROC curve (Figure A-1 D). Thus, according to this SDT account, specifying type 1 sensitivity (*d'*) and response bias (*c*) is already sufficient to determine response-specific type 2 sensitivity (i.e. the family of response-specific type 2 ROC curves).

Comparison of the current approach to that of Galvin et al (2003)

Before continuing with our treatment of SDT analysis of type 2 data, we will make some comparisons between this approach and the one described in Galvin et al. (2003).

SDT approaches to type 2 performance

Galvin et al were concerned with characterizing the *overall* type 2 ROC curve, rather than response-specific type 2 ROC curves. On their modeling approach, an (FAR_2, HR_2) pair can be generated by setting a single type 2 criterion on a type 2 decision axis. All internal responses that exceed this type 2 criterion are labeled "high confidence," and all others "low confidence." By systematically changing the location of this type 2 criterion on the decision axis, the entire overall type 2 ROC curve can be traced out.

However, if the internal response *x* is used to make the binary confidence decision in this way, the ensuing type 2 ROC curve behaves oddly, typically containing regions where it extends below the line of chance performance (Galvin et al, 2003). This suboptimal behavior is not surprising, in that comparing the raw value of *x* to a single criterion value essentially recapitulates the decision rule used in the type 1 task and does not take into account the relationship between *x* and the observer's type 1 criterion, which is crucial for evaluating type 1 performance. The solution is that some *transformation* of *x* must be used as the type 2 decision variable, ideally one that depends upon both *x* and *c*.

For instance, consider the transformation $t(x) = |x - c|$. This converts the initial raw value of the internal response, *x*, into the distance of *x* from the type 1 criterion. This transformed value can then plausibly be compared to a single type 2 criterion to rate confidence, e.g. an observer might rate confidence as high whenever $t(x) > 1$. Other transformations for the type 2 decision variable are possible, and the choice is not arbitrary, since different choices for type 2 decision variables can lead to different predictions for the type 2 ROC curve (Galvin et al, 2003). The optimal type 2 ROC curve (i.e. the one that maximizes area under the curve) is derived by using the likelihood ratio of the type 2 probability density functions as the type 2 decision variable (Galvin et al, 2003; Green & Swets, 1966).

We have adopted a different approach thus far. Rather than characterizing an overall (FAR_2, HR_2) pair as arising from the comparison of a single type 2 decision variable to a single type 2 criterion, we have focused on response-specific (FAR_2, HR_2) data arising from comparisons of the type 1 internal response *x* to separate type 2 decision criteria for "S1" and "S2" responses (e.g. Figure A-1 A). Thus, our approach would characterize the overall (FAR_2, HR_2) as arising from a pair of response-specific type 2 criteria set on either side of the type 1 criterion on the type 1 decision axis, rather than from a single type 2 criterion set on a type 2 decision axis. We have posited no constraints on the setting of these type

2 criteria other than that they stand in appropriate ordinal relationships to eachother. For the sake of brevity in comparing these two approaches, in the following we will refer to Galvin et al's approach as G and the current approach as C.

Type 2 decision rules and response-specific type 2 criterion setting

Notice that choosing a reasonable type 2 decision variable for G is equivalent to setting constraints on the relationship between type 2 criteria for "S1" and "S2" responses on C. For instance, on G suppose that the type 2 decision variable is defined as $t(x) = |x - c|$ and confidence is high if $t(x) > 1$. On C, this is equivalent to setting $t(\mathrm{c}_{2,\text{"S1"}}) = t(\mathrm{c}_{2,\text{"S2"}}) = |\mathrm{c}_{2,\text{"S1"}} - c| = |\mathrm{c}_{2,\text{"S2"}} - c| = 1$. In other words, assuming (on G) the general rule that confidence is high whenever the distance between *x* and *c* exceeds 1 requires (on C) that the type 2 criteria for each response type both satisfy this property of being 1 unit away from *c*. Any other way of setting the type 2 criteria for C would yield outcomes inconsistent with the decision rule posited by G. Similarly, if the type 2 decision rule is that confidence is high when type 2 likelihood ratio $\mathrm{LR}_2(x) > c_{\mathrm{LR2}}$, this same rule on C would require $\mathrm{LR}_2(\mathrm{c}_{2,\text{"S1"}}) = \mathrm{LR}_2(\mathrm{c}_{2,\text{"S2"}}) = c_{\mathrm{LR2}}$, i.e. that type 2 criteria for both response types be set at the locations of *x* on either side of *c* corresponding to a type 2 likelihood ratio of c_{LR2}.

On G, choosing a suboptimal type 2 decision variable can lead to decreased area under the overall type 2 ROC curve. This can be understood on C as being related to the influence of response-specific type 2 criterion placement on the response-specific type 2 (FAR, HR) points, which in turn affect the overall type 2 (FAR, HR) points. As shown above, overall type 2 FAR and HR are weighted averages of the corresponding response-specific type 2 FARs and HRs. But computing a weighted average for two (FAR, HR) pairs on a concave down ROC curve will yield a new (FAR, HR) pair that lies below the original ROC curve. As a consequence, more exaggerated differences in the response-specific type 2 FAR and HR due to more exaggerated difference in response-specific type 2 criterion placement will tend to drive down the area below the overall type 2 ROC curve. Thus, the overall type 2 ROC curve may decrease even while the response-specific curves stay constant, depending on how criterion setting for each response type is coordinated. This reduced area under the overall type 2 ROC curve on C due to response-specific type 2 criterion placement is closely related to reduced area under the overall type 2 ROC curve on G due to choosing a suboptimal type 2 decision variable.

For example, consider the SDT model where *d′* = 2, *c* = 0, $c_{2,\text{"S1"}}$ = -1, and $c_{2,\text{"S1"}}$ = 1. This model yields $\text{FAR}_{2,\text{"S1"}}$ = $\text{FAR}_{2,\text{"S2"}}$ = FAR_2 = .14 and $\text{HR}_{2,\text{"S1"}}$ = $\text{HR}_{2,\text{"S2"}}$ = HR_2 = .59. The type 1 criterion is optimally placed and the type 2 criteria are symmetrically placed around it. This arrangement of criteria on C turns out to be equivalent to using the type 2 likelihood ratio on G, and thus yields an optimal type 2 performance. Now consider the SDT model where *d′* = 2, *c* = 0, $c_{2,\text{"S1"}}$ = -1.5, and $c_{2,\text{"S1"}}$ = .76. This model yields $\text{FAR}_{2,\text{"S1"}}$ = .04, $\text{HR}_{2,\text{"S1"}}$ = .37, $\text{FAR}_{2,\text{"S2"}}$ = .25, $\text{HR}_{2,\text{"S2"}}$ = .71, and overall FAR_2 = .14, HR_2 = .54. Although *d′* and *c* are the same as in the previous example, now the type 2 criteria are set asymmetrically about *c*, yielding different outcomes for the type 2 FAR and HR for “S1” and “S2” responses. This has the effect of yielding a lower overall HR_2 (.54 vs .59) in spite of happening to yield the same FAR_2 (.14). Thus, this asymmetric arrangement of response-specific type 2 criteria yields worse performance on the overall type 2 ROC curve than the symmetric case for the same values of *d′* and *c*. On G, this can be understood as being the result of choosing a suboptimal type 2 decision variable in the second example (i.e. a decision variable that is consistent with the way the response-specific type 2 criteria have been defined on C).In this case, the asymmetric placement of the response-specific type 2 criteria is inconsistent with a type 2 decision variable based on the type 2 likelihood ratio.

A method for assessing overall type 2 sensitivity based on the approach of Galvin et al

In the upcoming section, we will discuss our methodology for quantifying type 2 sensitivity with meta-*d′*. meta-*d′* essentially provides a single measure that jointly characterizes the areas under the response-specific type 2 ROC curves for both “S1” and “S2” responses, and in this way provides a measure of overall type 2 sensitivity. However, in doing so, it treats the relationships of type 2 criteria across response types as purely a matter of criterion setting. However, as we have discussed, coordination of type 2 criterion setting could also be seen as arising from the construction of a type 2 decision variable, where the choice of decision variable influences area under the overall type 2 ROC curve. We take it to be a substantive conceptual, and perhaps empirical, question as to whether it is preferable to characterize these effects as a matter of criterion setting (coordinating response-specific type 2 criteria) or sensitivity (constructing a type 2 decision variable). However, if one were to decide that for some purpose it were better to view this as a sensitivity effect, then the characterization of type 2 performance provided by Galvin et al may be preferable to that of the current approach.

In the interest of recognizing this, we provide free Matlab code available online (see note at the end of the manuscript) that implements one way of using Galvin et al’s approach to evaluate an observer’s overall type 2 performance. Given the parameters of an SDT model, this code outputs the theoretically optimal[10] overall type 2 ROC curve—i.e. the overall type 2 ROC curve based on type 2 likelihood ratio, which has the maximum possible area under the curve. Maniscalco & Lau (2012), building on the suggestions of Galvin et al (2003), proposed that one way of evaluating an observer’s type 2 performance is to compare her empirical type 2 ROC curve with the theoretical type 2 ROC curve, given her type 1 performance. By comparing an observer’s empirical overall type 2 ROC curve with the theoretically optimal overall type 2 ROC curve based on type 2 likelihood ratios, the observer’s overall type 2 sensitivity can be assessed with respect to the SDT-optimal level. This approach will capture potential variation in area under the overall type 2 ROC curve that is ignored (treated as a response-specific criterion effect) by the meta-*d’* approach.

[10] Provided the assumptions of the SDT model are correct.

Advantages of the current approach

Our SDT treatment of type 2 performance has certain advantages over that of Galvin et al. One advantage is that it does not require making an explicit assumption regarding what overall type 2 decision variable an observer uses, or even that the observer constructs such an overall type 2 decision variable to begin with.[11] This is because our approach allows the type 2 criteria for each response to vary independently, rather than positing a fixed relationship between their locations. Thus, if an observer does construct an overall type 2 decision variable, our treatment will capture this implicitly by means of the relationship between the response-specific type 2 criteria; and if an observer does not use an overall type 2 decision variable to begin with, our treatment can accommodate this behavior. The question of what overall type 2 decision variables, if any, observers tend to use is a substantive empirical question, and so it is preferable to avoid making assumptions on this matter if possible.

A second, related advantage is that our approach is potentially more flexible than Galvin et al's in capturing the behavior of response-specific type 2 ROC curves, without loss of flexibility in capturing the overall type 2 ROC curve. (Since overall type 2 ROC curves depend on the response-specific curves, as shown above, our focus on characterizing the response-specific curves does not entail a deficit in capturing the overall curve.) A third advantage is that our approach provides a simple way to derive response-specific type 2 ROC curves from the type 1 SDT model, whereas deriving the overall type 2 ROC curve is more complex under Galvin et al's approach and depends upon the type 2 decision variable being assumed.

Characterizing type 2 sensitivity in terms of type 1 SDT: meta-*d'*

[11] Of course, our approach must at least implicitly assume a type 2 decision variable *within* each response type. In our treatment, the implicit type 2 decision variable for each response type is just the distance of *x* from *c*. However, for the analysis of response-specific type 2 performance for the equal variance SDT model, distance from criterion and type 2 likelihood ratio are equivalent decision variables. This is because they vary monotonically with eachother (Galvin et al 2003), and so produce the same type 2 ROC curve (Egan, 1975; Swets, Tanner, & Birdsall, 1961).

Since response-specific type 2 ROC curves can be derived directly from *d′* and *c* on the SDT model, this entails a tight theoretical relationship between type 1 and type 2 performance. One practical consequence is that type 2 sensitivity—the empirical type 2 ROC curves—can be quantified in terms of the type 1 SDT parameters *d′* and *c* (Maniscalco & Lau, 2012). However, it is necessary to explicitly differentiate instances when *d′* is meant to characterize type 1 performance from those instances when *d′* (along with *c*) is meant to characterize type 2 performance. Here we adopt the convention of using the variable names meta-*d′* and meta-*c* to refer to type 1 SDT parameters when used to characterize type 2 performance. We will refer to the type 1 SDT model as a whole, when used to characterize type 2 performance, as the meta-SDT model. Essentially, *d′* and *c* describe the type 1 SDT model fit to the type 1 ROC curve[12], whereas meta-*d′* and meta-*c* – the meta-SDT model—quantify the type 1 SDT model when used exclusively to fit type 2 ROC curves.

How do we go about using the type 1 SDT model to quantify type 2 performance? There are several choices to make before a concrete method can be proposed. In the course of discussing these issues, we will put forth the methodological approach originally proposed by Maniscalco & Lau (2012).

Which type 2 ROC curves?

As discussed in the preceding section “Comparison of the current approach to that of Galvin et al (2003),” we find the meta-SDT fit that provides the best simultaneous fit to the response-specific type 2 ROC curves for “S1” and “S2” responses, rather than finding a model that directly fits the overall type 2 ROC curve. As explained in more detail in that prior discussion, we make this selection primarily because (1) it allows more flexibility and accuracy in fitting the overall data set, and (2) it does not require

[12] When the multiple points on the type 1 ROC curve are obtained using confidence rating data, it is arguably preferable to calculate *d′* and *c* only from the (FAR, HR) pair generated purely by the observer’s type 1 response. The remaining type 1 ROC points incorporate confidence rating data and depend on type 2 sensitivity, and so estimating *d′* on the basis of these ROC points may confound type 1 and type 2 sensitivity.

making an explicit assumption regarding what type 2 decision variable the observer might use for confidence rating.

Which way of combining meta-d′ and meta-c?

A second consideration is how to characterize the response-specific type 2 ROC curves using meta-*d′* and meta-*c*. For the sake of simplifying the analysis, and for the sake of facilitating comparison between *d′* and meta-*d′*, an appealing option is to *a priori* fix the value of meta-*c* so as to be similar to the empirically observed type 1 response bias *c*, thus effectively allowing meta-*d′* to be the sole free parameter that characterizes type 2 sensitivity. However, since there are multiple ways of measuring type 1 response bias (Macmillan & Creelman, 2005), there are also multiple ways of fixing the value of meta-*c* on the basis of *c*. In addition to the already-introduced *c*, type 1 response bias can be measured with the relative criterion, *c′*:

$$c' = c \, / \, d'$$

This measure takes into account how extreme the criterion is, *relative to* the stimulus distributions.

Bias can also be measured as β*,* the ratio of the probability density function for S2 stimuli to that of S1 stimuli at the location of the decision criterion:

$$\beta = e^{cd\prime}$$

Figure A-2 shows an example of how *c*, *c′*, and β relate to the stimulus distributions when bias is fixed and *d′* varies. Panel A shows an SDT diagram for *d′* = 3 and *c* = 1. In panel B, *d′* = 1 and the three decision criteria are generated by setting *c*, *c′*, and β to the equivalent values of those exhibited by these

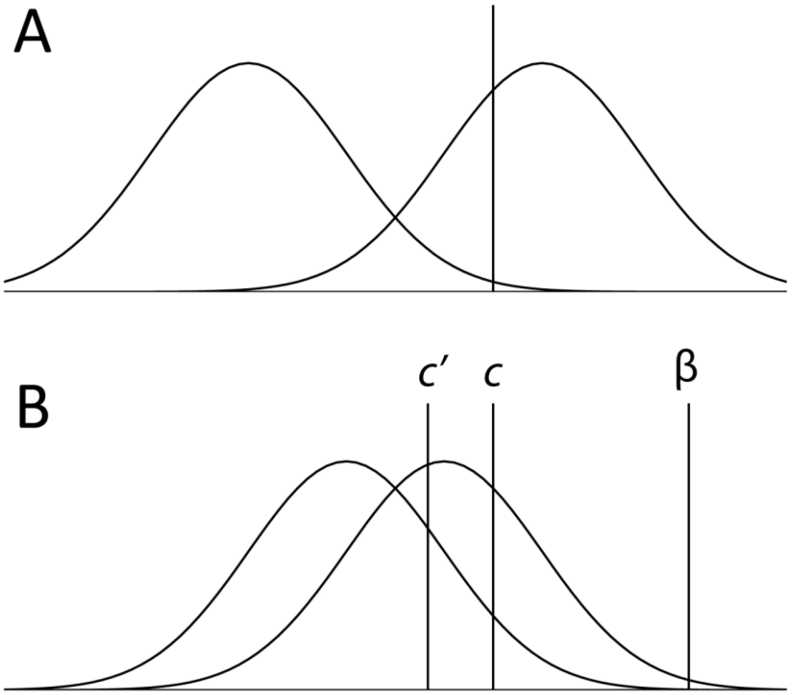

Figure A-2. Example behavior of holding response bias constant as *d′* changes for *c*, *c′*, and β. (A) An SDT graph where *d′* = 3 and *c* = 1. The criterion location can also be quantified as *c′* = *c* / *d′* = 1/3 and log β = *c* * *d′* = 3. **(B)** An SDT graph where *d′* = 1. The three decision criteria plotted here represent the locations of the criteria that preserve the value of the corresponding response bias exhibited in panel A. So e.g. the criterion marked *c′* in panel B has the same value of *c′* as the criterion in panel A (= 1/3), and likewise for *c* (constant value of 1) and β (constant value of 3).

measures in panel A. Arguably, *c′* performs best in terms of achieving a similar "cut" between the stimulus distributions in panels A and B. This is an intuitive result given that *c′* essentially adjusts the location of *c* according to *d′*. Thus, holding *c′* constant ensures that, as *d′* changes, the location of the decision criterion remains in a similar location with respect to the means of the two stimulus distributions.

By choosing *c′* as the measure of response bias that will be held constant in the estimation of meta-*d′*, we can say that when the SDT and meta-SDT models are fit to the same data set, they will have similar type 1 response bias, in the sense that they have the same *c′* value. This in turn allows us to interpret a subject's meta-*d′* in the following way: "Suppose there is an ideal subject whose behavior is perfectly described by SDT, and who performs this task with a similar level of response bias (i.e. same *c′*) as the actual subject. Then in order for our ideal subject to produce the actual subject's response-specific type 2 ROC curves, she would need her *d′* to be equal to meta-*d′*."

Thus, meta-*d′* can be found by fitting the type 1 SDT model to response-specific type 2 ROC curves, with the constraint that meta-*c′* = *c′*. (Note that in the below we list meta-*c,* rather than meta-*c′,* as a parameter of the meta-SDT model. The constraint meta-*c′* = *c′* can thus be satisfied by ensuring meta-*c* = meta-*d′* * *c′*.)

What computational method of fitting?

If the response-specific type 2 ROC curves contain more than one empirical (FAR_2, HR_2) pair, then in general an exact fit of the model to the data is not possible. In this case, fitting the model to the data requires minimizing some loss function, or maximizing some metric of goodness of fit.

Here we consider the procedure for finding the parameters of the type 1 SDT model that maximize the likelihood of the response-specific type 2 data. Maximum likelihood approaches for fitting SDT models to type 1 ROC curves with multiple data points have been established (Ogilvie & Creelman, 1968; Dorfman & Alf, 1969). Here we adapt these existing type 1 approaches to the type 2 case. The likelihood of the type 2 data can be characterized using the multinomial model as

$$L_{type\ 2}(\theta|data) \propto \prod_{y,s,r} Prob_{\theta}(conf = y \mid stim = s\ , resp = r)^{n_{data}(conf=y|stim=s\ ,resp=r)}$$

Maximizing likelihood is equivalent to maximizing log-likelihood, and in practice it is typically more convenient to work with log-likelihoods. The log-likelihood for type 2 data is given by

$$\log L_{type\ 2}\,(\theta|data) \propto \sum_{y,s,r} n_{data}\ \log Prob_{\theta}$$

θ is the set of parameters for the meta-SDT model:

$$\theta = (\mathrm{meta}d', \mathrm{meta}c, \mathrm{meta}\underline{\boldsymbol{c}}_{2,"S1"},\ \mathrm{meta}\underline{\boldsymbol{c}}_{2,"S2"})$$

$n_{data}(conf = y \mid stim = s\ , resp = r)$ is a count of the number of times in the data a confidence rating of *y* was provided when the stimulus and response were *s* and *r*.

y, *s*, and *r* are indeces ranging over all possible confidence ratings, stimulus classes, and stimulus classification responses, respectively.

$Prob_{\theta}(conf = y \mid stim = s\ , resp = r)$ is the model-predicted probability of generating confidence rating *y* for trials where the stimulus and response were *s* and *r*, given the parameter values specified in θ.

Calculation of these type 2 probabilities from the type 1 SDT model is similar to the procedure used to calculate the response-specific type 2 FAR and HR. For notational convenience, below we express these probabilities in terms of the standard SDT model parameters, omitting the “meta” prefix.

For convenience, define

$$\underline{\dot{\boldsymbol{c}}}_{2,"S1"} = \left(c, c_{2,"S1"}^{conf=2}, c_{2,"S1"}^{conf=3}, \dots, c_{2,"S1"}^{conf=H}, -\infty\right)$$

$$\underline{\dot{\boldsymbol{c}}}_{2,"S2"} = \left(c, c_{2,"S2"}^{conf=2}, c_{2,"S2"}^{conf=3}, \dots, c_{2,"S2"}^{conf=H}, \infty\right)$$

Then

$$Prob(conf = y \mid stim = S1\ , resp = "S1") = \frac{\Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S1"}(y), -\frac{d'}{2}\right) - \Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S1"}(y+1), -\frac{d'}{2}\right)}{\Phi\left(c, -\frac{d'}{2}\right)}$$

$$Prob(conf = y \mid stim = S2\ , resp = "S1") = \frac{\Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S1"}(y), \frac{d'}{2}\right) - \Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S1"}(y+1), \frac{d'}{2}\right)}{\Phi\left(c, \frac{d'}{2}\right)}$$

$$Prob(conf = y \mid stim = S1\ , resp = "S2") = \frac{\Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S2"}(y+1), -\frac{d'}{2}\right) - \Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S2"}(y), -\frac{d'}{2}\right)}{1 - \Phi\left(c, -\frac{d'}{2}\right)}$$

$$Prob(conf = y \mid stim = S2\ , resp = "S2") = \frac{\Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S2"}(y+1), \frac{d'}{2}\right) - \Phi\left(\underline{\dot{\boldsymbol{c}}}_{2,"S2"}(y), \frac{d'}{2}\right)}{1 - \Phi\left(c, \frac{d'}{2}\right)}$$

An illustration of how these type 2 probabilities are derived from the type 1 SDT model is provided in Figure A-3.

The multinomial model used as the basis for calculating likelihood treats each discrete type 2 outcome (conf=*y* | stim=*s* , resp=*r*) as an event with a fixed probability that occurred a certain number of times in the data set, where outcomes across trials are assumed to be statistically independent. The probability of the entire set of type 2 outcomes across all trials is then proportional to the product of the probability of each individual type 2 outcome, just as e.g. the probability of throwing 4 heads and 6 tails for a fair coin is proportional to $.5^4 * .5^6$.

Likelihood, *L*(θ), can be thought of as measuring how probable the empirical data is, according to the model parameterized with θ. A very low *L*(θ) indicates that the model with θ would be very unlikely to generate a pattern like that observed in the data. A higher *L*(θ) indicates that the data are more in line with the typical behavior of data produced by the model with θ. Mathematical optimization techniques can be used to find the values of θ that maximize the likelihood, i.e. that create maximal

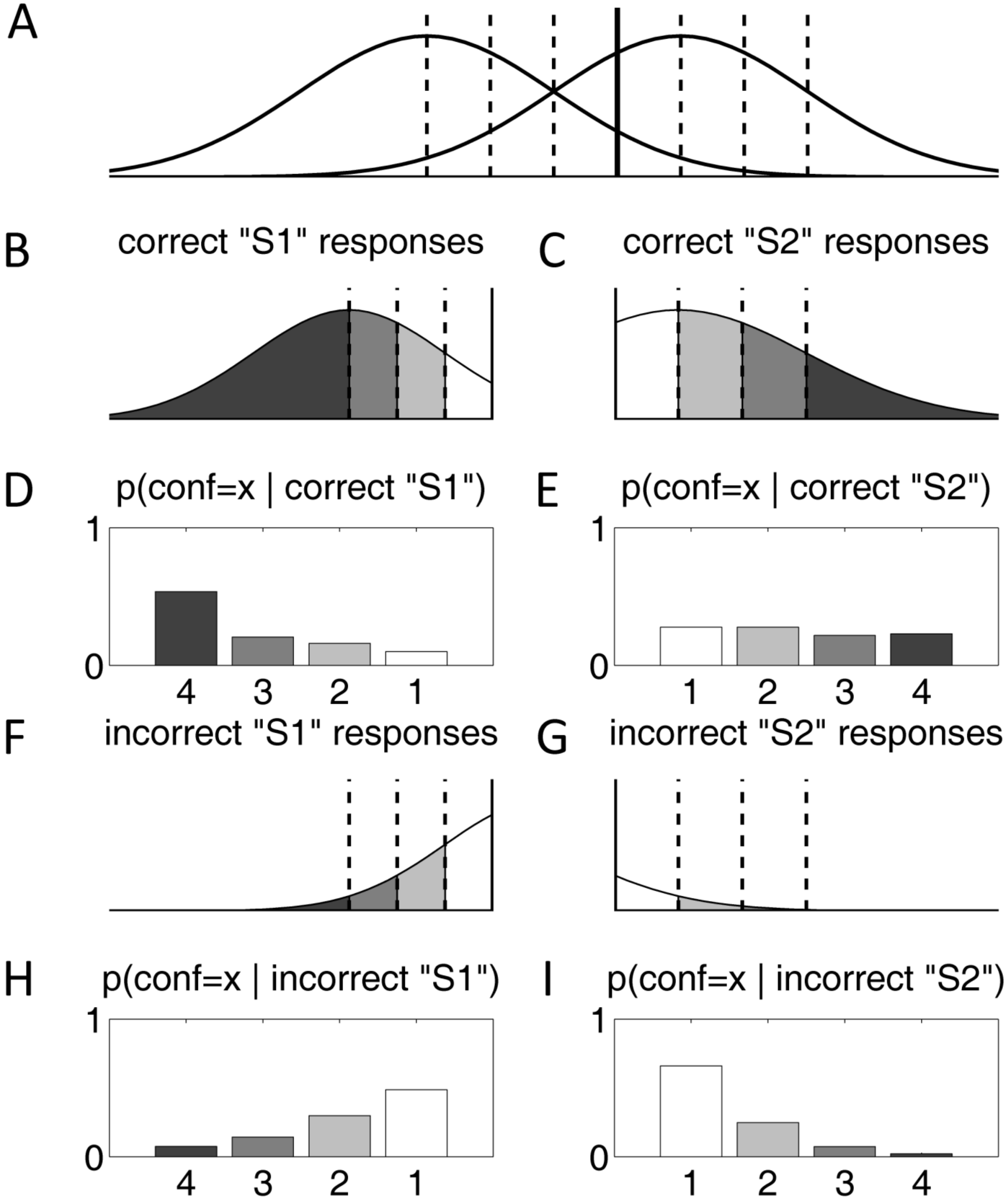

Figure A-3. Type 2 response probabilities from the SDT model. (A) An SDT graph with d' = 2 and decision criteria *c* = .5, $\underline{\boldsymbol{c}}_{2,\text{"S1"}}$ = (0, -.5, -1), and $\underline{\boldsymbol{c}}_{2,\text{"S2"}}$ = (1, 1.5, 2). The type 1 criterion (solid vertical line) is set to the value of 0.5, corresponding to a conservative bias for providing "S2" responses, in order to create an asymmetry between "S1" and "S2" responses for the sake of illustration. Seven decision criteria are used in all, segmenting the decision axis into 8 regions. Each region corresponds to one of the possible permutations of type 1 and type 2 responses, as there are two possible stimulus classifications and four possible confidence ratings. **(B-I) Deriving probability of confidence rating contingent on type 1 response and accuracy.** How would the SDT model depicted in panel (A) predict the probability of each confidence rating for correct “S1” responses? Since we wish to characterize “S1” responses, we need consider only the portion of the SDT graph falling to the left of the type 1 criterion. Since “S1” responses are only correct when the S1 stimulus was actually presented, we can further limit our consideration to internal responses generated by S1 stimuli. This is depicted in panel (B). This distribution is further subdivided into 4 levels of confidence by the 3 type 2 criteria (dashed vertical lines), where darker regions correspond to higher confidence. The area under the S1 curve in each of these regions, divided by the total area under the S1 curve that falls below the type 1 criterion, yields the probability of reporting each confidence level, given that the observer provided a correct “S1” response. Panel (D) shows these probabilities as derived from areas under the curve in panel (B). The remaining panels display the analogous logic for deriving confidence probabilities for incorrect “S1” responses (F, H), correct “S2” responses (C, E), and incorrect “S2” responses (G, I).

concordance between the empirical distribution of outcomes and the model-expected distribution of outcomes.

The preceding approach for quantifying type 2 sensitivity with the type 1 SDT model—i.e. for fitting the meta-SDT model—can be summarized as a mathematical optimization problem:

$$\theta^* = \underset{\theta}{\arg\max}\ L_{type\ 2}(\theta|data), \qquad \text{subject to:}\quad \mathrm{meta}c' = c',\quad \gamma\left(\mathrm{meta}c_{ascending}\right)$$

where type 2 sensitivity is quantified by $\mathrm{meta}d' \in \theta^*$.

$\gamma\left(\mathrm{meta}c_{ascending}\right)$ is the Boolean function described previously, which returns a value of “true” only if the type 1 and type 2 criteria stand in appropriate ordinal relationships.

We provide free Matlab code, available online, for implementing this maximum likelihood procedure for fitting the meta-SDT model to a data set (see note at the end of the manuscript).

Toy example of meta-*d'* fitting

An illustration of the meta-*d'* fitting procedure is demonstrated in Figure A-4 using simulated data. In this simulation, we make the usual SDT assumption that on each trial, presentation of stimulus S generates an internal response *x* that is drawn from the probability density function of S, and that a type 1 response is made by comparing *x* to the decision criterion *c*. However, we now add an extra mechanism to the model to allow for the possibility of added noise in the type 2 task. Let us call the internal response used to rate confidence x_2. The type 1 SDT model we have thus far considered assumes $x_2 = x$. In this example, we suppose that x_2 is a noisier facsimile of *x*. Formally,

$$x_2 = x + \xi, \quad \xi \sim N(0, \sigma_2)$$

Where $N(0, \sigma_2)$ is the normal distribution with mean 0 and standard deviation σ_2. The parameter σ_2 thus determines how much noisier x_2 is than *x*. For $\sigma_2 = 0$ we expect meta-*d'* = *d'*, and for $\sigma_2 > 0$ we expect meta-*d'* < *d'*.

The simulated observer rates confidence on a 4-point scale by comparing x_2 to response-specific type 2 criteria, using the previously defined decision rules for confidence in the type 1 SDT model.[13]

[13] Note that for this model, it is possible for *x* and x_2 to be on opposite sides of the type 1 decision criterion *c*. This is not problematic, since only *x* is used to provide the type 1 stimulus classification. It is also possible for x_2 to surpass some of the type 2 criteria on the opposite side of *c*. For instance, suppose that x = -0.5, x_2 = +0.6, *c* = 0, and $c_{2,"S2"}^{conf=h}$ = +0.5. Then *x* is classified as an S1 stimulus, and yet x_2 surpasses the criterion for rating “S2” responses with a confidence of *h*. Thus, there is potential for the paradoxical result whereby the type 1 response is “S1” and yet the type 2 confidence rating is rated highly due to the relatively strong “S2”-ness of x_2. In this example, the paradox is resolved by the definition of the type 2 decision rules stated above, which stipulate that internal responses are only evaluated with respect to the response-specific type 2 criteria that are congruent with the type 1 response. Thus, in this case, the decision rule would not compare x_2 with the type 2 criteria for “S2”

We first considered the SDT model with *d′* = 2, *c* = 0, $\underline{\boldsymbol{c}}_{2,\text{"S1"}}$ = (-.5, -1, -1.5), $\underline{\boldsymbol{c}}_{2,\text{"S2"}}$ = (.5, 1, 1.5) and $\sigma_2 = 0$. Because $\sigma_2 = 0$, this is equivalent to the standard type 1 SDT model. The SDT graph for these parameter values is plotted in Figure A-4 A. Using these parameter settings, we computed the theoretical probability of each confidence rating for each permutation of stimulus and response. These probabilities for “S1” responses are shown in panels C and D, and the corresponding type 2 ROC curve is shown in panel E. (Because the type 1 criterion *c* is unbiased and the type 2 criteria are set symmetrically about *c*, confidence data for “S2” responses follow an identical distribution to that of “S1” responses and are not shown.)

Next we simulated 10,000,000 trials using the same parameter values as the previously considered model, with the exception that $\sigma_2 = 1$. With this additional noise in the type 1 task, type 2 sensitivity should decrease. This decrease in type 2 sensitivity can be seen in the type 2 ROC curve in panel E. There is more area underneath the type 2 ROC curve when $\sigma_2 = 0$ than when $\sigma_2 = 1$.

We performed a maximum likelihood fit of meta-*d′* to the simulated type 2 data using the fmincon function in the optimization toolbox for Matlab (MathWorks, Natick, MA), yielding a fit with parameter values meta-*d′* = 1.07, meta-*c* = 0, meta-$\underline{\boldsymbol{c}}_{2,\text{"S1"}}$ = (-.51, -.77, -1.06), and meta-$\underline{\boldsymbol{c}}_{2,\text{"S2"}}$ = (.51, .77, 1.06). The SDT graph for these parameter values is plotted in Figure A-4 B.

Panels C and D demonstrate the component type 2 probabilities used for computing the type 2 likelihood. The response-specific type 2 probabilities for $\sigma_2 = 0$ are not distributed the same way as those for $\sigma_2 = 1$, reflecting the influence of adding noise to the internal response for the type 2 task. Computing meta-*d′* for the $\sigma_2 = 1$ data consists in finding the parameter values of the ordinary type 1 SDT model that maximize the likelihood of the $\sigma_2 = 1$ response-specific type 2 data. This results in a type 1 SDT model whose theoretical type 2 probabilities closely match the empirical type 2 probabilities for

responses to begin with. Instead, it would find that x_2 does not surpass the minimal confidence criterion for “S1” responses (i.e. $x_2 > c > c_{2,"S1"}^{conf=2}$) and would therefore assign x_2 a confidence of 1. Thus, in this case, the paradoxical outcome is averted. But such potentially paradoxical results need to be taken into account for any SDT model that posits a potential dissociation between *x* and x_2.

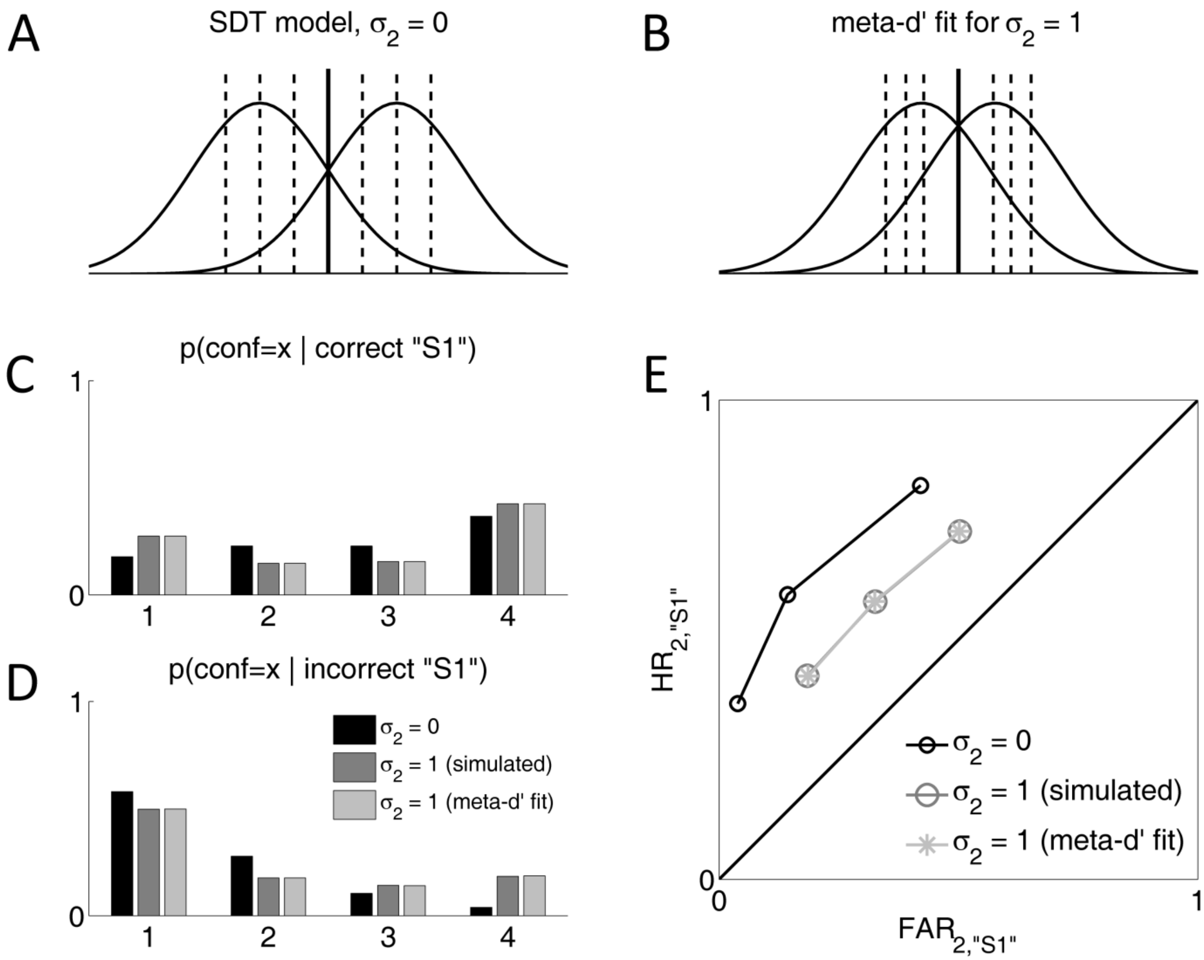

Figure A-4. Fitting meta-d' to response-specific type 2 data. (A) Graph for the SDT model where *d'* = 2 and σ_2 = 0 (see text for details). **(B)** A model identical to that in panel A, with the exception that σ_2 = 1, was used to create simulated data. This panel displays the SDT graph of the parameters for the meta-*d'* fit to the σ_2 = 1 data. **(C-D) Response-specific type 2 probabilities.** The maximum likelihood method of fitting meta-*d'* to type 2 data uses response-specific type 2 probabilities as the fundamental unit of analysis. The type 1 SDT parameters that maximize the likelihood of the type 2 data yield distributions of response-specific type 2 probabilities closely approximating the empirical (here, simulated) distributions. Here we only show the probabilities for "S1" responses; because of the symmetry of the generating model, "S2" responses follow identical distributions. **(E) Response-specific type 2 ROC curves.** ROC curves provide a more informative visualization of the type 2 data than the raw probabilities. Here it is evident that there is considerably less area under the type 2 ROC curve for the σ_2 = 1 simulation than is predicted by the σ_2 = 0 model. The meta-*d'* fit provides a close match to the simulated data.

the simulated $\sigma_2 = 1$ data (Figure A-4 C and D). Because type 2 ROC curves are closely related to these type 2 probabilities, the meta-*d′* fit also produces a type 2 ROC curve closely resembling the simulated curve, as shown in panel E.

Interpretation of meta-*d′*

Notice that because meta-*d′* characterizes type 2 sensitivity purely in terms of the type 1 SDT model, it does not explicitly posit any mechanisms by means of which type 2 sensitivity varies. Although the meta-*d′* fitting procedure gave a good fit to data simulated by the toy σ_2 model discussed above, it could also produce similarly good fits to data generated by different models that posit completely different mechanisms for variation in type 2 performance. In this sense, meta-*d′* is descriptive but not explanatory. It describes how an ideal SDT observer with similar type 1 response bias as the actual subject would have achieved the observed type 2 performance, rather than explain how the actual subject achieved their type 2 performance.

The primary virtue of using meta-*d′* is that it allows us to quantify type 2 sensitivity in a principled SDT framework, and compare this against SDT expectations of what type 2 performance *should have been*, given performance on the type 1 task, all while remaining agnostic about the underlying processes. For instance, if we find that a subject has $d' = 2$ and meta-$d' = 1$, then (1) we have taken appropriate SDT-inspired measures to factor out the influence of response bias in our measure of type 2 sensitivity; (2) we have discovered a violation of the SDT expectation that meta-$d' = d' = 2$, giving us a point of reference in interpreting the subject's metacognitive performance in relation to their primary task performance and suggesting the possibility of suboptimal metacognition; and (3) we have done so while making minimal assumptions and commitments regarding the underlying processes.

Another important point for interpretation concerns the raw meta-*d′* value, as opposed to its value in relation to *d′*. Suppose observers A and B both have meta-$d' = 1$, but $d'_A = 1$ and $d'_B = 2$. Then there is a sense in which they have equivalent metacognition, as their confidence ratings are equally sensitive in discerning correct from incorrect trials. But there is also a sense in which A has superior metacognition, since A was able to achieve the same level of meta-*d′* as B in spite of a lower *d′*. In a sense, A is more metacognitively ideal, according to SDT. We can refer to the first kind of metacognition, which depends only on meta-*d′*, as “absolute type 2 sensitivity,” and the second kind, which depends on the relationship between meta-*d′* and *d′*, as “relative type 2 sensitivity.” Absolute and relative type 2 sensitivity are distinct constructs that inform us about distinct aspects of metacognitive performance.

Code for implementing overall and response-specific meta-*d′* analysis

We provide free Matlab scripts for conducting type 1 and type 2 SDT analysis, including functions to find the maximum likelihood fits of overall and response-specific meta-*d′* to a data set, at http://www.columbia.edu/~bsm2105/type2sdt

Appendix B

Comparison of dual channel SDT models in Chapter 1 and Del Cul et al (2009)

Although our SDT implementation of the independent dual channel model described in Chapter 1 (call it DC_{SDT}) is intended to capture the primary computational features of the model described in Del Cul, Dehaene, Reyes, Bravo, and Slachevsky (2009) (call it DC_{accum}), it is not identical. In this supplement, we discuss the relevant similarities and differences between the models. Despite their technical differences, they share core computational features that encapsulate a general theory of the sensory processing structures underlying perceptual decision making and reports of conscious awareness. Here we discuss the conceptual similarities between the models and demonstrate that the core behavioral patterns captured by DC_{accum} in Del Cul et al can also be produced by DC_{SDT}. Thus, DC_{SDT} is an acceptable stand-in for DC_{accum} in the present work, capturing the core features of its sensory processing architecture.

Computationally, both models describe the perceptual decision making process as arising from the comparison of noisy sensory evidence to a decision criterion. The primary difference between the models is that DC_{accum} is an accumulator model that explicitly describes the dynamics of this perceptual decision process within the level of individual trials (Laming, 1968; Link, 1992; Ratcliff & Rouder, 1998), whereas DC_{SDT} does not.

The full details of DC_{accum} are described in the supplementary material of Del Cul et al. (2009). Here, we provide a brief summary. In DC_{accum}, two separate processing channels, one "conscious" and the other "unconscious," accumulate noisy sensory evidence over time for each possible stimulus identification response (e.g. evidence for responding "squares" and evidence for responding "diamonds" would be accumulated simultaneously). Response alternatives "race" each other in each channel, such that a behavioral identification of the stimulus is produced as soon as a response alternative in one of the channels achieves a threshold level of evidence. The first channel to achieve this threshold level of evidence for a response alternative within a predefined time period δ emits the corresponding stimulus identification response. If the stimulus identification response arises from the conscious channel, the observer additionally reports that the stimulus was “seen.” Conversely, if the stimulus identification response arises from the unconscious channel, the observer reports that the stimulus was “not seen.” If no stimulus identification response achieves the threshold level of evidence within δ, then the unconscious channel emits a response on the basis of the stimulus alternative that currently has the most sensory evidence.

Crucially, independent sources of noise contaminate sensory processing in the two channels, so that noisy fluctuations in evidence accumulation in one channel are not reflected in the noisy fluctuations of the other. These independent sources of noise correspond to the physiological notion that sensory computations are taking place in separate, unconnected processing streams.

DC_{SDT} does not explicitly describe the temporal dynamics corresponding to evidence accumulation within a trial. However, for two main reasons, it remains conceptually comparable to DC_{accum}:

(1) DC_{SDT} and DC_{accum} share the same core computational principle, namely that trials associated with subjective reports of high and low stimulus visibility are associated with distinct processing channels which are subject to independent sources of noise. This general principle is what we wish to assess by including DC_{SDT} in the model comparison analysis described in the manuscript.

(2) One major advantage of accumulator models over atemporal SDT models is that the former can model patterns in behavioral response time data. However, although Del Cul et al’s model posits a dynamic evidence accumulation process, they did not constrain their model fitting procedure with response time data. Rather, they fit the model only to measures of task performance and reports of stimulus visibility. Thus, the same behavioral indeces of performances used to fit DC_{accum} to the data in

Del Cul et al were also used to fit DC_{SDT} to the data in the current manuscript. It is thus not the case that DC_{accum}, as described in Del Cul et al, accounts for a wider range of behavioral data than does DC_{SDT}, in spite of their computational differences.

In summary, DC_{SDT} and DC_{accum} are broadly comparable in that they posit the same core computational principle in order to account for patterns in the same sorts of behavioral data.

Nonetheless, it remains possible that the technical differences between the models are sufficient to make them different in the particular patterns of behavioral data that they can account for. Below, we perform a simulation analysis that reproduces the key patterns of behavioral data in patient and control populations used to support DC_{accum} in Del Cul et al, and show that DC_{SDT} can also generate these patterns. Thus, the two models are comparable not just at a broad conceptual level but also in the specific patterns of data that they can generate.

Simulation

$\underline{DC_{accum}}$

In Del Cul et al, subjects performed a digit identification task. One of ten digits (0 – 9) was displayed on every trial, followed by a mask. The target-mask stimulus onset asynchrony could take on one of eight possible values. Subjects verbally indicated whether the masked digit was seen or not, and then provided a forced-choice judgment regarding the digit identitiy. The authors computed measures of task performance and subjective report for frontal lesion patients and healthy controls, and fit DC_{accum} to this data set.

DC_{accum} has 5 free parameters:

σ_i : standard deviation of the noise added to sensory signals at every time step during the “input” stage; this noise is shared by both channels

σ_r : standard deviation of the noise added to sensory signals at every time step for the unconscious channel

σ_w : standard deviation of the noise added to sensory signals at every time step for the conscious channel

θ : the decision threshold describing how probable a given response alternative must be in order for the corresponding behavioral response to be emitted

δ : the period of time during which evidence accumulation occurs; if no response is generated within δ time units, then the best-supported stimulus identification response in the unconscious channel is emitted as the behavioral response

In their data fitting, Del Cul et al constrained θ to be equal in the conscious and unconscious channels, and *a priori* set δ = 6 (corresponding to the number of SOAs; although data was collected at 8 SOAs, the model was fit to data from the first 6 SOAs only). They first fit DC_{accum} to the data from the healthy subjects, yielding parameter values of $\sigma_i = 3.44$, $\sigma_r = 9.07$, $\sigma_w = 1.12$, and $\theta = 0.893$. They then fit the model to the patient data, using the same parameter values and allowing only σ_w to vary. The value of σ_w obtained for patients was 2.53.

We created a simulation aiming to reproduce the relevant features of the behavioral data in Del Cul et al by using DC_{accum}. In all simulations, we specified the parameter values of DC_{accum} *a priori*, simulated 20,000 trials at each SOA for each subject group (patients / controls), and subsequently computed mean levels of task performance and subjective report as a function of SOA and subject group.

In order to compare DC_{accum} to DC_{SDT}, it was necessary to adjust DC_{accum} to account for 2 stimulus alternatives rather than 10. All other specifications of the model were set as described in Del Cul et al.

Initially, we set the parameter values in our simulation equal to those found for healthy subjects in Del Cul et al, as described above. However, this had the effect of yielding near-chance levels of task performance in the simulated data. It is possible that this occurred due to the fact that, after each time step, accumulated evidence is normalized and converted into a probability value. The normalization factor grows larger with the number of response alternatives. Thus, with only 2 stimulus alternatives, evidence for response alternatives may more rapidly arrive at the decision threshold than it would in the 10 stimulus case, even in spite of having the same level of sensory noise. This reduction in the time devoted to evidence accumulation would have the effect of reducing task performance.

In order to correct for this, we adjusted the parameter values as follows. We defined constraints on the parameter values such that

$\sigma_r = (9.07/3.44) * \sigma_i$

$\sigma_{w,\text{ control}} = (1.12/3.44) * \sigma_i$

$\sigma_{w,\text{ patient}} = (2.53/3.44) * \sigma_i$

and manually searched for the value of σ_i that would yield a similar profile of behavioral data to that found in Del Cul et al. Thus, we tuned the parameter values of DC_{accum} to be suitable for adaptation to the 2-stimulus case, while still preserving the ratios between parameter values found by Del Cul et al.

We found that setting $\sigma_i = 0.2$ yielded a satisfactory result, as displayed in Figure B-2 (compare with our Figure B-1, which is a reproduction of the results of model fitting originally presented in Del Cul et al's Figure S3). In particular, our implementation of DC_{accum} captured the primary patterns of interest in the behavioral data as identified by Del Cul et al:

- Overall task performance (p(correct)) and subjective report (p(seen)) increase with SOA, and these measures are higher in controls than in patients (Figure B-2 A, B)

- Task performance conditioned on subjective report (p(correct | seen) and p(correct | not seen)) is virtually identical for patients and controls (Figure B-2 C)
- Subjective report conditioned on task performance (p(seen | correct) and p(seen | incorrect)) is larger for controls than for patients (Figure B-2 D)

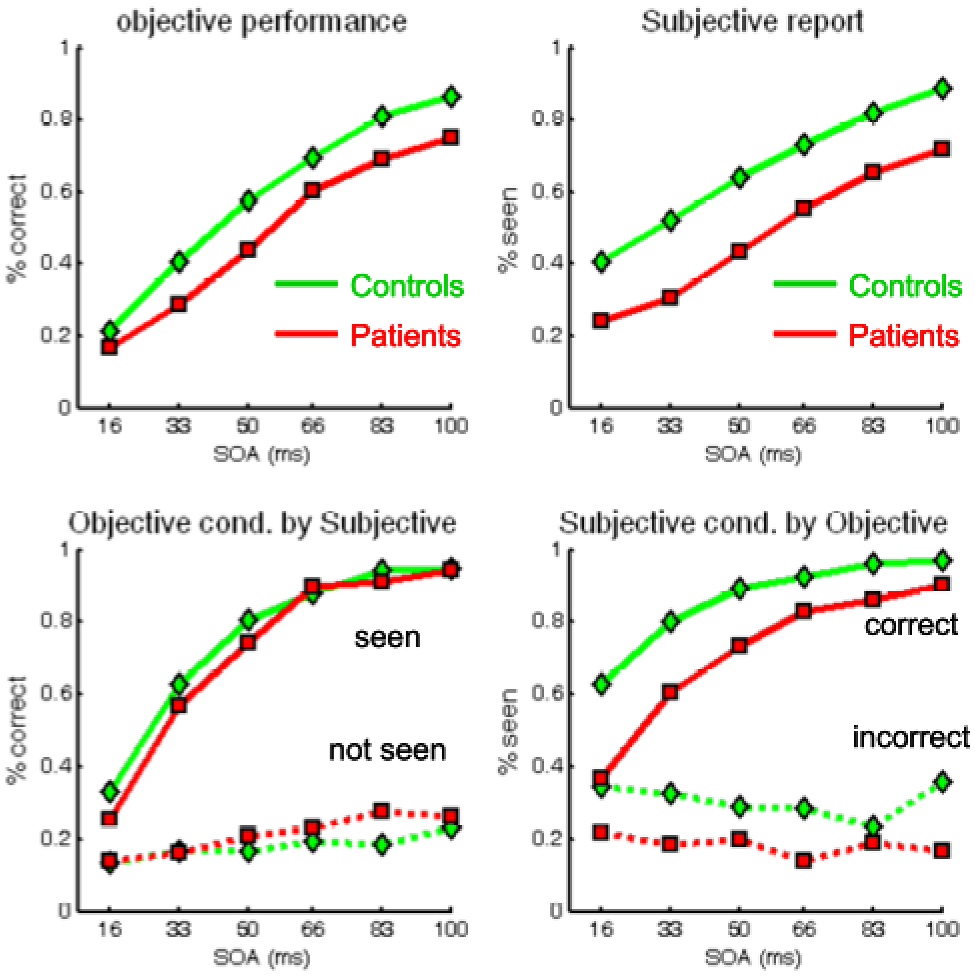

Figure B-1. The model fit of DC$_{accum}$ to patient and control data originally described in Del Cul et al (2009). All displayed curves are generated by the model (i.e. actual data from patients and controls, to which the model was fit, are not displayed in this figure). Reproduced from Figure S3 of Del Cul et al.

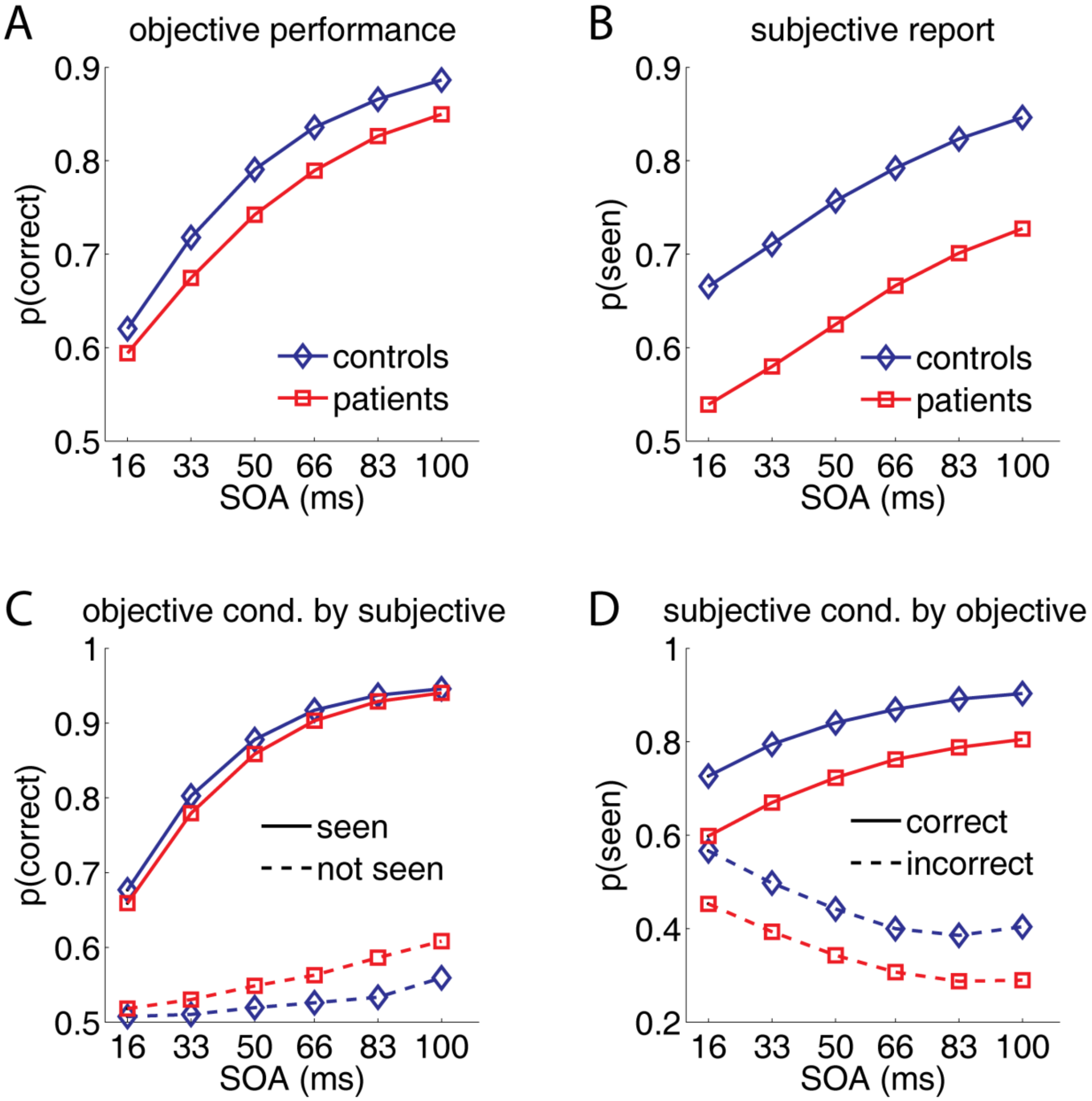

Figure B-2. Adaptation of DC$_{accum}$ to the 2-stimulus case. We adapted the model described in Del Cul et al (2009) to apply to tasks with 2 response alternatives rather than 10, but otherwise adhered to the model specifications described in the supplemental material in Del Cul et al (see Appendix B for details). Here we show that our 2-stimulus implementation of the model can produce results similar to those found by Del Cul et al in the 10-stimulus case (Figure B-1).

$\underline{DC_{SDT}}$

Next, we investigated whether DC_{SDT} could produce a pattern of results similar to those of DC_{accum}, as originally found by Del Cul et al (Figure B-1) and reproduced in our adaptation to the 2-stimulus case (Figure B-2). DC_{SDT} was implemented as described under “Independent Dual Channel” in the “Model descriptions” section of the manuscript, with minor differences.

- Because we needed to simulate 6 SOA levels rather than 8 as in the manuscript, we required only 6 levels of $\mu_{diff\ C}$ and $\mu_{diff\ U}$.
- To simplify modeling, we assumed that stimulus identification responses arising from the unconscious channel were unbiased, and so we set $c_U = 0$.
- Because we needed to simulate 2 levels of stimulus visibility (“not seen” and “seen”) rather than 4 as in the manuscript, we required only 2 levels of the decision criteria for subjective report, c_C. To further simplify the model, we also assumed that these two criteria were set symmetrically about 0, the point of unbiased responding, so that we needed to specify only one value of c_C.

We adopted a heuristic approach to converge upon an appropriate set of parameter values, as follows.

First, we computed the mean values of p(correct) at each SOA level for the patient and control groups in the DC_{accum} simulation (Figure B-2). We converted these to *d’* values by assuming unbiased responding using the formula *d’* = 2*z[p(correct)], where z is the inverse of the normal cumulative distribution function (Macmillan & Creelman, 2005). Call these *d’* values d'_{SOA}. For the simulated control group, we set $\mu_{diff\ C,\ control} = d'_{SOA}$ at each SOA level. This ensured that, for the control group, p(correct) as a function of SOA resembled the corresponding curve generated by DC_{accum}. The values of $\mu_{diff\ C,\ control}$ thus obtained were 0.54, 1.03, 1.45, 1.78, 2.04, and 2.23.

To further constrain the search space, we ensured that the following relationships between parameter values were enforced at each level of SOA:

$\mu_{\text{diff U, control}} = w_1 * \mu_{\text{diff C, control}}$

$\mu_{\text{diff C, patient}} = w_2 * \mu_{\text{diff C, control}}$

$\mu_{\text{diff U, patient}} = \mu_{\text{diff C, control}}$

The first constraint characterizes sensitivity in the unconscious channel as some constant fraction of the sensitivity in the conscious channel for controls. The second constraint characterizes sensitivity in the conscious channel for patients as some constant fraction of the sensitivity in the conscious channel for controls. The third constraint is that sensitivity in the unconscious channel for patients and controls is equal, mirroring the assumption of equally noisy unconscious channels posited by Del Cul et al in their model fitting.

We manually adjusted values of w_1, w_2, $c_{\text{C, control}}$, and $c_{\text{C, patient}}$ so as to produce a pattern of results resembling that produced by DC_{accum}, as depicted in Figure B-1. We found that setting $w_1 = .14$, $w_2 = .9$, $c_{\text{C, control}} = .45$, and $c_{\text{C, patient}} = .65$ accomplished this result well (Figure B-3). The key patterns in the data originally found in Del Cul et al are reproduced by the DC_{SDT} model, and the specific numerical values for all plotted curves produced by DC_{SDT} (Figure B-3) closely approximate the numerical values for the curves produced by DC_{accum} (Figure B-2).

Thus, not only do DC_{accum} and DC_{SDT} share core assumptions about the computational architecture underlying perceptual decision making, but they also can generate (and therefore account for) similar patterns of behavioral data. We therefore are justified in using DC_{SDT} as a model encompassing the core assumptions of DC_{accum} when comparing SDT models in the main manuscript.

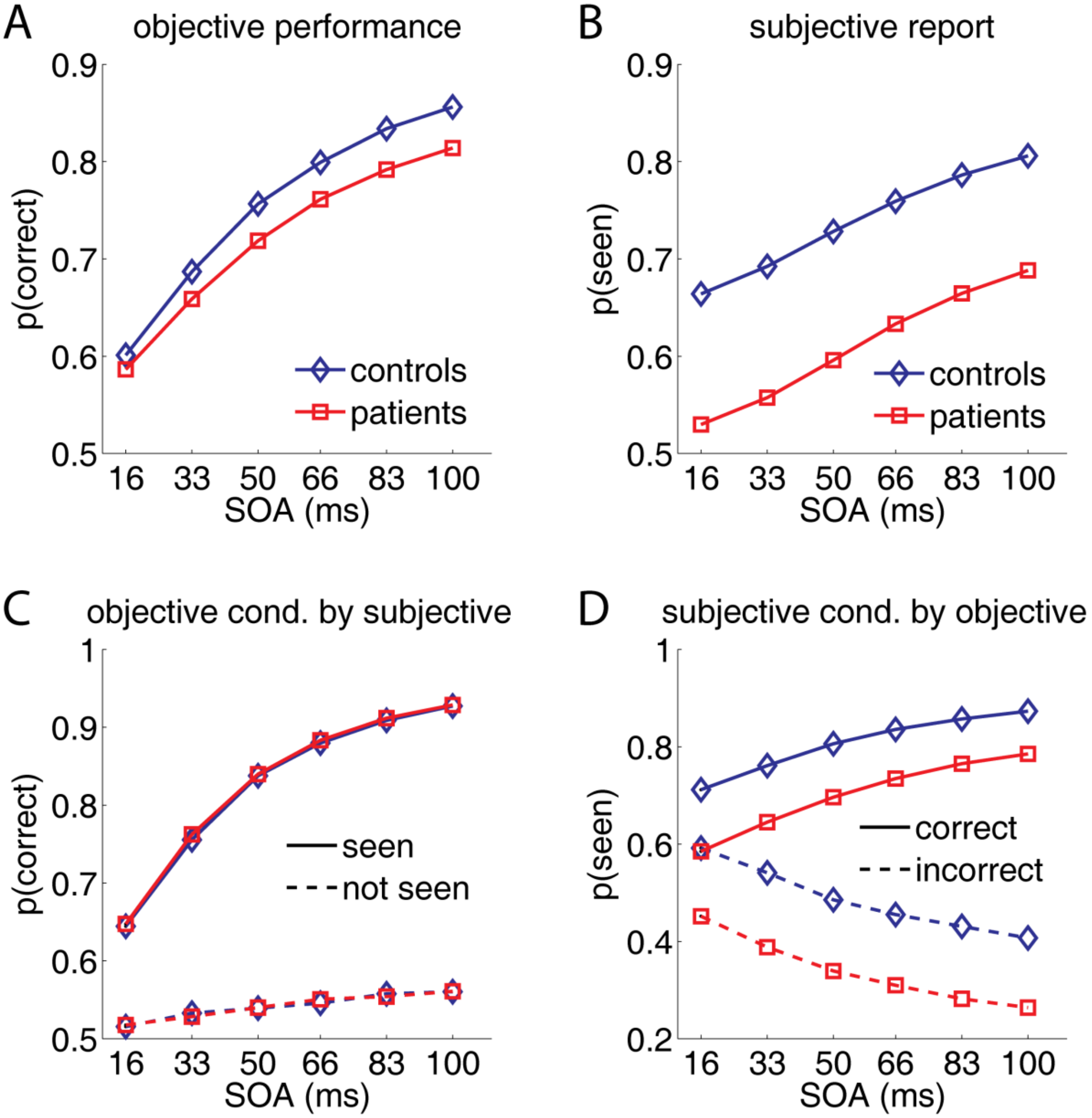

Figure B-3. Reproduction of DC_{accum} results with DC_{SDT}. Our SDT implementation of the independent dual channel model, DC_{SDT}, was able to produce results closely matching those produced by the accumulator model of dual channel processing, DC_{accum}, posited by Del Cul et al (2009) (compare this figure with Figure B-2).