Confidence Ratings and Non-Diagnostic Information – A Theoretical Account of Confidence for Recognition Memory and Perceptual Discrimination Tasks

by
Amelia Turner Kohl
Bachelor of Arts (Psychology)(Hons)
School of Psychological Sciences | Faculty of Health

Submitted in fulfilment of the requirements for the Doctor of Philosophy (Psychology) University of Tasmania
June, 2022

Statement of Co-Authorship

The following people and institutions contributed to the publication of work undertaken as part of this thesis:
Candidate – Amelia Turner Kohl, University of Tasmania
Author 1 – Dr James Sauer, University of Tasmania
Author 2 – Dr Matthew Palmer, University of Tasmania
Author 3 – Prof Andrew Heathcote, University of Newcastle
Author 4 – Talira Kucina, University of Tasmania
Author 5 – Jasmin Brooks, University of Tasmania

Contribution of work by co-authors for each paper:

PAPER 1: Located in Chapter 3
Kohl, A. T.., Sauer, J. D., & Palmer, M. (2022). *The Theoretical Basis of Recognition-Ratings: Inferential vs Psychophysical Approaches to Confidence*. Manuscript in preparation.
Author contributions:
Conceived and designed experiment: Candidate, Author 1, Author 2
Performed the experiments: Candidate
Analysed the data: Candidate, Author 1
Wrote the manuscript: Candidate, Author 1

PAPER 2: Located in Chapter 4
Kohl, T. A., Sauer, J. D., Palmer, M., Brooks, J., & Heathcote, A., (2022) *The Effects of Non-Diagnostic Information on Confidence and Decision Making*. Manuscript under review.
Author contributions:
Conceived and designed experiment: Candidate, Author 1, Author 2, Author 3
Performed the experiments: Candidate, Author 5
Analysed the data: Candidate, Author 3
Wrote the manuscript: Candidate, Author 1, Author 2, Author 3

PAPER 3: Located in Chapter 5
Kohl, T. A., Kucina, T., Sauer, J. D. & Palmer, M., (2022) *A Rating-Only Approach to Perceptual Discrimination*. Manuscript in preparation.
Author contributions:
Conceived and designed experiment: Candidate, Author 4, Author 1, Author 2
Performed the experiments: Candidate, Author 4
Analysed the data: Candidate, Author 4, Author 1
Wrote the manuscript: Candidate, Author 4, Author 1, Author 2

Endorsement

We, the undersigned, endorse the above stated contribution of work undertaken for each of the published (or submitted) peer-reviewed manuscripts contributing to this thesis:

Signed:

	Amelia Turner Kohl	Dr James Sauer	Prof Lisa Foa
	Candidate	Supervisor	Head of School
	School of Psychological Sciences	School of Psychological Sciences	School of Psychological Sciences
	University of Tasmania	University of Tasmania	University of Tasmania
Date:	09/03/2021	09/03/2021	10/03/2022

Declarations

This thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

I declare that this is my own work and has not been submitted in any form for another degree or diploma at any university of other institution of tertiary education, aside from the following exceptions: (1) the data used in Experiment 1 of Chapter 3, and (2) the data used in Experiment 1 of Chapter 4. Both of these experiments were conducted as a requirement for Honours theses within the School of Psychological Sciences at the University of Tasmania. The data were re-analysed and interpreted for the purpose of this Doctoral Thesis. Information derived from the published or unpublished work of others has been duly acknowledged in the text and a list or references is given.

The research associated with this thesis abides by the international and Australian codes on human and animal experimentation, the guidelines by the Australian Government's Office of the Gene Technology Regulator and the rulings of the Safety, Ethics and Institutional Biosafety Committees of the University. Ethics Approval No/s H0018677 and H0018198.

Signed: Amelia Turner Kohl

Date: 03/03/2022

Acknowledgements

This thesis would not have come to fruition without the support and guidance of an incredible team. First, I would like to thank my primary supervisor, Dr Jim Sauer, for making all of this possible. Your passion for good science is surpassed only by your knowledge of cognitive psychology, and I have been so lucky to have been able to learn from you over the past five years. To my co-supervisor, Dr Matt Palmer; your feedback and attention to detail have improved the overall calibre of my research. I would also like to extend my sincerest of thanks to Prof Andrew Heathcote; I am so grateful for the time and effort that you have put into our research. Finally, to Matthew Gretton; without your programming know-how (and endless offers of help), some of the experiments detailed in this thesis simply never would have existed.

I have been lucky enough to have been surrounded by a group of likeminded researchers in the Tas Cog Lab, as well as the broader Utas psychology PhD cohort, who have motivated (and at times commiserated) with me throughout this experience. I would like to take the time to explicitly thank Talira Kucina, who has become not just one of my most valued collaborators, but also one of my most valued friends.

A massive thank you is in order to all my friends and family who have supported me throughout the past four years; particularly to my partner, Angus, who has helped me stay calm and collected as we approached the pointy end of the submission processes. I would also like to especially thank my parents, Penny and Brian, for always making me feel as though I could achieve anything I set my mind to. Without your support and guidance, I would not be where I am today.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Abstract

In a variety of domains, ratings have been used (in the absence of a binary yes/no or old/new judgement) to index recognition. More recently, this approach has been implemented in the face recognition and eyewitness identification literature. Such an approach requires participants to provide a confidence rating (e.g., from 0-100%; Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012) as to whether they have viewed a specific face before, without asking for a binary identification judgement. Evidence suggests that such ratings can be used to indicate the likelihood that an image or lineup member has been seen before, and may even provide a more sensitive index of recognition than categorical responses (e.g., Brewer, Weber, Wootton & Lindsay, 2012; Cleary & Greene, 2000; Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012). However, there has been little direct investigation of the theoretical mechanisms that underlie these ratings.

Research into the mechanisms that shape Feelings of Knowing (judgements that non-recalled information might be cued in a later recognition task; Koriat, 1997) and Judgements of Learning (assessments of the likelihood of future recall for studied information; Koriat, 1993) suggest that metacognitive judgements can be shaped by non-memorial cues, such as heuristics pertaining to perceptions of fluency at test, or naïve theories about how the learning/encoding conditions are likely to affect later recall or recognition. Such heuristic processes reflect perceptions of the experiences associated with encoding and retrieval, rather than reflecting evidence gleaned directly from memory. Thus, they are vulnerable to systematic distortions. Conversely, theories with their origins in the domain of psychophysics suggest that confidence indexes stimulus discriminability and/or memory strength (e.g., signal detection theory, Bernbach, 1971; Egan, Schulman, & Greenberg, 1959; Green & Swets, 1966; or accumulator models using a balance of evidence mechanism to account for confidence, Vickers, 1970; Van Zandt, 2000).

The primary focus of this thesis is to test the extent to which rating-only judgements (i.e., confidence in recognition or perceptual discrimination) are reflective of the physical properties of the stimuli (as per psychophysical models of confidence) or shaped by heuristic processes (i.e., as per inferential models of metacognition; appraisals of the reliability of, or our confidence in, our own cognitions). Specifically, we are interested in whether confidence reflects the diagnosticity of the information at test (i.e., the extent to which the accumulated evidence favours one response over another) or simply the amount information available at test. Further, we are interested in whether these effects are consistent across both recognition memory tasks and perceptual discrimination tasks. While psychophysical models of confidence assume that effects observed in psychophysics should generalise to recognition tasks, the reliance on memory adds noise to the process, opening the door for more inferential processes associated with fluency etc.

Chapter 3 reports three experiments testing the theoretical mechanisms that shape recognition confidence ratings. Experiment 1 uses a face recognition paradigm, Experiment uses a landscape/house recognition paradigm to test whether effects from Experiment 1 generalise beyond faces to other complex stimuli that might be less reliant on holistic processing, and Experiment 3 uses a perceptual discrimination task asking participants to indicate their confidence that a dynamic grid was primarily blue or primarily orange. Extending upon findings by Busey, Tunicliff, Loftus and Loftus (2000) suggesting that the presence of non-diagnostic information at test can inflate confidence, we in turn implemented a manipulation of non-diagnostic information. In Experiments 1 (N = 60) and 2 (N = 113), this was operationalised through the use of full and partial images. To elaborate, participants may have viewed the top half of a face (or house/landscape) during the study phase, but been presented with the full version of the image during the test phase. In this example, we would view the additional bottom-half of the image as “non-diagnostic” information, as it was not

present at encoding. In Experiment 3 (N = 86), participants were asked to indicate their confidence as to whether a dynamic grid was predominantly blue or predominantly orange (i.e., "Indicate your confidence that the flashing grid is predominantly **orange**", with the target colour differing between blocks of trials). Non-diagnostic information was included through the addition of white pixels to the grid.

To find support for an inferential model of confidence, we would expect to see an increase in confidence associated with the addition of non-diagnostic information because, according to inferential models, metacognitive judgements are based at least in part on the volume of information available at test regardless of the accuracy or diagnostic value of that information. To find support for psychophysical models, we would need to see no changes in confidence associated with the addition of non-diagnostic information at test, as the amount of diagnostic information is unaffected by the addition of non-diagnostic information. Instead, what we discovered across all three experiments was that confidence decreased when the presence of non-diagnostic information increased. These findings align with a little-used psychophysical theory: Baranski and Petrusic's (1998) doubt-scaling model, which suggests that confidence is inversely proportionate to the amount of non-diagnostic information present at the time of decision.

Having seemingly found support for Baranski and Petrusic's (1998) doubt-scaling model – a model first formally proposed in 1998 and then largely ignored – we further explored the nature of the relationship between non-diagnostic information and confidence. Thus, we conducted a series of perceptual tasks (Chapter 4) that allowed us to examine the effects of different levels of non-diagnostic information on confidence, reaction time (RT) and accuracy in a highly controlled experimental paradigm. In Experiment 1 (N = 56), participants completed 720 trials, in which they were presented with a dynamic grid consisting of flashing blue, orange and white pixels and asked to indicate whether the grid

was predominantly blue or orange (using a response scale ranging from low confidence blue, moderate confidence blue, high confidence blue, low confidence orange etc.), with the white pixels constituting non-diagnostic information. Unlike the experiments detailed in Chapter 3, this study asked participants to make a simultaneous confidence/decision judgment rather than a standalone confidence judgement to allow us to observe the effects of non-diagnostic information on accuracy, response time (RT), and confidence. The results showed a reduction in both confidence and accuracy associated with an increase in the proportion of non-diagnostic information, together with an increase in RTs. Further, CAF analyses (which plot accuracy as a function of RT) showed that as the proportion of non-diagnostic information increased, the accuracy of fast responses decreased in comparison to the accuracy of slow responses.

To determine whether non-diagnostic information has a direct effect on accuracy (as opposed to the decrease in accuracy being a by-product of the simultaneous confidence judgement and decision judgment), we ran a second experiment (N = 21). This time, participants viewed the same stimuli, but were asked only to indicate whether they believed the dynamic grid to be majority blue or majority orange. We found a pattern of results similar to those for Experiment 1 regarding the relationship between non-diagnostic information and accuracy (RTs). Again, CAFs showed that the accuracy of fast responses decreased in comparison to the accuracy of slow responses as the proportion of non-diagnostic information increased. Much like the experiments outlined in Chapter 3, these results add to the existing literature by demonstrating a negative relationship between non-diagnostic information and confidence. In line with predictions made by Baranski and Petrusic, the current study demonstrated a negative relationship between non-diagnostic information and accuracy, as well as a general slowing of RTs as the proportion of non-diagnostic information increased.

Chapter 5 reports two experiments that tested whether a rating-only, confidence approach (similar to that used in Chapter 3) could be used in an applied visual search task to indicate the likelihood of a target being present. In Experiment 1 (N = 24), we confirmed that a rating-only approach could be used in a basic perceptual paradigm to indicate the likelihood of a dynamic grid being majority blue or majority orange (the stimuli was similar to that used in Chapters 3 and 4, but without the additional white pixels). We found that participants were well calibrated in their responding, indicating that they were able to use the rating-only scale to convey the likelihood that the stimuli is majority blue or majority orange. In Experiment 2 (N = 99) we tested whether this could be extended to an applied visual search task: Searching luggage X rays for evidence of a weapon. Specifically, we tested whether confidence ratings could discriminate trials containing a weapon from those that did not, and whether using ratings could mitigate the low prevalence effect (where shifts in decision criterion mean rarely-occurring targets often go undetected). The results showed that despite increases in overconfidence associated with lower prevalence rates, a rating-only approach effectively discriminated target present from target absent trials, even when target prevalence was low.

In sum, the experiments outlined here expand the existing literature on confidence judgements in two main ways. First, our research suggests that confidence judgements (whether accompanied by a decision judgment or not) are influenced by the presence of non-diagnostic information, providing support for Baranski and Petrusic's (1998) doubt-scaling model of confidence. Second, our findings suggest that a rating-only approach can be used in applied perceptual discrimination tasks to indicate the likelihood of a target being present in a complex array.

Table of Contents

List of Figures

Chapter V

List of Tables

Chapter III

Chapter V

Chapter I

An Overview

Chapter I

Confidence ratings have long been used in the decision-making literature to index perceived accuracy (e.g., Angell, 1907; Baranski & Petrusic, 1998; Brewer & Wells, 2006; Gigerenzer, Hoffrage & Kleinboelting, 1991; Henmon, 1911; Juslin, Winman & Olsson, 2000; Palmer, Brewer, Weber, & Nagesh, 2013). Examples of confidence being used in the psychophysics literature date back over a century, with early work focussing on participants' confidence in discriminating, for example, the longer of two lines (e.g., Henmon, 1911). Confidence is also commonly used to index perceived accuracy for recognition memory tasks, with such research demonstrating a positive relationship between confidence and accuracy, at least when participants make a positive recognition judgement (e.g., Brewer & Wells, 2006; Gigerenzer, et al., 1991; Juslin et al., 2000; Palmer et al., 2013; Sauer et al., 2010; Sauer & Brewer, 2015; Wixted & Wells, 2017; Weber & Brewer, 2004).

The relationship between confidence and accuracy has important applied implications. For example, when a witness to a crime later identifies a lineup member as the culprit, the confidence with which the identification is made has been proposed as a potential tool for evaluating the likely accuracy of that decision. The relationship between confidence and accuracy has even been acknowledged in the legal field. In the ruling for *Neil v. Biggers* (1972), the U.S. Supreme Court identified an eyewitness' confidence as being one of the key criteria for assessing eyewitness identification evidence. The National Academy of Sciences Report (2014) and the Technical Working Group for Eyewitness Evidence (1999) also recommend that eyewitness confidence be presented alongside an identification judgement in court. Surveys of those involved in the legal process demonstrate that police, lawyers, and jurors all believe that there is a positive relationship between confidence and accuracy (Deffenbacher & Loftus, 1982; Potter & Brewer, 1999), a point that is further strengthened by mock juror studies, which consistently show that eyewitness confidence influences

perceptions of the defendant's guilt (e.g., Bradfield & Wells, 2000; Brewer & Burke, 2002; Cutler, Penrod & Dexter, 1990; Sauer, Palmer, & Brewer, 2017). Recent work has even shown that confidence can be used in the absence of a binary decision to communicate the likelihood that a face has been seen before in a recognition memory task (Brewer et al., 2012; 2020; Cleary & Greene, 2000; Sauer et al., 2008; 2012), with such ratings proving to be less susceptible to factors that impair discriminability for binary identification judgements, including retention interval and distinctiveness (Brewer et al., 2020; Bruer et al., 2017; Sauer et al., 2012).

Though eyewitness identification provides one clear example of an applied context where the relationship between confidence and accuracy is important, early theoretical accounts of confidence originated as a means of explaining confidence in perceptual judgements. For example, early accumulator models of confidence were tested using perceptual tasks in which participants would have to determine which of two lines were longer (Vickers, 1970), or the orientation of a stimulus consisting of a group of lines (Smith & Vickers, 1988). Despite the psychophysical origins of such theories having little in common with the recognition memory tasks to which they are now applied, little has been done to directly compare whether the underlying mechanisms that shape confidence in perceptual judgements also shape confidence in recognition memory in the same way. Admittedly, in general, predictions drawn from these psychophysical theories are borne out in confidence-accuracy data from basic (e.g., face recognition) and more applied (e.g., eyewitness identification) recognition paradigms. However, there are alternative frameworks for understanding the bases of metacognitive judgments (i.e., appraisals of the reliability of, or our confidence in, our cognition processes). These theories often suggest metacognitive judgments (e.g., judgments of learning and feelings of knowing) are based on inferences drawn from the phenomenology of the task experience (e.g., with confidence being positively

associated with the fluency with which information is processed, or retrieved from memory, or the amount of information retrieved in response to a memory probe).

The current research aimed to bring together theory that originated in the psychophysics literature (i.e., psychophysical models of confidence) and theory that originated in the broader metacognitive literature (i.e., models that propose a more inferential basis for confidence; hereafter referred to as "inferential models" for convenience) to allow us to further explore the mechanisms that underlie confidence ratings in both recognition memory and perceptual discrimination tasks. Thus, we aimed to answer two overarching questions: First, to what extent are confidence responses reflective of the physical properties of the stimuli (i.e., psychophysical models of confidence) compared to the extent to which they are shaped by inferential processes (i.e., inferential models of confidence)? Second, what scale is best for recording confidence ratings in the absence of a binary decision judgement?

While Chapter 2 provides some grounding in the relevant frameworks that inform this research. Given the nature of a PhD by publication (i.e., where Chapters 3-5 present multi-study manuscripts either in preparation, or submitted, for review), there will be some repetition and/or elaboration of content in subsequent chapters.

Chapter 3, which consists of a series of three experiments examining the role of non-diagnostic information on confidence ratings, addresses both of our key questions. Based on the assumption of inferential models of metacognition that the presence of additional information at test should inflate confidence ratings (regardless of the utility of said information; see Koriat, 1993), as well as findings by Busey, Tunnicliff, Loftus and Loftus (2000) suggesting that the presence of additional, non-diagnostic information at the time of decision may inflate confidence ratings relative to accuracy in a recognition memory task (this will be explored in more detail in Chapter 2), we tested how manipulating the presence of non-diagnostic information affected confidence in a face recognition task (Experiment 1),

a house/landscape recognition task (Experiment 2), and a perceptual discrimination task (Experiment 3). For this study, we were interested specifically in confidence given in the absence of a binary yes/no or old/new decision judgement, referred to as "recognition-ratings" (see Chapter 2 for a more detailed rationale). Experiments 1 and 2 use a combination of "full" and "partial" images to operationalize non-diagnostic information, with trials of interest being those in which a participant studied a partial image, but were tested with the full version of the same image (making the additional half of the image presented at test "non-diagnostic" to the decision). Experiments 1 and 2 also included a between-subjects manipulation of scale-type, with half of the participants using a fine-grained numerical scale (consisting of 11 buttons labelled 0%, 10%, 20% … 100%), and half using a coarse grained verbal scale (consisting of "Low Confidence", "Moderate Confidence", "High Confidence"), allowing us to observe the influence of scale type on confidence-ratings. While there is an obvious confound between the two scales (by combining verbal vs numerical and fine-grained vs coarse-grained), we were interested in these measurement scales in particular due to their utility in terms of confidence accuracy characteristic analysis (CAC; Mickes, 2015). In short, CAC analyses allow for confidence to be plotted against accuracy on a curve. Where confidence is generally collected on a 0-100% scale, it is then typically collapsed into 3 categories (e.g., Carlson et al, 2016; Mickes, 2015; Sauerland et al., 2016), sparking our interest in whether collecting data in three categories to begin with would provide a more direct measure of confidence. Had patterns of results differed as a function of scale-type, we would have teased apart these factors in follow up experiments, however no effect was observed. Experiment 3 deviated from the use of recognition memory tasks, instead testing the effects of non-diagnostic information in a colour discrimination task. Participants provided confidence ratings pertaining to whether a dynamic grid consisting of blue, orange, and (sometimes) white pixels was predominantly blue, or predominantly orange (with white

operating as non-diagnostic information). Shifting to this perceptual discrimination not only allowed us greater control over the amount of non-diagnostic information present in each trial (and a cleaner understanding of the effects of non-diagnostic information on confidence), it also allowed us to directly test the extent to which mechanisms that underlie confidence in recognition memory tasks generalise to perceptual discrimination tasks.

To foreshadow the results detailed in Chapter 3, we found evidence supporting Baranski and Petrusic's (1998) doubt-scaling model of confidence. Given how little focus in the literature has been paid to the doubt-scaling model, Chapter 4 aims to further clarify the nature of the relationship between non-diagnostic information, and directly test predictions drawn from the doubt-scaling model, by using a similar perceptual discrimination task to that of Chapter 3, Experiment 3. While the aforementioned experiment asked participants to provide a confidence-only recognition rating, Experiment 1 of Chapter 4 required participants to provide a simultaneous decision judgement and confidence rating (e.g., "Low confidence blue", "Moderate confidence blue", "High confidence blue", "Low confidence orange" etc.). This allowed us to observe the effects of non-diagnostic information on not only confidence, but also on accuracy and reaction time (RT). Experiment 2 differs in that it asks participants for a decision judgment (i.e., majority blue/majority orange) only, forgoing the confidence judgement. This will clarify whether the effects on accuracy and RT observed in Experiment 1 persist when participants are no longer considering their confidence in their decision.

Pivoting from the focus on non-diagnostic information, the research reported in Chapter 5 aimed to determine whether a rating-only approach (similar to the recognition-rating approach used in Chapter 3) could be used in an applied visual search task to (a) indicate the likelihood of a target being present in a complex visual array, and (b) mitigate "low prevalence effects" (i.e., a criterion-related failure to identify targets when target-prevalence is low; Wolfe et al., 2007) that have been previously observed in visual search

data. By forgoing a decision judgement in favour of just a confidence rating (indicating the likelihood that a target was present), we hoped to bypass bias effects on response criteria that sometimes lead to errors in categorical decision making (see Chapter 2 for a more comprehensive account). Thus, we believed that a rating-only scale may provide a more informative measure of target presence in both low-target-prevalence and high-target-prevalence trials than a binary identification decision coupled with a retrospective confidence rating, and that such an approach might help avoid "miss" errors that commonly result from low-prevalence effects inducing decision-makers to set overly conservative response criteria. Finally, Chapter 6 provides a summary of the key findings from the experimental chapters.

Chapter II

Theories of Confidence

Chapter II

As mentioned in Chapter 1, there are two dominant schools of theory that are used to account for confidence metacognitive judgements pertaining to memory-related tasks: psychophysical models, that generally suggest that confidence is reflective of the amount of evidence in favour of the chosen response option (e.g., Signal detection theory; Green & Swets, 1966; Accumulator models; Vickers, 1970), and inferential models, that suggest that confidence is shaped by inferential cues (that can be independent of stimulus discriminability; e.g., Koriat, 1993; 1997). Where psychophysical models have typically been relied upon to explain confidence in recognition, inferential models have been relied upon to explain metacognitive judgments (such as judgements of learning; JOLs and feeling of knowing; FOKs; Koriat 1993; 1997). The question raised when considering recognition-ratings (i.e., confidence ratings made in the absence of a binary recognition decision) as opposed to a more traditional measurement is as follows: to what extent do theories developed to account for confidence in the context of binary decision making explain recognition-ratings? And could theories developed to account for other similar metacognitive judgements (i.e., JOLs and FOKs) provide a more accurate picture of the processes driving such ratings? This chapter explores the extent to which recognition-ratings may be explained by psychophysical models of confidence versus inferential models of confidence, and how the predictions put forward by each school of theory would differ for recognition-ratings compared to a more traditional measure of confidence.

A Rating-Only Approach to Assessing Recognition

Confidence ratings have shown to provide useful evidence of recognition even when provided in the absence of a decision judgement (Brewer et al., 2012; 2020; Cleary & Greene, 2000; Sauer et al., 2008; 2012). In fact, early signal detection-based work often favoured confidence-based rating scales (e.g., a 6-point scale ranging from “sure old” to “sure

new", as used by Ratcliff, McKoon, & Tindall, 1994) over binary yes/no or old/new judgments. A similar rating-only approach to assessing recognition, which we will refer to as "recognition-ratings", has also been applied in the eyewitness identification literature, where ratings have been found to be reflective of the likelihood that a lineup member had been seen before (Brewer et al., 2012; 2020; Sauer et al., 2008; 2012).

It has been suggested that recognition-ratings may provide a more diagnostic account of the evidence accumulation process by reducing the influence of bias effects on decision criteria, and more directly indexing the evidence in favour of a given response option (Sauer et al., 2008. See page 17 for a more thorough explanation). For example, a security screener examining a luggage x-ray where threats are unlikely may set a more conservative criterion (i.e., require more evidence to identify a threat due to their rarity), knowing that the chances of a target-present trial are objectively low. Similarly, a witness viewing a lineup may set a conservative response criterion, not wanting to mistakenly implicate an innocent lineup member. In such cases, a witness may see the culprit in the lineup, experience a sense of familiarity or recognition, but ultimately decide that this sense of familiarity in insufficient to support an identification. In both cases, this can lead to costly errors, with targets being missed. If recognition-ratings are truly an unbiased reflection of evidence accumulation (i.e., less affected by factors operating on criteria placement), then they may provide a more sensitive representation of diagnostic evidence. Sauer et al. (2008) found that participants were able to effectively communicate whether they had studied a face previously by responding using a 0-100% confidence scale without an accompanying old/new judgement. Additionally, participants who provided a rating-only measure of confidence provided more diagnostic information than those in a control group (who provided a binary recognition judgement without an accompanying confidence rating). Moreover, when participants were asked to provide a recognition-rating for *every lineup member* rather than a single

identification decision (either identifying a lineup member as the culprit or rejecting the lineup as whole), patterns of confidence ratings could be used to identify a target among foils in a lineup more often than when a standard identification procedure (i.e., a binary decision judgement) was used.. Further, performance in target absent trials was comparable between the recognition-rating and binary conditions, suggesting that the increase in accurate identifications was not solely due to an increased propensity to "choose" (as that would also manifest in an increase in false identifications in the target absent condition). It has also been shown that factors that are known to impair discriminability for binary identification judgements, such as retention interval and distinctiveness, have less of an effect on confidence ratings (Sauer et al., 2012). Similar findings have been reported by Brewer et al. (2012) and Bruer et al., (2017).

In a scenario where a witness views a lineup and sees two equally plausible lineup members, they may end up rejecting the lineup, opting not to identify a lineup member because they cannot choose between the two plausible options and do not want to guess. However, if one of the favoured lineup members is in fact the suspect, indicating that this lineup member is one of two equally plausible alternative may still be diagnostic of suspect guilt. In this scenario, the fact that there is another highly plausible candidate need not undermine the evidence against the suspect, assuming that the eyewitness demonstrates an ability to discriminate between lineup members on the whole. Thus, Brewer et al. (2020) tested the utility of multiple recognition-ratings in an eyewitness identification paradigm as an alternative to a traditional lineup task, with participants instead providing recognition-ratings (i.e., their confidence from 0-100% that the lineup member was in fact the perpetrator) for every member of the lineup.

Overall, there was a positive relationship between confidence given to a lineup member and likelihood of guilt. Further, likely guilt varied systematically with the extent to

which the suspect was favoured over the next-best alternative. Of particular interest in this study were the "max" confidence ratings; with these being the highest confidence rating that each individual witness gave to any one (or more) lineup member(s), indicating that that said lineup member(s) was the closest match to their memory of the perpetrator. Brewer et al., found that, compared to a traditional lineup paradigm (i.e., a lineup including an identification decision), the recognition-rating approach provided a better discriminated guilty from innocent suspects in scenarios in which participants provided max ratings for multiple lineup members, as well as when participants provided the max rating for a foil and gave the second highest rating to the suspect. This approach to interpreting recognition-ratings differs from early work in the area, which imposed decision criterion upon recognition ratings based on individual participants responses (e.g., Koriat & Goldsmith, 1996; Sauer et al., 2008; Brewer et al., 2012), and further highlights the potential advantages of ratings-based (cf. categorical) approaches to indexing recognition.

Thus, there is a growing body of literature demonstrating that (a) confidence can discriminate targets from foils in basic face recognition and more complex lineup tasks, even in the absence of a categorical identification/recognition decision, (b) that these ratings might be less sensitive than categorical responses to factors that impair discriminability, and (c) that these ratings might be more *informative* than a categorical response, offering diagnostic information of suspect guilt in cases where a categorical decision may simply lead to a lineup rejection. However, researchers have not really explored the theoretical bases for these ratings. The work is grounded in, and motivated by, models that originated in the psychophysics literature (e.g., Signal detection theory; Green & Swets, 1966; Accumulator models; Vickers, 1970 etc.) that tie confidence to stimulus discriminability. However, although researchers have emphasized that a variety of psychophysical models predict that these ratings will have diagnostic value, researchers have not interrogated the applicability of

individual theoretical frameworks (i.e., testing predictions derived from different theoretical frameworks to see which best account for these ratings).

The extant literature suggests the applied value of ratings-based approaches to measuring recognition. However, if a rating-only measure of recognition is to be used in applied (or experimental) settings, it is important to understand the theoretical mechanisms that shape them, and the types of evidence that contribute to them. Such understanding would provide insight into scenarios in which these ratings may no longer index the diagnostic evidence on which we hope decisions would be based, and may instead be vulnerable to systematic distortions. There are two main schools of theory that contribute to our current understanding of traditional measures of confidence and associated metacognition judgements: psychophysical models (e.g., those derived from Signal Detection Theory; Green & Swets, 1966; or accumulator models coupled with a balance of evidence hypothesis; Smith & Vickers, 1998: Vickers & Lee, 1998) and inferential models (e.g., the cue utilization model; Koriat, 1997).

Psychophysical Models of Confidence

Psychophysical models of confidence assume that confidence is, primarily, reflective of the physical properties of the stimulus, such as the length of a line, or the brightness of dots (e.g., Signal detection theory; Green & Swets, 1966; Accumulator models; Vickers, 1970). According to these models, confidence indexes stimulus discriminability; the intensity of a stimulus, or the difference in intensity between two or more stimuli. Psychophysical models can be divided (coarsely) into signal detection-based models (which generally hold that confidence and decision arise concurrently from the same process) and accumulator-based models (which generally hold that confidence, though indexing the evidence upon which the decision is based, emerges post-decision). As mentioned previously, these models originated in the domain of psychophysics, where they were developed to account for

decision making and confidence for perceptual discrimination tasks. They have since been adapted to provide a framework for understanding confidence (and decision making) in recognition tasks.

Signal Detection Models

Signal detection theory (SDT; e.g., Bernbach, 1971; Egan, Schulman, & Greenberg, 1959; Green & Swets, 1966) breaks the processes behind decision making into two components: sensitivity and bias. While sensitivity refers to the ability to discriminate targets from foils, bias refers to a person's general likelihood to categorise any given stimuli as a target, foil or lure (Green & Swets, 1966; Macmillan & Creelman, 1991). To use a face recognition task as an example: a participant studies a series of faces, after which they are presented with a series of test trials. Each test trial presents a single face, and asks the participant to determine whether that face was part of the study-set. Information from the face-stimulus produces a sense of stimulus intensity (e.g., strength of recognition, brightness etc.) that falls somewhere along a continuum from weak evidence to strong evidence. The participant sets a decision criterion somewhere along this continuum, with stimulus intensity that exceeds the criterion triggering an affirmative response (e.g., an "old" response; indicating that the face was studied previously). A failure to exceed the criterion, on the other hand, generates a negative response (e.g., a "new" response; indicating that the face was not studied previously). Generally, previously studied faces will produce higher "values" of stimulus intensity than lures (i.e., non-studied faces). Thus, a decision is based on the evidence derived from the stimulus (and the observer's sensitivity to this evidence), and the response criterion (which can be sensitive to external and internal factors that might make a responder more or less cautious to provide a particular type of response). Confidence is then thought to index the extent to which the value for the stimulus exceeds the criterion (in the case of old responses),or falls short of the criterion (in the case of new responses).

Factors that relate to the physical properties of the stimuli and/or the strength of encoding, such as exposure time or the perceived similarity between a foil and target, can influence sensitivity (Green & Swets, 1966). Bias, on the other hand, is generally influenced by the payoff for making a correct decision vs the consequences of making an incorrect decision, or pre-existing ideas about the likelihood that a test item will have been presented at study. Circling back to the face recognition example, if the experimenter were to decrease the retention interval for each image, it would be expected that the participant would show decreased sensitivity (due to a decrease in memory quality). Further, if participants were being rewarded monetarily for correct identifications without being penalised for incorrect identifications, we would expect to see a liberal shift in criteria associated with bias toward picking to maximise the number of correct identifications made (Lynn & Barett, 2014). In contrast, as mentioned previously, an eyewitness viewing a real lineup might be concerned about falsely identifying an innocent suspect – understanding that false identifications are a major contributor to wrongful conviction – and may therefore set a conservative response criterion to avoid this costly error.

SDT suggests that a decision is made when the signal strength hits a predetermined criterion. The *criterion hypothesis* then holds that confidence indexes the difference between strength of the memory signal and the decision criteria, with stronger signals further exceeding the criterion and resulting in higher confidence ratings (Petrusic & Baranski, 2003). In sum, SDT suggests that factors that lead to increased accuracy (memory strength, physical properties of the stimuli that enhance discriminability, etc.) will also lead to increased confidence. Thus, confidence should be useful for discriminating target from foil stimuli. Moreover, given that changes in response criterion – that may often be independent of stimulus discriminability – can fundamentally change a decision and contribute to error (i.e., the same stimulus intensity can lead to a hit or a miss, depending on criterion

placement), SDT would suggest that using ratings might preserve diagnostic information, and attenuate recognition error. Although SDT provides an intuitive theoretical account of the processes that shape confidence ratings, it has been suggested that it fails to adequately explain some patterns of confidence for recognition memory tasks. For example, across two experiments – the first of which altered stimulus probabilities, and the second payoffs for correct answers – Van Zandt (2000) tested the assumption if confidence is scaled from stimulus strength/familiarity, that zRoc curves should be invariant under different levels of response bias. While bias may shift the confidence range higher/lower, the points should still fall upon the same, straight, zRoc function. This is not what Van Zandt found, suggesting that the criterion hypothesis – while generally quite successful – may oversimplify the basis of confidence judgments. Thus, despite the popularity of SDT in the confidence literature and the long-standing utility of this framework more broadly, it is important to consider other theoretical accounts if we hope to paint the most comprehensive picture.

Accumulator Models

Vickers' (1970) accumulator model of two choice discrimination was developed to explain the process by which individuals come to a decision in scenarios where there are two potential outcomes. For example, if they are asked to determine which of two lines is longer, or if a face has or has not been studied before. It posits that there are two simultaneous accumulation processes: one for evidence favouring option A (A>B) and another for evidence favouring option B (B>A). A decision is made in favour of A>B or B>A when the evidence in the relevant accumulator reaches the predetermined threshold. According to this approach, confidence indexes the "balance of evidence" between the two accumulators when a decision is reached (Smith & Vickers, 1998: Vickers & Lee, 1998; Van Zandt, 2000). This means that the larger the discrepancy is between A>B and B>A, the more confident the person will be in their decision. Essentially, confidence indexes stimulus discriminability.

In sum, Vicker's accumulator model suggests that decision making (and therefore accuracy) reflects the amount of evidence in favour of the chosen response, and confidence reflects the strength of evidence in favour of the chosen option *compared to* the strength of evidence in favour of the alternative option. Given this, an accumulator account would assume recognition-ratings to also be reflective of the difference in evidence accumulated for option A vs option B. As discussed before in regard to SDT, we would still expect biases that influence criterion (or in this case, threshold) to influence recognition-ratings. This shift in responses, however, should have less impact in applied settings when represented on a more a fine-grained confidence scale as opposed to a binary decision judgment (where a small shift in criterion could lead to a person responding "yes" instead of "no").

The Doubt-Scaling Model

SDT and Accumulator models tend to assume – implicitly or explicitly – that decisions and confidence are based on an index of diagnostic information (e.g., signal strength or stimulus discriminability). However, a relatively overlooked theoretical account – Baranski and Petrusic's (1998) doubt-scaling model of confidence – suggests that non-diagnostic information (i.e., information present at the time of decision that *does not* provide evidence in favour of any given response option) plays an important role in shaping confidence ratings and decision making. The doubt-scaling model proposes that there are in fact three accumulation processes occurring simultaneously during the decision-making process: A>B, B>A, and A=B. In this scenario, A=B represents information present at the time of decision that does not support A>B or B>A (i.e., non-diagnostic information). Unlike the traditional balance of evidence hypothesis, the doubt-scaling model posits that confidence is not reflective of the difference in accumulation between A>B and B>A, but rather that it is inversely proportionate to the amount of information accumulated for A=B (i.e., the more non-diagnostic information accumulated, the less confident the responder will be in their

response). Further, the model also suggests that if the accumulator for A=B reaches threshold before the evidence in favour of A>B or B>A, then a "guess" response will be triggered. This means as the amount of non-diagnostic information present at the time of decision increases, that both confidence ratings and decision accuracy (due to an increased reliance on guessing) will decrease accordingly.

While SDT and Vicker's (1970) accumulator model suggest that confidence is reflective of the evidence in favour of the chosen response option, the doubt-scaling model suggests that non-diagnostic information drives confidence processing. Where low confidence responses would typically be indicative of an increased likelihood of a "guess" response (and therefore lower accuracy), a high confidence response is not necessarily reflective of the amount of evidence in favour of the chosen response compared to the decision threshold and/or the alternative response. Instead, we would expect a high confidence response to be reflective of an absence of non-diagnostic information at the time of decision. Thus, under this theoretical framework, there may be scenarios in which a shift in decision criteria that would typically lead to a change in binary responding would not result in a shift in recognition-ratings (see Chapter 3 for a more thorough examination of this concept including examples).

Though the aforementioned psychophysical models differ regarding the specific mechanisms by which they propose confidence is determined, they do share one common assumption: that confidence reflects the evidence base for the initial decision and stimulus discriminability. In the current work, our initial goal was not to contrast these individual psychophysical theories, but rather to contrast the common predictions shared by these models with predictions drawn from theories that suggest a more inferential basis of confidence.

Inferential Models of Confidence

Unlike psychophysical models of confidence, which were conceptualised to account for decision making in relation to perceptual tasks, inferential models were born out of an attempt to explain how heuristics and inferential cues shape metacognitive judgements and account for dissociations between metacognitive judgements and task performance. Specifically, inferential accounts have been built upon research on Feelings of Knowing (FOKs; the feeling that information in memory might be cued by a later recognition test despite being presently unretrievable; Koriat, 1993) and Judgements of Learning (JOLs; referring to judgments about the likelihood that studied information will be retrievable in a later test; Koriat, 1997). Given the suggestion by Van Zandt (2000) that the criterion-hypothesis of confidence fails to capture some of the non-stimulus factors that influence confidence ratings, it could be that an inferential approach to confidence provides a more holistic account.

While FOKs and JOLs do have some conceptual similarities to confidence ratings – they all index metacognitive uncertainty – there are also some key differences. Primarily, both JOLs and FOKs require predictions related to future performance, whereas confidence ratings reflect current task performance (e.g., ratings of the likelihood that a test stimulus was presented at study, or the likelihood that a made decision is correct). Although it may seem intuitive to apply models that account for other metacognitive judgements to confidence ratings (given that all metacognitive judgements involve some assessment of cognitive processes), it is important to test whether the same mechanisms apply. It could be that inferential mechanisms are less influential when the test stimulus is present (allowing an assessment of stimulus discriminability), than when future performance must be inferred based on current memory strength. To understand the potential generalizability of these inferential metacognitive frameworks to confidence in recognition or perceptual

discrimination, we must first consider the contexts in which these inferential models emerged.

Koriat (1997) aimed to determine how both intrinsic and extrinsic factors would influence JOLs for a word association task. In their first experiment (which we have chosen to focus on as a clear example of a typical JOL task), participants were presented with 50 word pairs, ranging from easier associations (e.g., cow-milk) to harder associations (e.g., citizen-fox). After each pair was presented, participants were asked to provide a JOL (from 0-100%) reflecting the likelihood that they would be able to recall the second word of the pair (i.e., the target) if they were to be presented with the first word (i.e., the cue). Participants were then tested on their recall. This process was repeated, with participants in the experimental condition receiving the same set of stimuli for the second study phase. The control condition received a new set of word pairs. Koriat found that studying the word pair twice (as compared to the control group who studied the list once) had larger effects on recall accuracy than it did JOL. Further, the difficulty manipulation (hard vs easy word pairs) had a larger effect on JOLs than it did recall accuracy. This dissociation between factors that influence memory strength and factors that influence JOLs suggest that JOLs are to some extent influenced by non-memorial, inferential cues.

These inferential cues were summarized in Koriat's cue-utilization approach (1997) to metacognitive judgments. The model proposes that there are three different types of "cues" that inform confidence ratings: intrinsic, extrinsic, and mnemonic. Intrinsic cues relate directly to the item studied and impact how easy/difficult the participant finds learning the item (e.g., the semantic relatedness of the two words when recalling word pairs, or the concreteness of a word). Extrinsic cues are not related to the item studied, but to the conditions in which the item was learned, and the ways in which the participant engaged with the item in order to encode it (e.g., stimulus repetition or exposure duration). Mnemonic cues

refer to internal cues that suggest that an item has been learned well. Although there will often be some overlap between these cues and stimulus discriminability, many of these cues rely on inferences driven by participants' heuristics relating to the types of stimuli, as well as encoding and test conditions, that are likely to be associated with enhanced memory performance

In another paper, this time focussing on FOKs, Koriat (1993) tested participants using four letter consonant-strings known as "tetragrams" (e.g., RDFK). Each trial consisted of two parts: recall and recognition. For the recall part of the trial, participants completed a short four-item stroop task before being presented with a tetragram for 1000ms, followed by another short stroop task. Participants were then asked to recall the tetragram, receiving one "point" for every correct letter (with no penalties for incorrect letters). Finally, they provided a FOK judgement in relation to the tetragram that they had just recalled. In the second part of the experiment, participants were tested on how well they would recognize the correct tetragram when presented alongside seven other tetragrams of varying levels of similarity. For trials in which the participant did not recall the entire tetragram, Koriat found that future recognition was best predicted by the *accuracy* of the partial information. Surprisingly, however, the magnitude of participants' FOKs increased in line with the *total amount* of information recalled, regardless of whether the letters were accurate. That is, it was the *amount* of partial information retrieved, not the accuracy/diagnostic value of the partial information retrieved, that drove FOKs. Essentially, additional non-diagnostic information inflated metacognitive judgements.

From this work, Koriat's accessibility model of FOKs emerged (1993). The theory posits that metacognitive judgements infer reliability (and therefore performance) from the accessibility of information, regardless of the accuracy/diagnostic value of said information. This theory is seemingly at odds with SDT and accumulator-based models of confidence, and

is directly contradictory to the doubt-scaling model of confidence (which suggests that additional, non-diagnostic information should *reduce* confidence rather than inflate it).

Non-Diagnostic Information and Confidence

Clearly, one area of difference between psychophysical models and inferential models of metacognition is the presumed influence on confidence of additional, non-diagnostic information at test. In general, psychophysical models would suggest that providing additional non-diagnostic information in the test phase of a recognition memory task should have no effect on confidence, as it does not affect the evidence accumulated in favour of competing response options (assuming that the presentation of the non-diagnostic information does not interfere with the accumulation of diagnostic information). Baranski and Petrusic's doubt-scaling model (1998), however, suggests that confidence is inversely proportionate to the amount of non-diagnostic information present at the time of decision making. Thus, according to this model, additional non-diagnostic information at test would be expected to decrease confidence in recognition. Finally, inferential accounts of confidence suggests that confidence may reflect the total amount of information recalled, and therefore that additional, non-diagnostic information should lead to higher confidence in recognition.

Some initial empirical evidence comes from Busey et al. (2000), who conducted a face recognition task in which they altered the luminance levels of facial images at study and test, asking participants to make a retrospective confidence judgment alongside a yes/no identification at the test phase. They found that while accuracy was highest when test luminance matched study luminance, confidence was highest when test luminance was at its brightest, regardless of encoding luminance. This suggest two things. First, that confidence may not be driven by the same physical evidence that informs accurate decision making, and second, that providing additional, non-diagnostic information may actually inflate confidence for recognition memory tasks. While this finding does not support general, psychophysical

accounts of confidence, and directly contradicts the doubt-scaling model, it may be explained by an inferential account. Specifically, that the total amount of information present, regardless of its diagnostic value, may determine confidence (with more information leading to greater confidence).

Measuring Recognition

Although the primary focus of this thesis is to investigate how different theoretical models account for confidence - and particularly for the contribution of additional, non-diagnostic information to confidence - a second area of interest lies in how to best measure and communicate confidence. While this refers in part to our work directly comparing the results of participants using a rating-only confidence scale to those who were asked to provide a decision judgement followed by a confidence rating (see Chapter 5), it also refers to our interest in whether fine-grained numerical scales (e.g., 0-100%) and more coarse-grained verbal scales (e.g., "Not Confident", "Moderately Confident" etc.) are differentially vulnerable to noise in confidence ratings that might be produced by additional, non-diagnostic information (Chapter 3).

Confidence ratings are often measured using probabilistic scales ranging from 0-100% (e.g., Brewer & Wells, 2006; Koriat, 2011; Sauer, Weber & Brewer, 2012), or more concise verbal scales (e.g., with labels such a "Sure Old", "Sure New" etc.; Ratcliff, McKoon, & Tindall, 1994). Although probabilistic scales lend themselves easily to measures of metacognitive accuracy typically of interest to researchers (e.g., calibration and over/underconfidence), some have argued that probabilistic ratings are not representative of how people typically conceptualize their own uncertainty (Windschitl & Wells, 1996). If we are to rely on measures of confidence to provide an assessment of the likely accuracy of a decision, or to replace the decision judgement altogether, then it is important to ensure that the method used to record confidence ratings is truly reflective of the intended meaning.

Previous research focusing on how to best measure metacognitive uncertainty (of which confidence is an example) has provided some support for the argument against numerical/probabilistic representations of confidence, with verbal scales having being more responsive to changes in context and framing, better at predicting choices or preferences when uncertainty was involved, and better at predicting behavior intentions than numerical scales (Windschitl & Wells, 1996). Weber and Brewer (2008), however, found only negligible differences between how participants used an 11-point numerical scale (with points ranging from 0%-100%) to an 11-point verbal scale (ranging from "Impossible" to "Certain").

There is, however, the potential for non-trivial variance in regards to how people interpret verbal measures of uncertainty. For example, interpretations of verbal weather forecasts differ between experts (Handmer & Proudley, 2007). It has also been found that although participants report that they prefer to communicate their own uncertainty to other people verbally, they would rather receive information regarding uncertainty in numerical form (Erev & Cohen, 1990; Wallsten, Budescu, Zwick & Kemp, 1993). The preference for receiving numerical expressions of uncertainty is supported by Mansour's (2020) work comparing the use of a numerical confidence scale (0-100%) to free-response verbal expressions of confidence in relation to a recognition memory task. Mansour found both measures of confidence to be reflective of accuracy, but reported that the verbal expressions were often difficult to interpret. This demonstrated preference for receiving numerical ratings may suggest that the assumption that people do not conceptualise uncertainty in numerical terms is flawed.

Another question that arises when considering how to best record confidence data relates to the optimum number of response options. Using a 0-100% scale as an example, participants could be asked to provide an estimate that lies anywhere within that range, or

they could be asked to respond on an 11-point Likert-type scale, with points labelled 0%, 10%, 20% etc. While the more fine-grained option may allow participants more flexibility with their responses, Benjamin, Tullis and Lee (2013) argue that increasing the number of response options in a scale can result in an increased measurement noise. Reducing the amount measurement noise is not only advantageous from an analyses standpoint, but also from an internal validity standpoint. In cases like this, where it could be argued that the noise arises from problems in mapping one's internal state onto such a specific external scale; the final results may not faithfully reflect the participant's internal uncertainty (Benjamin et al., 2013; Hanczakowski et al., 2013). Tekin and Roediger (2017), however, suggested that the number of points on a measurement scale makes no difference in terms of their predictive accuracy. When they compared 4-point, 5-point, 20-point and 100-point numerical scales across two recognition memory tasks (one using a word list, one using face images), they found no discernible differences in response patterns. Given these contradicting results, we are interested in determining whether patterns of results will differ when collected on a fine-grained or coarse-grained measurement scale, and whether scales that allow for more criterion-related noise show increased effects of non-diagnostic information.

Although comparing a fine-grained numerical measure to a coarse-grained verbal measure raises an obvious confound, the decision to initially use these two measure types stems from data collected for confidence accuracy characteristic analysis (CAC; Mickes, 2015). CAC allows researchers to plot confidence against accuracy in the form of a CAC curve. While the confidence data used for CAC analyses is typically collected as a percentage (i.e., 0-100%), it is then typically collapsed into three categories, with the first category containing ratings from 0-60%, the next ratings from 70-80%, and the final containing ratings from 90-100 (e.g., Mickes 2015; Carlson et al, 2016; Sauerland et al., 2016). This tendency to collapse numerical representations of uncertainty into three categories piqued our interest,

especially given the literature suggesting that fine-grained measures are more susceptible to noise, and verbal measures may be more reflective of internal conceptualisations of uncertainty. Thus, we decided to initially compare the two measure to determine whether a more simplistic scale may provide informative recognition-ratings that do not require further manipulation before analyses. If the scales produced different patterns of results, we planned to tease apart the effects of numerical vs verbal labels, and more vs fewer response options, in subsequent experiments. To foreshadow upcoming chapters, this did not prove necessary.

Summary

In sum, across three manuscripts and seven studies, this work tests how confidence is affected by non-diagnostic information across a variety of task types (comparing predictions from psychophysical and inferential models of metacognition), provides direct tests of a largely over-looked psychophysical model, and begins an examination of how ratings-based approaches might be applied to counter bias-related errors in applied visual search tasks. We also provide an initial investigation of potential scale-related effects on the efficacy of ratings-based approaches to recognition.

Chapter III

The Theoretical Basis of Recognition-Ratings: Inferential vs Psychophysical Approaches to Confidence

The Theoretical Basis of Recognition-Ratings: An Inferential Approach vs A Psychophysical Approach to Confidence.

Amelia T. Kohl[1], James D. Sauer[1], Matthew A. Palmer[1]

[1] The School of Psychological Sciences

The University of Tasmania, Australia

Manuscript in preparation for submission to the Journal of Experimental Psychology: General.

This research was supported by funding from the Australian Research Council (grant DP200100655 to A. Heathcote, J. Sauer, M. Palmer et al.).

Corresponding Author:

Amelia Kohl, School of Psychological Sciences, University of Tasmania, Locked Bag 30, Hobart, Tasmania 7001, Australia

Email: Amelia.Kohl@Utas.edu.au

Abstract

In the recognition memory literature, confidence ratings have been used in the absence of a binary yes/no or old/new judgement to effectively discriminate whether an image has been previously studied. The theoretical mechanisms that underpin such ratings, however, are currently unexplored. While it may seem intuitive that the same models used to account for confidence in the psychophysics literature (e.g., signal detection theory, accumulator models etc.) would also apply to recognition-ratings, it could be that models developed to explain metacognitive judgements such as judgements of learning (JOLs) and feelings of knowing (FOKs) provide a better account. To test this, we investigated how the presence of non-diagnostic information during the decision making process would influence confidence ratings. While inferential theories (i.e., those used to account for JOLs and FOKs) posit that confidence should increase when there is more information present at the time a decision is made, regardless of whether the information is diagnostic, psychophysical models suggest that confidence should remain constant regardless of the presence of non-diagnostic information. One psychophysical model that provides an alternative prediction is Baranski and Petrusic's doubt-scaling model of confidence (1998), which suggests that confidence should decrease as the amount of non-diagnostic information present increases. Here, we report three experiments testing the theoretical mechanisms that shape recognition-ratings. Experiment 1 used a face recognition paradigm, Experiment 2 a landscape/house recognition paradigm (to test whether effects from Experiment 1 generalised beyond faces to other complex stimuli that might be less reliant on holistic processing), and Experiment 3 used a perceptual discrimination task asking participants to indicate their confidence that a dynamic grid was primarily blue or primarily orange. In both Experiments 1 and 2, non-diagnostic information was manipulated by presenting both full and partial versions of the same image (i.e., if a participant encoded a partial image but was tested with the full version of the same

image, this would constitute the presence of non-diagnostic information at test). In Experiment 3, non-diagnostic information was operationalized through the inclusion of white pixels within the grid, with both the proportion of white pixels and the total number of diagnostic (i.e., blue/orange) pixels bring manipulated. Across all three experiments, we found that confidence decreased as the amount of non-diagnostic information increased, providing support for the doubt-scaling model of confidence.

Keywords: confidence, doubt-scaling, decision making, metacognition

Understanding the processes that underlie decision making has been a constant point of interest in the psychological literature, with early work in the area dating back over a century (e.g., Angell, 1907; Baranski & Petrusic, 1998; Henmon, 1911; Horry & Brewer, 2016). Understanding the basic process by which individuals accumulate and interpret evidence is particularly useful for conditions requiring individuals to make important decisions under conditions of uncertainty (e.g., an eyewitness choosing a perpetrator from a lineup, or a security screener examining luggage x-rays for contraband), as it may help in (a) understanding the causes of error and (b) identifying potential indices of decision accuracy.

For example, considering the basic architecture that underlies decision-making also points to confidence as a potential index of accuracy. It has been shown that across a variety of domains - ranging from basic perceptual discrimination tasks through to complex, real-world decision-making tasks - confidence is generally positively associated with decision accuracy (e.g., Brewer & Wells, 2006; Hertzog et al., 1990; Gigerenzer, et al., 1991; Juslin et al., 2000; Palmer et al., 2013; Sauer et al., 2010; Sauer & Brewer, 2015; Wixted & Wells, 2017). There are two main schools of theory that are used to explain the underlying processes that shape confidence and associated metacognitive judgements (i.e., judgments that evaluate the reliability of cognitive processes). First, there are psychophysical models, which originated to account for decision-making and confidence in basic perceptual judgements and were later extended to recognition memory tasks (e.g., those derived from Signal Detection Theory; Bernbach, 1971; Egan, Schulman, & Greenberg, 1959; Green & Swets, 1966; Wickelgren & Norman, 1966; or accumulator models coupled with a balance of evidence hypothesis; Vickers, 1979; Van Zandt, 2000). Second, there are what we will refer to as inferential models, which are commonly applied in the broader metacognition literature and often concerned with how people assess their own learning and the likelihood of future remembering (e.g., the cue utilization model; Koriat, 1997). While both schools of theory

propose that memory strength and/or the degree of match between what was seen at encoding vs. retrieval plays an integral part in forming confidence ratings, the key difference between these types of accounts lies in the emphasis placed on the role of other, non-memorial factors. The primary focus of this paper is to determine the extent to which these different approaches can account for confidence ratings for recognition memory tasks.

One area of application for which confidence ratings are particularly useful is the eyewitness identification domain. Mistaken identifications – where a witness identifies an innocent suspect as the culprit – are both common and costly; being a frequent contributing factor to documented cases of wrongful conviction (Brewer & Williams, 2017; The National Academy of Sciences Report, 2014; Technical Working Group for Eyewitness Evidence; 1999). Thus, researchers have focused considerable attention on attempting to identify reliable and informative markers of identification accuracy. Confidence is one of the most researched of these potential markers. Consistent with findings in basic decision making tasks, a large and robust body of literature demonstrates a positive relationship between confidence and identification accuracy when a witness identifies someone from a lineup (provided confidence is collected under appropriate conditions, see Wixted & Wells, 2017 for a review). Further, mock juror studies have consistently demonstrated that eyewitness confidence influences perceptions of defendant guilt (e.g., Bradfield & Wells, 2000; Brewer & Burke, 2002; Cutler et al., 1990; Sauer et al., 2017), and surveys have found that police, lawyers, and jurors all believe that confidence is predictive of accuracy (Deffenbacher & Loftus, 1982; Potter & Brewer, 1999). Further, the U.S. Supreme Courts *Neil v. Biggers* (1972) ruling identified as confidence a key criteria against which identification accuracy should be assessed and, more recently, The National Academy of Sciences Report (2014) and the Technical Working Group for Eyewitness Evidence (1999) have also recommended that

eyewitness confidence be presented in court to help jurors assess the likely accuracy of an identification.

Despite the theoretical and empirical support for a confidence-accuracy relationship, and the findings that confidence is a compelling marker of accuracy in important applied settings, it is also the case that metacognitive judgements (of which confidence is one example) can, in general, be vulnerable to non-diagnostic influences (i.e., factors unrelated to stimulus discriminability). Thus, we designed a series of experiments to explicitly test the effect of additional non-diagnostic information on confidence ratings given in the absence of a binary recognition judgement (i.e., recognition-ratings). Although researchers interested in confidence have typically focused on the utility of confidence in diagnosing the accuracy of a decision, it has been demonstrated more recently that confidence ratings can be useful in discriminating target from foil stimuli in the absence of a categorical decision (e.g., Sauer et al., 2008, 2012; Brewer et al., 2012, 2020), and argued that such an approach may help attenuate errors associated with decision biases in applied recognition tasks (Sauer et al., 2012).

In Experiments 1 and 2, we investigated whether providing additional non-diagnostic information at the test phase of a recognition experiment affected participants' recognition-ratings (i.e., confidence that a test stimulus had presented during the encoding phase) when making judgements pertaining to whether they had been exposed to a face (Experiment 1), house or landscape (Experiment 2) in an earlier study phase. Further, these experiments tested whether the effects of our key manipulation (i.e., the amount of non-diagnostic information at test) on recognition-ratings differed for participants who responded using a more fine-grained probabilistic scale compared to those who used a more coarse-grained verbal scale.

In Experiment 3, participants completed a perceptual discrimination task in which both the proportion of non-diagnostic information and the total amount of diagnostic information were manipulated to extend the findings of Experiment 1 and 2 in a more controlled paradigm. Specifically, participants were presented with a dynamic grid consisting of blue, orange and (sometimes) white pixels. They were then asked to indicate their confidence as to whether the grid was primarily blue/orange, with the white pixels serving as non-diagnostic information. Together, these experiments allowed us to determine whether recognition-ratings are best explained by psychophysical models – which argue that confidence is primarily a property of stimulus discriminability – or by inferential models – which argue that metacognitive judgements can be influenced by the amount of information available when a judgment is made, regardless of the diagnosticity of the information.

Recognition-Ratings

Ideally, a recognition decision would be based on the extent to which a test stimulus matches the individual's memory of a previously viewed stimulus. However, in both basic and applied tasks, factors that operate largely or solely on response bias can affect decisions and contribute to error. For example, if the payoff for a correct decision is larger than the consequence for an incorrect decision, it can lead to a more liberal response criterion and increase false alarms (e.g., Lynn & Barett, 2014). Thus, a method of indexing recognition that more directly reflects the degree of match between the stimulus encoded in memory and the test stimuli, and is potentially less vulnerable to effects of response bias, could reduce costly errors (e.g., Brewer et al., 2012, 2020; Sauer et al., 2008).

Confidence in recognition memory is generally measured in one of three ways: retrospectively (after a decision has been made), simultaneously (at the same time as a decision is made), or even prospectively (before a decision has been made). An alternative approach is to ask participants to provide a confidence rating in the absence of a categorical

decision judgement (e.g., asking participants to forego a categorical decision and simply "Provide your confidence from 0-100% that you have seen this face before"). Sauer et al. (2008) suggested that such a method – which we will refer to as "recognition-ratings" – may provide a more informative measure of memory strength by reducing the impact of factors that influence response bias or criterion placement to provide a more direct measure of evidence accumulation.

Research investigating the utility of recognition-ratings for face recognition tasks has found that participants were better able to discriminate a target among foils compared to a control, and that recognition-ratings provide a richer source of information than retrospective confidence ratings (e.g., Sauer et al., 2008, 2012; Brewer et al., 2012, 2020). Further, factors such as retention interval and distinctiveness had less impact on participants' ability to discriminate between previously studied and unstudied faces when using recognition-ratings rather than a binary decision (Sauer et al., 2012). While early work looking at recognition ratings classified confidence ratings as "old" or "new" by establishing a criterion for each individual participant that maximized the amount of correct old and new responses (e.g., Sauer et al., 2008; Koriat & Goldsmith, 1966), more recent work applying a different classification approach has provided promising results.

Brewer et al. (2020) conducted an eyewitness identification task, in which participants viewed a mock-crime video followed by a 12 person lineup. They were then asked to provide a recognition-rating from 0-100% for every lineup member regarding their confidence as to whether they were the perpetrator. Brewer et al. then examined the "max" ratings; that is, the highest confidence rating that the witness gave to any one lineup member (therefore indicating that that suspect was the closest match to their memory). They found max confidence ratings to be predictive of guilt, with suspects who received the max confidence rating being more likely to be the guilty party, and suspects who did not receive the max

being more likely to be innocent. These findings were consistent across both child and adult samples. Of particular interest are trials in which participants provided max confidence ratings for multiple lineup members. Where traditional (i.e., including a binary decision) lineup tasks may lead to the participant rejecting the lineup if two lineup members appear to be equally plausible, resulting in a "miss", recognition-ratings allow for participants to still provide a measure of recognition. If one of the lineup members who receives the max confidence rating is in fact the suspect, then the recognition-rating can provide diagnostic evidence independent of the fact that another lineup member also seemed to resemble the perpetrator, something that is unlikely in a traditional lineup task.

When motivating the above work, researchers have been clear that such outcomes are consistent with predictions drawn from a variety of psychophysical models of confidence in recognition (i.e., models that tie confidence to stimulus discriminability; e.g., SDT; Green & Swets, 1966; accumulator models; Van Zandt, 2000; Vickers, 1970). The predictions are also consistent with models of metacognitive judgements in other domains, such as Koriat's (1997) cue utilization model of judgements of learning and (1993) accessibility model of feeling of knowing (i.e., models that hold that reliable, diagnostic information can be accessed in the absence of an explicit decision to inform metacognitive inferences; see Sauer et al., 2008). However, previous work has not attempted to directly compare predictions drawn from these competing models to see which school of theory best accounts for recognition ratings.

A Thurstonian vs a Brunswikian Approach

An examination of Thurstonian and Brunswikian accounts of uncertainty (Juslin & Olsson, 1997) provides a framework for considering how a recognition memory task fits within a psychophysical framework compared to a more typical, general metacognitive framework. Similar to psychophysical models of confidence, the Thurstonian approach to

uncertainty characterises uncertainty that is caused by noise in a person's information processing system rather than any issues relating to the stimuli itself. The Brunswikian approach, on the other hand, characterises uncertainty that is based upon the imperfect nature of the relationship between what is currently known, and what is currently unknown and/or what will occur in the future. Unlike the Thurstonian account, the Brunswikian approach to uncertainty is heavily reliant on cues, much like the inferential accounts of metacognition where future performance (e.g., likelihood of retrieval) must be inferred based on current memory states. Juslin and Olsson (1997) suggest that the Thurstonian approach is used primarily in sensory discrimination tasks where the stimuli to be compared are presented at the same time; unlike the Brunswikian approach which dominates cognitive tasks, such as JOLs and FOKs, where individuals are using their current knowledge to make predictive judgements about the future. For example, when comparing two lines to determine which is longer, all the necessary information is present in front of the observer. Thus, consistent with a Thurstonian approach, any uncertainty relates to limitations on the processing of that information. In contrast, when judging whether studied information will be retrievable in the future, an individual must make inferences based on the current retrievability of information. However, it is unclear how well the current retrievability will predict the future retrievability of information. Thus, consistent with a Brunswikian approach, there is uncertainty/noise inherent in the information upon which the decision is based.

It is difficult to class a recognition task – in which participants make decisions and confidence judgements by comparing a stimulus (or stimuli) in front of them to their memory of a previously viewed stimulus – as a task that is characterized solely by either Brunswikian or Thurstonian uncertainty. In the literature, researchers have tended to apply the Thurstonian approach as models of confidence in recognition tend to draw on theoretical frameworks designed for perceptual discrimination tasks (e.g., SDT and Accumulator models). However,

given that memory represents an imperfect source of information and this adds noise to the comparison process (cf. having two stimuli available for direct comparison), Brunswikian accounts (which emphasise the role of inference in assessments of uncertainty based on incomplete information) may also have something to offer. Relying solely on psychophysical models, ignoring the inferential mechanisms common among models designed to account for tasks involving Brunswikian uncertainty, might fail to appreciate some of the complexity of metacognitive judgements for recognition memory tasks. Having a solid theoretical understanding of how such judgements are formed will provide a better understanding of conditions under which recognition-ratings may become dissociated from memory strength (i.e., when non-diagnostic cues are likely to affect confidence).

Psychophysical Models of Retrospective Confidence

Originating in the psychophysics literature to account for how participants make decisions regarding perceptual stimuli, psychophysical models of confidence posit that confidence indexes the physical properties of the stimuli (e.g., signal strength or stimulus discriminability). Although individual models vary, psychophysical theories are typically grounded in, and rely on mechanisms proposed by, signal detection theory where confidence indexes signal strength relative to the response criterion (Bernbach, 1971; Egan, Schulman, & Greenberg, 1959; Green & Swets, 1966; Wickelgren & Norman, 1966;) or accumulator models (Van Zandt, 2000; Vickers, 1970) coupled with a balance of evidence mechanism, where confidence indexes the difference in evidence accumulated for competing response options. Despite their origins in psychophysics, psychophysical models of confidence have been adapted in the recognition memory literature to account for confidence in recognition judgments.

Signal Detection Based Models

Signal detection theory (SDT; e.g., Bernbach, 1971; Egan, Schulman, & Greenberg, 1959; Green & Swets, 1966) suggests that a decision is made when signal strength (e.g., the strength of recognition) reaches a predetermined criterion. In a recognition memory task, a participant compares a presented test stimulus against the memory for previously viewed stimuli. This comparison generates a signal strength (e.g., feeling of familiarity). If this signal strength hits/passes criterion (e.g., when the physical properties of the stimuli provide enough of a match with the version encoded to memory), the response is either classed a "hit" (if the decision is correct) or a "false alarm" (if the decision was incorrect). If the signal strength fails to reach criterion, it is classed as a "miss" (if the decision was incorrect) or a correct rejection (if the decision was correct; Macmillan & Creelman, 1991). Confidence is then thought to index the difference between the strength of the memory signal and the decision criteria, with stronger signals further exceeding the criterion and resulting in higher confidence ratings (Petrusic & Baranski, 2003). Importantly, criterion placement can be vulnerable to the effects of bias, such as the payoff for making a correct identification vs the consequences of making an incorrect decision, or pre-existing ideas about the likelihood that a test item will have been presented at study (Lynn & Barett, 2014).

To put this in concrete terms, consider a recognition task where participants study a series of face images. At test, they are then presented with a series of faces one by one, and are asked to indicate for each (a) whether they studied the face previously, and (b) their confidence in their decision. When presented with a face that was studied (i.e., a "target"), signal strength will generally fall toward the higher end of the strength continuum. When presented with a face similar to one they studied (i.e., a "lure"), there will be some degree of signal strength, however it will not be as strong as for the target. When presented with a face that they have not seen before that does not resemble any of the previously studied faces (i.e., a "foil), signal strength will be negligible. This account helps explain how confidence, and

likely accuracy, both tie to the strength of the underlying memory signal and, therefore, how confidence can be a useful index of recognition, even in the absence of a categorical decision: by indexing signal strength. While a shift in response criteria would influence recognition-ratings in a similar way to that of a categorical response, it would likely have less of a noticeable effect. For example, when presented with a face and asked whether they have seen the image before, a conservative criterion shift may push a participant from responding "yes" to responding "no". If asked instead to provide their confidence as to whether they had seen the face before, the conservative criterion shift may lead to a response of 60% confidence becoming a response of 50% confidence. While the result is still influenced by criterion, this change is less exagerated than that provided in the binary choice example.

Accumulator Based Models

Vickers' (1970) accumulator model of two choice discrimination suggests that there are two simultaneous accumulation processes: one for evidence favouring option A and another for evidence favouring option B. When evidence in favour of Option A or Option B reaches a predetermined decision threshold, a choice is made in favour of that option. Within this framework, confidence is thought to reflect the "balance of evidence" between the two response options when the decision is made, with high confidence responses being associated with large discrepancies between the evidence in favour of option A and option B (and vice-versa; Smith & Vickers, 1998: Vickers & Lee, 1998; Van Zandt, 2000). For example, in a face recognition task, a participant is presented with a face and asked to (a) decide whether they have seen the face before, and (b) indicate their confidence in the accuracy of their decision. The features of the face that are familiar to the participant would lead to evidence accumulating in favour of option A (i.e., that they have seen the face before), whereas any feature that is different to their memory of the face would result in evidence accumulating for option B (i.e., that they have not seen the face before). Once enough information has

accumulated to hit the decision threshold for either option, the participant responds and their confidence in the decision will be informed by how much more evidence is in the winning accumulator compared to the other accumulator.

Much like SDT, accumulator models suggest that both confidence and accuracy reflect the same underlying evidence base. Specifically, that confidence indexes the extent to which one decision option "won out" over the other. In this scenario, we would expect recognition-ratings to be of particular use in scenarios in which the amount of evidence in favour of option A and option B is similar. To use a recognition memory task as an example: a participant is presented with a test image and are asked to identify whether they previously studied the image. Evidence then accumulates in support of option A, that they have seen the image before, and option B, that they have not seen the image before. For this example, evidence in favour of option A eventually wins out. In a traditional binary decision scenario, the participant would simply indicate that they have seen the image before. Recognition-ratings, on the other hand, would reflect the amount by which evidence in favour option A won out over option B; providing a potential insight into the likelihood that the image had been previously studied.

Inferential Models of Metacognition

Originating in the metacognition literature, inferential models suggest that metacognitive judgements (of which confidence is an example) are shaped not only by the physical properties of the stimulus and/or the conditions of encoding, but also by inferential cues and heuristics. Given the earlier suggestion that recognition memory tasks most likely involve both Thurstonian uncertainty – often used in psychophysics tasks – and Brunswikian uncertainty – often used in metacognitive tasks – these inferential models may provide further insight into the processes that shape recognition ratings.

Research focusing upon judgements of learning (JOLs; judgments about the likelihood that studied information will be retrievable in a later test) and feelings of knowing (FOKs; the feeling that information that cannot currently be recalled might be cued by a later recognition test) has provided evidence in favour of an inferential basis for metacognition, suggesting that intrinsic, extrinsic and mnemonic cues can all have an effect (Hertzog et al., 1990; Koriat, 1993; 1997). Koriat (1997) defined intrinsic cues as those related directly to the studied item. For example, participants might estimate the likelihood of future recall based on the perceived difficulty of learning the item. When learning word pairs, pairs of semantically related words (e.g., salt-pepper) may be perceived as easier to learn, and therefore more likely to be remembered, than pairs of non-related words (e.g., stop-basket). Extrinsic cues are not related to the studied item, but to the conditions in which the item was learned, and the ways in which the participant engaged with the item in order to encode it. For example, participants may feel the likelihood of future retrievable is increases with repeated or longer exposure at encoding. Finally, mnemonic cues refer to internal cues that suggest that an item has been learned well. These include cues such as the ease of processing, and ease of retrieval (e.g., retrieval fluency) of the item and associated information when asked to predict future recall or recognition.

Of particular interest to this paper is an experiment by Koriat (1993) examining the mechanisms that underly FOKs. Participants were tested using four letter consonant-strings known as “tetragrams” (e.g., RDFK). Each trial consisted of two parts: recall and recognition. For the recall part of the trial, participants completed a short four-item stroop task before being presented with a tetragram for 1000ms, followed by another short stroop task. Participants were then asked to recall the tetragram, receiving one “point” for every correct letter (with no penalties for incorrect letters). Finally, they provided a FOK judgement in relation to the tetragram that they had just recalled. In the second part of the experiment,

participants were tested on how well they would recognize the correct tetragram when presented alongside seven other tetragrams of varying levels of similarity. For trials in which the participant did not recall the entire tetragram, Koriat found that future recognition was best predicted by the accuracy of the partial information recalled (i.e., how many of the recalled letters were present in the studied tetragram). However, that the magnitude of FOKs increased in line with the total amount of information recalled, regardless of whether that information (i.e., the individual letters) was accurate. Thus, these findings suggest that metacognitive judgements are shaped by the total amount of information retrieved, and non-diagnostic information inflates metacognitive judgements. Based on the conceptual similarities between metacognitive judgements that rely on gauging future memory performance – such as FOKs – and recognition memory tasks (i.e., that both likely involve an element of Brunswikian uncertainty), additional, non-diagnostic information at test may also inflate recognition-ratings.

Non-Diagnostic Information and Confidence in Recognition

Busey et al. (2000) provide some evidence consistent with the applicability of inferential models of metacognition to recognition memory tasks. Busey et al. altered the luminance levels of facial images at study and test, asking participants to make a retrospective confidence judgment alongside a yes/no identification at the test phase. Psychophysical models of confidence would suggest that confidence should be highest when luminance levels are kept the same at study and test, as that is when the degree of match between the test image and the memorial image of the studied image is highest. However, although accuracy was highest when luminescence at test was the same as at encoding, confidence was positively associated with luminance at test, regardless of luminance at study. Thus, confidence increased when more information was available at test, even when that additional information was non-diagnostic and did not increase the degree of match between

the item at study and test. These results align with inferential model of confidence, as the dissociation between confidence and accuracy suggests that the presence of additional, non-diagnostic information at test influences confidence without having a corresponding effect on stimulus discriminability. Thus, participants seem to be making inferences about the quality of their memory – or the strength of their recognition – based on the sheer amount of information available at test; a result that parallels Koriat's (2003) findings relating to the processes that shape FOKs.

Non-diagnostic information is often present in applied settings where people are making judgements based upon their recognition memory. To use an eyewitness identification example, consider a scenario where a person witnesses a crime in which the perpetrator is wearing a mask covering half of their face. When later asked by police to view a lineup and attempt to identify the culprit, they are presented with a series of full, unmasked faces. This additional featural and configurational facial information would be considered non-diagnostic, as the eyewitness has no ability to match this information to their memory (as the information was not present at encoding). This scenario has been tested empirically, with Manley et al. (2019) conducting a series of four experiments in which the amount of facial information was varied between study and test. They found that for trials in which the participants encoded a partial face, confidence ratings were highest when tested with the same partial face (cf., the corresponding full face). These findings are at odds with predictions made by any of the models that we have covered so far in this paper. While inferential models would suggest that confidence should be higher when additional non-diagnostic information is present at test for a recognition memory task (given the suggestion that metacognitive judgements are based on the total amount of information present regardless of diagnosticity), psychophysical models would suggest that the presence of non-diagnostic information at test would not affect confidence in a recognition memory task

(given that the non-diagnostic information did not reduce the amount of available diagnostic information). While psychophysical models of confidence (e.g., SDT or accumulator based models) generally do not have a mechanism to account explicitly for non-diagnostic information, there is one lesser-known theory that may explain why an increase in non-diagnostic information would result in a decrease in confidence: Baranski and Petrusic's doubt-scaling model (1998).

The Doubt-Scaling Model of Confidence

With their doubt-scaling model of confidence (1998), Baranski and Petrusic posit that there is a third accumulator process that occurs alongside the accumulation of evidence in favour of option A and evidence in favour of option B: the accumulation of non-diagnostic information. An evolution of slow and fast guessing theory (Petrusic, 1992), the doubt-scaling model suggests that if evidence in favour of the non-diagnostic accumulator reaches threshold before the other accumulators, then a "guess" response is triggered.

The doubt-scaling model suggests that regardless of the decision outcome (i.e., a decision in favour of option A, a decision in favour of option B, or a guess), confidence is inversely proportionate to the amount of non-diagnostic information that has accumulated during the decision-making process. Thus, an increase in non-diagnostic information at test would lead to a corresponding decrease in confidence.

Unlike the other psychophysical models mentioned in this paper, the doubt-scaling model does not suggest that confidence is primarily based upon the same factors that predict decision accuracy (i.e., stimulus discriminability). While low confidence responses may be reflective of an increased likelihood of a guess response - and therefore lower accuracy - high confidence responses are not primarily reflective of the amount of evidence in favour of the chosen response option. Rather, they indicate a relative lack of non-diagnostic accumulated evidence. Thus, there are scenarios in which evidence in favour of option A and option B

may be comparable, but due to a lack of non-diagnostic information, confidence ratings for the chosen response will be high. If there truly is a doubt-scaling mechanism involved in decision making, it could be that recognition-ratings given when non-diagnostic information is present at test are not reflective of evidence accumulation in the way that other psychophysical models would suggest they are. Although it might be hard to picture a case where this may occur in regards to recognition memory, it is easy enough to operationalize independent manipulations of (a) evidence favouring the competing response option and (b) non-diagnostic information in a perceptual discrimination paradigm. For example, a participant may be presented with a dynamic grid consisting of blue, orange and white pixels. They are then asked to indicate their confidence as to whether the grid is predominantly blue (or predominantly orange), with the white pixels operating as non-diagnostic information. In this paradigm, the amount of diagnostic information (i.e., number of blue vs orange pixels) can be manipulated independently of the amount of non-diagnostic information (i.e., white pixels). A participant may undertake a trial in which the difference between the number of blue and orange pixels is negligible, however there are few non-diagnostic pixels present. According to the doubt-scaling model, in this scenario, their confidence in their decision will be high, despite evidence in favour of the chosen response being similar to that of the alternative option. Vice-versa, in a trial in which there is ample evidence in favour of the chosen response option compared to the alternative, confidence ratings may be low if a large number of non-diagnostic pixels are present.

Despite being the only model to clearly specify the effects of non-diagnostic information on confidence, the doubt-scaling model is directly at odds with the results of Busey et al.'s (2000) findings that non-diagnostic information increased confidence. Thus, we aim to directly test the idea to investigate whether providing additional, non-diagnostic information at test will inflate recognition-ratings (in-line with inferential models of

metacognition), reduce recognition-ratings (in-line with the doubt-scaling model), or leave recognition-ratings unaffected (in-line with mainstream psychophysical models of confidence). In a third experiment, we use a perceptual discrimination paradigm to (a) test the possibility of non-diagnostic information affecting confidence independent of variations in diagnostic evidence, and (b) to rule out an additional explanation - relating to transfer appropriate processing - for patterns observed in experiment 1 and 2.

Measuring Recognition

As a secondary area of interest, we examined throughout these experiments how our approach to measuring recognition-ratings affected their utility as indices of recognition. Although there is a wealth of literature that relies on confidence ratings, there is some debate over how it is best measured. When it comes to indexing confidence in memory, most previous research investigating applied memory (i.e., researchers interested memory and metacognition in applied domains, as opposed to researchers using more basic memory tasks) has recorded confidence (retrospective or in the absence of a decision) as a percentage ranging from 0-100% (e.g., Brewer & Wells, 2006; Sauer et al., 2008; 2012 etc.). However, Windshcitl & Wells (1996) argued (a) against using numerical scales for recording confidence, as numerical scales do not reflect the way people typically conceptualize or discuss uncertainty, and (b) that verbal scales would be better suited to measuring uncertainty because verbal anchors (“certain”, “probable”, “unlikely”, etc.) better reflect the way individuals typically think and talk about confidence. In fact, Windshcitl & Wells found that verbal measures were more responsive to changes in context and framing, better at predicting choices or preferences when uncertainty was involved, and better at predicting behavior intentions than numerical scales. More recent work, however, has suggested that verbal scales and numerical scales may be comparable in terms of usability. One study that compared confidence for a recognition memory task given on an 11-point numeric scale (with points

ranging from 0%-100%) to an 11-point verbal scale (ranging from “Impossible” to “Certain”) found negligible differences between the two (Weber & Brewer, 2008). Further, Mansour (2020) compared numerical ratings of confidence (0-100%) to free report verbal expressions of confidence, finding that both methods were indicative of accuracy – with the numerical ratings being easier to interpret than the verbal expressions. Given that we are advocating for the use of recognition-ratings as a potential alternative to traditional decision judgements, it is important that we ensure our measure is indicative of participants beliefs (i.e., whether they have or have not seen the stimuli before).

Benjamin, Tullis and Lee (2013) also argued that increasing the number of response options presented as part of a scale results in an increase in the amount of noise in the measurement. As expanded upon before in our discussion of Thurstonian and Brusnwikian approaches to uncertainty, recognition memory tasks are already impacted by noise that is independent of memory strength. Thus, if recognition-ratings are to be used to infer the likelihood that a stimuli has been seen before, it is important to reduce the amount of additional noise associated with measurement to allow for the most accurate reflection of the participant’s memory.

Experiment 1

Given that recognition-ratings originated as a means to reduce identification error (e.g., Sauer et al., 2008; 2012), our interest in theoretical foundations for such ratings stemmed from their potential utility in that context. Thus, we chose to initially test our predictions using a face recognition task. Although such a task is not fully generalizable to an eyewitness identification context, it allows us to draw some parallels while also providing a paradigm in which we can efficiently collect large amounts of data, and make estimates based on repeated measures data (reducing some of the noise associated with between-participant manipulations). Also, if should be noted that previous research suggests that face recognition

paradigms provide confidence-accuracy data that generally translate to more realistic identification tasks (e.g., Weber & Brewer, 2004; Sauer et al., 2008).

Experiment 1 examined whether providing additional, non-diagnostic information to participants during the test phase of a face recognition task affected their confidence that they had seen a test stimulus before. This basic paradigm required participants to study a series of faces then, after a delay, complete a series of test trials where, for each test face, they rated their confidence that the face had been presented at study. Images at study were presented either as a full face or the top half of a face (including the top of the nose, the eyes etc.). Test faces consisted of half and full faces, that either had or had not been viewed at study. Participants rated their confidence that the individual depicted had been seen at study. Thus, for the purpose of this experiment, a partial and a full version of the same face constitutes an "old" or studied face (see *Figure 1*).

Figure 1

Face Stimuli for Experiment 1

Note. An example of a full face (left) and the corresponding half face (right).

Participants responded on using either a fine-grained numerical scale (0-100%) or a coarse-grained verbal scale ("low confidence", "moderate confidence", "high confidence"). Our decision to compare a fine-grained numerical scale to a coarse-grained verbal scale stems from confidence accuracy characteristic analysis (CAC; Mickes, 2015). While CAC data is generally collected numerically (i.e., 0-100%), it is then transformed to a categorical measure before being graphed. Typically, researchers have collapsed 0-60% as one category, 70-80% as the next, and 90-100 as the final category (e.g., Mickes 2015; Carlson et al, 2016;

Sauerland et al., 2016). Being able to collect data in its intended form (i.e., in three categories) may remove potential noise from the measurement process that arises from including additional response options, as well as allowing the participants to map their confidence onto verbal labels that better match their internal uncertainty. While there is a confound when it comes to statistically comparing a fine-grained numerical measure to a coarse-grained verbal measure, we will be able to observe overall differences in response patterns. This paradigm allowed us to test a number of hypotheses:

Hypothesis 1: Based on the extant literature, we expected that previously studied faces would receive higher confidence ratings than non-studied faces (Sauer et al., 2008; 2012; Brewer et al., 2012; 2020).

The main interest of this study, however, was to test how additional, non-diagnostic information affected these ratings, and whether effects best aligned with psychophysical or inferential accounts of confidence. Specifically, we were interested in whether (as per inferential accounts) the provision of additional (albeit non-diagnostic) information at test inflated confidence. To answer this question, we focussed primarily on comparing ratings provided when participants viewed a partial face at study followed by the same partial face at test, compared to when they view a partial face at study followed by the corresponding full face at test. Based on the existing literature pertaining to the theoretical basis of metacognitive judgements, as well as the prominent theoretical accounts of confidence from the psychophysics literature, we have determined three competing hypotheses.

Hypothesis 2a: Based on the literature demonstrating the inferential basis of both JOLs and FOKs (Koriat, 1993; Koriat, 1997), and the findings by Busey, Tunnicliff, Loftus & Loftus (2000), we would expect that when participants are provided with additional, non-diagnostic information at test, they will provide higher confidence ratings than when they were provided with the same level of information at test compared to study. Specifically, we

expected confidence ratings to be higher in trials in which participants were shown a full face at test after viewing the corresponding partial face at study, compared to instances in which they are shown a partial face at test after viewing the corresponding partial face at study.

Hypothesis 2b: According to most psychophysical approaches grounded in signal detection and accumulator / balance of evidence frameworks (where confidence indexes stimulus discriminability and the degree of match between a test stimulus and memory for studied stimulus), we would not expect confidence ratings to change with the inclusion of additional non-diagnostic information at test, because the amount of diagnostic information has not changed. Therefore, we would not expect any significant differences in confidence for trials where participants view a partial face at study followed by the corresponding full face at test, or a partial face at study and the same partial face at test.

Hypothesis 2c: Differing from other psychophysical accounts, Baranksi and Petrusic's (1998) doubt scaling model suggests that confidence is inversely related to the amount of non-diagnostic information accumulated. Thus, according to this model, an increase in non-diagnostic information would be expected to produce a decrease in confidence. According to this perspective, we would expect confidence levels to be lower in instances in which participants view a partial face at study followed by the corresponding full face at test, compared to when they view a partial face at study followed by the same partial face at test.

Hypothesis 3: Given the suggestion that different scale types might respond differently to manipulations of uncertainty (Windshcitl & Wells, 1996), we investigated whether any effects of our manipulation differed between participants responding on scales using fine-grained probabilistic values and coarse-grained verbal labels.

Method

Design

We used a 2 (face type at study: full face or partial face) x 2 (face type at test: full face or partial face) x 2 (test face status: old or new) x 2 (confidence scale type: fine-grained or coarse-grained) mixed design, with scale type as the between-participants factor. The dependent variable was participants' recognition rating (confidence) at test. All participants viewed an equal number of new and old faces, and of full and partial faces.

Participants

Sixty participants (43 female), aged 16 to 75 years (M=30.63, SD=14.42), participated in the experiment. We decided on this number to allow for a minimum of 20 observations per between-subjects cell (see Simmons et al., 2011), a decision that guided participant recruitment for all experiments outlined in this paper. First year psychology students received one research credit for their participation, whilst other participants received a $15 gift voucher. Participants were randomly allocated to use either a fine-grained or course-grained scale type for recording confidence.

Stimuli

We used 333 colour photographs of male and female (predominantly Caucasian) faces obtained from databases at Flinders University, the University of Sterling, and the AR Face Database (Martinez & Benavente, 1998). Each of these faces was edited to produce a corresponding "partial face". Partial face versions of the stimuli showed the same facial image but were cropped so only the top half of the face was visible (i.e., showing the tip of the nose and above; see Figure 1). All stimuli and instructions were presented via computer, using purpose-developed experimental software.

Procedure

Participants were tested in groups of up to five people but completed the task individually. To ensure that participants understood their task was to attempt to recognize the individual in the image rather than determine whether the same version of the image had been

studied, during the instruction phase, participants were shown an example of a matching full and partial face (similar to Figure 1), accompanied with the following instructions: “Although the images look different, these faces constitute a correct match as they show the same person”. On the next screen, they were shown a series of three possible “correct matches” (i.e., a pair consisting of two matching partial faces, a pair consisting of two matching full faces, and a pair consisting of one partial face and the matching full face) with the following instructions “In all of the cases here, the correct answer is a “Yes” (i.e., that the test face was studied). You should indicate this through a high confidence rating”. The experiment was divided into 6 blocks of trials, with each block consisting of a study phase and a test phase. In the study phase, participants viewed a sequence of 24 faces each presented for 500ms, with a 500ms inter-stimulus interval (ISI). In the test phase, each participant was shown 48 facial stimuli (half new, half old). For each image, participants indicated how confident they were that they had seen the face at study. There were no time restrictions in the test phase. Participants viewed an equal number of full and partial faces, randomly ordered within blocks, at study and test.

Confidence ratings were recorded on either a fine-grained or a coarse-grained confidence scale, with participants using a mouse to click the on-screen button that corresponded to their level of confidence. The fine-grained scale consisted of 11 numerical points that represented 0-100% (i.e., with on-screen buttons for 0%, 10%, etc.). The coarse-grained scale consisted of a three-point verbal scale, with buttons for “low confidence”, “moderate confidence”, and “high confidence”.

Results

We used linear mixed effects models to analyze our data. Using this approach allowed us to include participants and stimuli as random factors in the models, which allowed random intercepts for these effects. We used both the lme4 package (Bates et al., 2013) and the car

package (Fox & Weisberg, 2019) in R, an open-source language and environment for statistical computing (R Core Team, 2013), to compute the models. The outcomes of these analyses can be interpreted as per a standard linear regression, with the coefficient values in Tables 1 and 2 representing the change in outcome per one unit change in the predictor. We set the reference point for comparison (i.e., the intercept) as non-studied faces, presented as a partial at test. Coefficients represent how recognition-ratings changed relative to this reference point, when faces *were* studied (as either a full or partial face [*FS* and *TS*, respectively]) and when the test face was full (*FT*, cf. partial). The interaction term TS x FT refers to trials in which participants studied a partial face but were tested with the corresponding full face, whereas FS x FT refers to trials in which participants were presented with a full face at study and test. It is not possible to provide errors bars for figures produced based on mixed effects models. Thus, *Figure 2* and *Figure 3* are provided for descriptive purposes only (i.e., to illustrate the patterns in the data). We encourage readers to base their interpretations on the coefficients and associated indices of variance provided in the relevant tables, as these coefficients and indices of variance indicate whether main effects or interactions apparent in the figures are statistically meaningful.

Our primary aim was to determine whether additional non-diagnostic information would increase ratings (as per inferential models of metacognition), have no effect on ratings, (as per psychophysical models of confidence), or decrease ratings. Evidence of an inferential approach would be provided if, for the data points labeled “Top only” on the X-axis of *Figure 1* and *Figure 2* (i.e., referring to nature of the stimulus at study), confidence is higher for the “Test: full face” bar than the “Test: top only” bar (as, compared to partial face stimuli, full face stimuli provide *more* information at test). This represents trials in which participants viewed a partial face at study, and the corresponding full face at test. Evidence consistent with the more standard psychophysical approach (i.e., where confidence indexes stimulus

discriminability) will be provided if, for the data points labeled “Top only” on the X-axis of *Figure 1* and *Figure 2*, confidence is the same for the “Test: full face” and the “Test: top only” bar (because the amount of diagnostic information present in the test stimuli is the same for the full and partial test faces). Evidence for the doubt-scaling model (Baranski & Petrusic, 1998) will be provided if, for the data points labeled “Top only” on the X-axis of *Figure 2* and *Figure 3*, confidence is lower for the “Test: full face” bar than the “Test: top only” bar (as these stimuli present information not encountered at study, and this additional information is therefore ambiguous).

However, before addressing this question, we first confirmed that our data demonstrated typical patterns of confidence-based discrimination between studied and non-studied faces. Consistent with Hypothesis 1, and findings previously reported in the literature (Sauer et al., 2008, 2012; Brewer et al., 2012, 2020), for both scale types confidence ratings were higher for test stimuli that had been viewed at study than for stimuli that had not. This is visible in Figures 2 and 3, for responses on the fine- and coarse-grained confidence scales, respectively.

To determine whether recognition-ratings were best accounted for by an inferential or psychophysical account, the specific trials of interest are those in which participants viewed a partial face at study followed by the same partial face or the corresponding full face at test. The results of these trials are labeled “Top only” on the X-axes of Figures 2 and 3 for the fine- and coarse-grained scales, respectively. Both figures show that in trials where participants viewed a partial face at study followed by the same partial face at test, their confidence ratings were notably higher than in trials where they viewed a partial face at study followed by the corresponding full face at test. Thus, consistent with the (psychophysical) doubt scaling model, the presentation of additional, non-diagnostic information at test

reduced confidence. These effects are represented by the *TS* × *FT* coefficients in Tables 1 and 2. However, as discussed below, this initial support comes with an important caveat.

Table 1

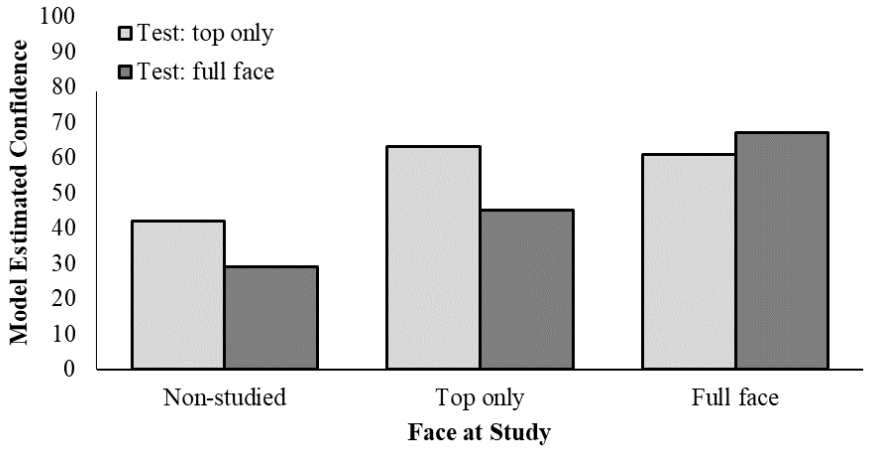
Fixed effect coefficients for linear mixed-effects model predicting confidence on fine- and coarse-grained (verbal) scales.

Factors	Coefficient	*SE*	*t*	*p*	95% CI
			Fine-grained scale		
Intercept	41.50	1.70	24.294	<.001***	[38.10, 44.87]
Top at study (TS)	21.68	1.10	19.574	<.001***	[19.55, 23.88]
Full at study (FS)	19.09	1.10	17.263	<.001***	[16.91, 21.22]
Full at test (FT)	-12.08	0.90	-13.352	<.001***	[-13.81, -10.30]
TS × FT	-5.97	1.56	-3.809	<.001***	[-9.10, -2.97]
FS × FT	18.22	1.56	11.636	<.001***	[15.17, 21.30]
			Coarse-grained scale		
Intercept	0.64	0.04	16.75	<.001***	[0.57, 0.71]
Top at study (TS)	0.47	0.03	16.36	<.001***	[0.41, 0.52]
Full at study (FS)	0.42	0.03	14.76	<.001***	[0.36, 0.48]
Full at test (FT)	-0.22	0.02	-9.47	<.001***	[-0.27, -0.18]
TS × FT	-0.17	0.04	-4.23	<.001***	[-0.25, -0.09]
FS × FT	0.35	0.04	8.60	<.001***	[0.27, 0.43]

Note. * p <.05, ** p <.01, *** p <.001.

Figure 2

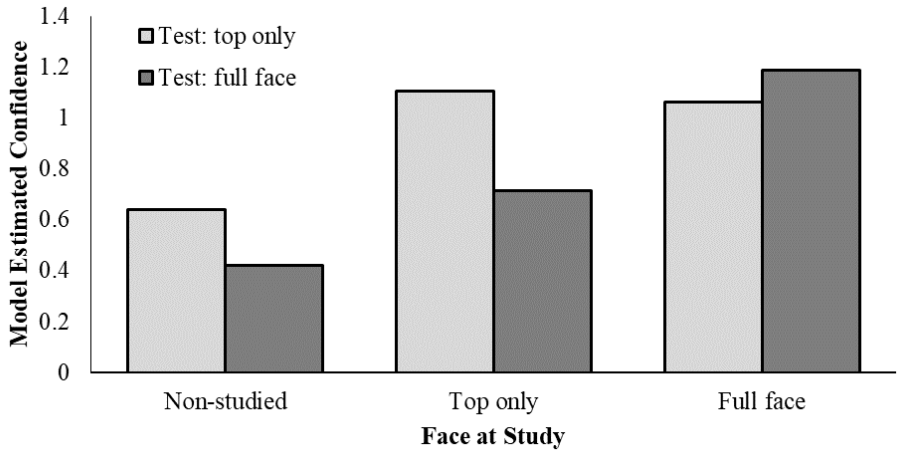
Confidence for the Fine-Grained Scale



Note. The model-estimated mean confidence ratings for participants using the fine-grained (numeric) scale, based upon the amount of information provided at study and test.

Figure 3

Confidence for the Coarse-Grained Scale



Note. The model-estimated mean confidence ratings for participants using the coarse-grained (verbal) scale, based upon the amount of information provided at study and test.

Interestingly, although a direct inferential comparison between the fine- and coarse-grained scales is problematic, visual inspection of Figures 2 and 3 and the coefficients in

Tables 1 and 2 suggest that the patterns of results are very similar across scale-type conditions.

Additional exploratory results

The following results, though not necessary to test our key hypotheses, provide important context for interpreting our results. Confidence ratings for new faces were lower for full face stimuli than partial face stimuli (see Figures 2 and 3), demonstrating that participants were better able to distinguish that they had not studied the face before when the full stimulus was presented at test. Confidence was significantly higher in trials where participants viewed a full face followed by the same full face, compared to trials in which they saw a full face followed by the corresponding partial face (see Table 2). This may suggest a potential effect of holistic facial processing or transfer appropriate processing. We return to this issue, and the broader implications it has for interpreting our findings, in the Discussion.

Table 2

Fixed effect coefficients for linear mixed-effects model comparing confidence between that were partial at study vs full at study when a full face was presented at test, on fine- and coarse-grained (verbal) scales

Factors	Coefficient	*SE*	*t*	*p*	95% CI
			Fine-grained scale		
Intercept	60.83	2.31	26.33	<.001***	[56.30, 65.26]
Full at study (FS)	5.84	1.32	4.42	<.001***	[3.23, 8.46]
			Coarse-grained scale		
Intercept	1.06	0.05	20.84	<.001***	[0.96, 1.16]
Full at study (FS)	0.12	0.04	3.5	<.001***	[0.06, 0.19]

Note. * p <.05, ** p <.01, *** p <.001.

Discussion

Experiment 1 investigated whether recognition-ratings were best accounted for by psychophysical or inferential models of metacognition. Inferential models of metacognition would propose that provided the degree of match (i.e., diagnostic information) remained constant, any additional information available at test, regardless of the diagnosticity of that information, would result in higher recognition-ratings (Busey et al., 2000; Koriat, 1993; 1997). In contrast, psychophysical models generally suggest that confidence indexes stimulus discriminability (e.g., Green & Swets, 1966; Vickers, 1970; Wixted, 2007). Thus, additional non-diagnostic information should not increase confidence. However, the doubt-scaling model argues that non-diagnostic information should decrease confidence by increasing ambiguity (Baranksi & Petrusic, 1998).

In short, Experiment 1 demonstrated that when participants viewed a partial face at study followed by the corresponding full face at test, their confidence ratings were generally lower than when they viewed a partial face at study followed by the same partial face at test. Thus, consistent with Baranski and Petrusic's (1998) doubt scaling model but counter to predictions drawn from inferential models (that tie confidence to the *amount* of information available) and typical psychophysical models (which posit that confidence indexes diagnostic information), additional non-diagnostic information at test *reduced* confidence in recognition. This finding appears to apply to both response scales.

Interestingly, our findings are at odds with those reported by Busey et al. (2000), who found that providing additional non-diagnostic information at test (i.e., by increasing the luminance of the test stimulus; a manipulation that did not increase accuracy) increased confidence ratings compared to when this additional information was not present. We have identified several potential explanations for this discrepancy.

Task demands. The simplest explanation may lie in the way in which non-diagnostic information was operationalized in our experiment compared to that of Busey et al. By changing the luminescence of the entire stimuli, Busey et al. changed the *nature* of the entire stimuli: changing the appearance of the stimulus *as a whole*. In contrast, by adding or removing the bottom half of the face while always leaving the top half (i.e., the part necessary to determine whether the stimuli was old or new) intact, we provided additional information without affecting the degree of match between the stimulus at study and the studied portion of the test face (i.e., the top half of the face, Figure 1). It could be that changing the entire stimulus changed the perceived diagnostic value of the additional information, compared to our manipulation, which simply added information. Without a direct comparison of these manipulations, however, this explanation is speculative. Another potential explanation for the difference between our results and Busey et al.'s could lie in the way in which they manipulated luminance. The brightest stimuli were scaled to 80 cd/m2, which Busey et al. described themselves as being "essentially white" (p.32), and the darkest stimuli were scaled to 10 cd/m2, which resulted in a very dark image (Busey et al. describe 5 cd/m2 as "essentially black", p.32). This extreme variance in luminance could have produced stimuli at the brightest and darkest ends of the manipulation that contained very little diagnostic information, which may have left participants almost entirely reliant on inferential cues.

Holistic facial processing. Compared to other stimuli, faces tend to rely more on holistic processing (Maurer, Le Grand & Mondloch, 2002; Richler & Gauthier, 2014; Tanaka & Gordon, 2011). Face processing relies not only on an individual's ability to recognize individual facial features, but their ability to recognize the configural relations between features (Maurer, Le Grand & Mondloch, 2002). This reliance on holistic processing may also contribute to the differences observed between Busey et al.'s (2000) findings and ours.

By altering the luminescence of the facial stimuli, Busey et al. were trying to replicate an environmental variable (i.e., changes in brightness in the environment between study and test conditions). In contrast, we were modifying the physical makeup of the face by removing certain facial features. When participants were comparing a full face at test with their memory of the corresponding partial face at study, they not have perceived the test stimulus as a "matching stimulus with some additional information" but as a stimulus that, *as a whole,* did not match well with the partial image encoded at study. This decrease in holistic match could account for the lower confidence ratings observed. Thus, according to this view, it is the reduction in match (in a holistic sense) rather than the addition of specifically non-diagnostic information, that affected ratings.

That ratings were significantly lower in trials where participants viewed a full face at study and a partial face at test compared to when they saw full face at study and a full face at test further supports the idea that holistic facial processing (and/or transfer appropriate processing, discussed below) may be playing a part (see Tables 3 and 4). If the holistic nature of facial processing caused a mismatch between stimuli in partial face at study and full face at test trials, the same effect should be evident for full face at study and partial face at test trials. However, this is not what we observed. The decrease in recognition ratings associated with adding non-diagnostic information at test (partial study-full test vs partial study-partial-test) was greater than the decrease in ratings associated with removing information at test (full study-partial test vs full study-full test; see *Figure* 2 and *Figure* 3). This may suggest a combined effect holistic facial processing and a doubt-scaling mechanism on participants' confidence ratings. To elaborate, the doubt-scaling model would predict that confidence should be higher in full face at study and partial face at test trials than partial face at study and full face at test trials, as there is no non-diagnostic information present in the full face at study and partial face at test trials. Therefore, the finding that full face at study and partial

face at test and partial face at study and full face at test produced lower confidence ratings than full face at both study and test and partial face at both study and test respectively could be a result of holistic facial processing. However, the fact that partial face at study and full face at test trials showed a larger reduction in confidence than full face at study and partial face at test trials (compared to partial face at both study and test and full face at both study and test respectively) suggests a doubt-scaling effect.

Transfer appropriate processing. The concept of transfer appropriate processing (TAP) suggests that participants' memory for a stimulus should be improved if they are tested under the same conditions in which they encoded the stimulus (Franks, Bilbrey, Lien & McNamara, 2000; Morris, Bransford & Franks, 1977). According to this account, we would expect confidence to be lower for all trials in which the amount of face information provided at test differed from that at study, regardless of whether participants were gaining non-diagnostic information or losing diagnostic information. This was reflected in our results, which showed that confidence was higher when participants studied a partial face and were tested with a partial face (cf., tested with a full face), and in trials in which participants studied a full face and were tested with a full face (cf., tested with a partial face).

Due to these alternative yet plausible explanations for our results, Experiments 2 and 3 tested whether holistic facial processing (Experiments 2 and 3) and transfer appropriate processing (Experiment 3) contributed to the observed pattern of results.

Experiment 2

The results from Experiment 1 were consistent with a doubt-scaling mechanism, where confidence is negatively associated to the amount of non-diagnostic information at test. However, we cannot rule out the contribution of a mechanism relating to the role of holistic processing in face perception. Consequently, Experiment 2 aimed to attenuate the potential contribution of holistic processing effects and test whether the results persist.

Experiment 2 retains the basic design from Experiment 1 but, whereas Experiment 1 used facial images, Experiment 2 uses images of houses and landscapes. This approach allowed us to isolate the effects of non-diagnostic information at test (i.e., independent of holistic processing effects), as neither houses nor landscapes should rely on holistic processing. While images of houses were chosen due to their conceptual similarity to faces (clearly defined edges, roughly symmetrical features, etc.; Filliter et al., 2016), landscapes were chosen due to their undefined edges and lack of symmetry. Thus, the likelihood of holistic processing for such an image is low. Replicating the pattern of results from Experiment 1 with stimuli that rely less on holistic processing would provide more compelling evidence for a doubt-scaling model of confidence in recognition. We retained the manipulation of scale type to see if the generality of findings across the two scales replicates.

Method

Design

We used a 2 (stimuli type: houses or landscapes) x 2 (stimuli type at study: full image or partial image) x 2 (stimuli type at test: full image or partial image) x 2 (test image status: old or new) x 2 (confidence scale type: fine-grained or coarse-grained) mixed design, with scale type and stimuli as the between-participants factors. The dependent variable was participants' recognition rating at test. All participants viewed an equal number of new and old images, and of full and partial images.

Participants

113 participants (75 female), aged 16 to 48 years participated in the experiment. First year psychology students received one research credit for their participation, whilst others received a $15 gift voucher. Participants were randomly allocated to use either a fine-grained or course-grained scale type for recording confidence.

Stimuli

432 colour photographs of landscapes as well as 422 photographs of houses were obtained from Google Images. Each of these photographs was edited to produce a corresponding "partial" stimulus. All stimuli and instructions were presented via computer, using purpose-developed experimental software.

Figure 4

Landscape Stimuli

Note. An example of a full landscape (left) and the corresponding half landscape (right).

Figure 5

House Stimuli

Note. An example of a full house (left) and the corresponding half house (right).

Procedure

The procedure was identical to that described for Experiment 1. To reiterate the key points: the experiment was divided into 6 blocks of trials, with each block consisting of a study phase and a test phase. In the study phase, participants viewed a sequence of 24 images each presented for 500ms, with a 500ms inter-stimulus interval (ISI). In the test phase, each participant was shown 48 images (half new, half old). For each image, participants indicated

how confident they were that they had seen the image at study. There were no time restrictions in the test phase. Participants viewed an equal number of full and partial images, randomly ordered within blocks, at study and test.

Confidence ratings were recorded on either a fine-grained or a coarse-grained confidence scale, with participants using a mouse to click the on-screen button that corresponded to their level of confidence. The fine-grained scale consisted of 11 numerical points that represented 0-100% (i.e., with on-screen buttons for 0%, 10%, etc.). The coarse-grained scale consisted of a three-point verbal scale, with buttons for "low confidence", "moderate confidence", and "high confidence".

Results

We used the same analytical approach as outlined for Experiment 1. Like Experiment 1, the outcomes of our analyses can be interpreted as per a standard linear regression, with the coefficient values in Tables 3 and 4 representing the change in outcome per one unit change in the predictor. We set the reference point for comparison (i.e., the intercept) as non-studied images, presented as a partial at test. Coefficients represent how recognition-ratings changed relative to this reference point, when images *were* studied (as either a full or partial image) and when the test image was full (c.f., partial). It is not possible to provide errors bars for figures produced based on mixed effects models. Thus, *Figure 6* and *Figure 7* are provided for descriptive purposes only (i.e., to illustrate the patterns in the data). Like Experiment 1, we encourage readers to base their interpretations on the coefficients and associated indices of variance provided in the relevant tables, as these coefficients and indices of variance indicate whether main effects or interactions apparent in the figures are statistically meaningful.

The primary aim of this experiment was to determine if the previous results consistent with the doubt-scaling model generalized to a task where participants were less reliant on

holistic processing. As a reminder, evidence for the doubt-scaling model would be provided if, for the columns labeled “Top only” on the X-axis of Figure *6* and *Figure 7*, confidence is lower for the “Test: full stimulus” bar than the “Test: top only” bar, as this would indicate that providing additional non-diagnostic information reduces confidence. Both *Figure 5* and *Figure 6* demonstrate this effect, indicating initial support for a doubt-scaling model of confidence.

Table 3

Fixed effect coefficients for linear mixed-effects models predicting confidence for landscape stimuli on both a fine-grained and coarse-grained scale.

Factors	Coefficient	*SE*	*t*	*p*	95% CI
			Fine-grained scale		
Intercept	33.48	2.21	15.11	<.001***	[29.14, 37.95]
Top at study (TS)	28.60	1.18	24.25	<.001***	[26.33, 30.91]
Full at study (FS)	21.37	1.18	18.13	<.001***	[19.13, 23.68]
Full at test (FT)	-6.82	0.96	-7.08	<.001***	[-8.68, -4.94]
TS × FT	-8.41	1.67	-5.05	<.001***	[-11.69, -5.09]
FS × FT	18.80	1.67	11.26	<.001***	[15.58, 21.99]
			Coarse-grained scale		
Intercept	0.62	0.04	14.18	<.001***	[0.53, 0.71]
Top at study (TS)	0.66	0.03	22.93	<.001***	[0.61, 0.72]
Full at study (FS)	0.52	0.03	18.02	<.001***	[0.46, 0.57]
Full at test (FT)	-0.17	0.02	-7.14	<.001***	[-0.21, -0.12]
TS × FT	-0.16	0.04	-3.91	<.001***	[-0.24, -0.08]
FS × FT	0.39	0.04	9.66	<.001***	[0.32, 0.47]

Note. * p <.05, ** p <.01, *** p <.001.

Table 4

Fixed effect coefficients for linear mixed-effects models predicting confidence for house stimuli on both a fine-grained and coarse-grained scale.

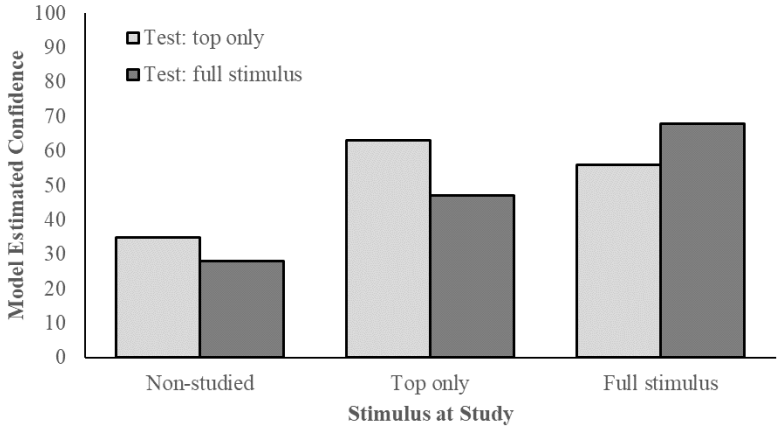
Factors	Coefficient	*SE*	*t*	*p*	95% CI
			Fine-grained scale		
Intercept	38.70	2.15	17.96	<.001***	[34.58, 42.85]
Top at study (TS)	17.51	1.26	13.94	<.001***	[15.06, 20.02]
Full at study (FS)	19.02	1.26	15.14	<.001***	[16.52, 21.40]
Full at test (FT)	-6.10	1.02	-5.96	<.001***	[-8.08, -4.12]
TS × FT	-1.26	1.78	-0.71	0.477	[-4.74, 2.23]
FS × FT	7.03	1.78	3.96	<.001***	[3.64, 10.53]
			Coarse-grained scale		
Intercept	0.66	0.06	11.82	<.001***	[0.54, 0.76]
Top at study (TS)	0.42	0.03	14.74	<.001***	[0.36, 0.48]
Full at study (FS)	0.40	0.03	13.86	<.001***	[0.34, 0.45]
Full at test (FT)	-0.10	0.02	-4.48	<.001***	[-15, -0.06]
TS × FT	-0.09	0.04	-2.13	0.034*	[-0.16, -0.00]
FS × FT	0.27	0.04	6.68	<.001***	[0.19, 0.35]

Note. * p <.05, ** p <.01, *** p <.001.

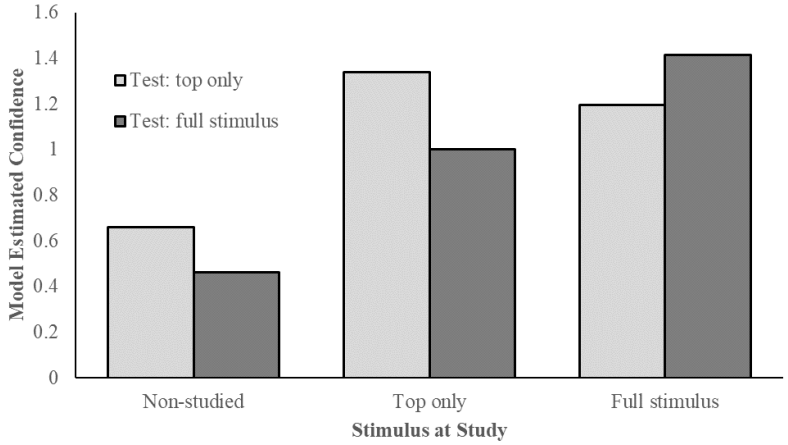
Figure 6

Confidence for Landscape Stimuli

Fine-Grained Scale



Coarse-Grained Scale

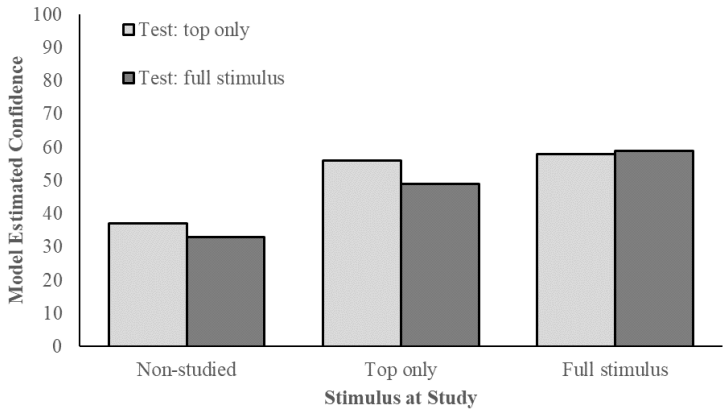


Note. The model-estimated mean confidence ratings for participants who viewed landscape images using the fine-grained (numeric) scale and a coarse-grained (verbal) scale based upon the amount of information provided at study and test.

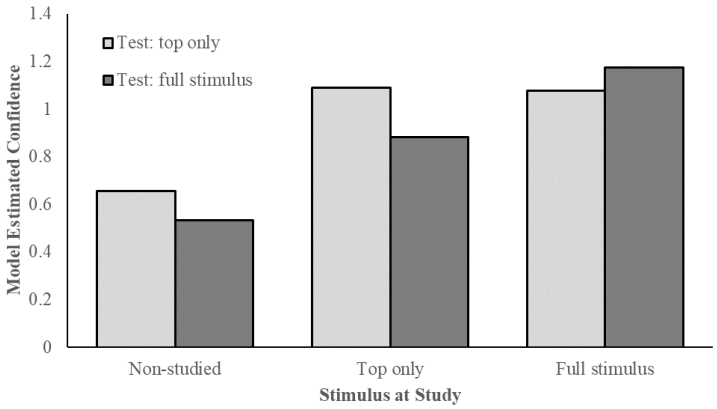
Figure 7

Confidence for House Stimuli

Fine-Grained Scale



Coarse-Grained Scale



Note. The model-estimated mean confidence ratings for participants who viewed house images using the fine-grained (numeric) scale and a coarse-grained (verbal) scale based upon the amount of information provided at study and test.

Additional exploratory results

Again, while these results are not necessary to our key hypotheses, they provide important context with which to interpret our results As in Experiment 1, recognition ratings for new faces were lower for full test stimuli than partial test stimuli (see *Figures 6 and 7*), demonstrating again that participants were better able to distinguish that they had not studied the image before when the full stimuli was presented at test.

Confidence was significantly higher in trials where participants viewed a full landscape image followed by the same full landscape image compared to trials in which they saw a full landscape followed by the corresponding partial landscape (see Table 5), however this effect was not significant for houses in the fine-grained condition (see Table 6). The results for participants in the landscape condition suggest that transfer appropriate processing may still be playing a role, something addressed in Experiment 3.

Table 5

Fixed effect coefficients for linear mixed-effects model comparing confidence for landscape images that were partial at study vs full at study when a full landscape presented at test, on fine- and coarse-grained (verbal) scales

Factors	Coefficient	*SE*	*t*	*p*	95% CI
			Fine-Grained		
Intercept	54.80	2.47	23.06	<.001	[50.11, 59.46]
Full at study (FS)	12.15	1.49	8.15	<.001	[9.18, 15.17]
			Coarse-Grained		
Intercept	1.15	4.90	23.40	<.001	[1.05, 1.24]
Full at study (FS)	2.25	3.52	6.39	<.001	[1.64, 3.01]

Note. * p <.05, ** p <.01, *** p <.001.

Table 6

Fixed effect coefficients for linear mixed-effects model comparing confidence for house images that were partial at study vs full at study when a full house was presented at test, on fine- and coarse-grained (verbal) scales

Factors	Coefficient	*SE*	*t*	*p*	95% CI
			Fine-Grained		
Intercept	58.81	2.47	23.39	<.001	[52.96, 62.74]
Full at study (FS)	8.68	1.58	0.55	0.55	[-21.50, 39.44]
			Coarse-Grained		
Intercept	1.05	5.71	18.40	<.001	[0.93, 1.16]
Full at study (FS)	1.57	3.51	4.47	<.001	[0.89, 2.25]

Note. * p <.05, ** p <.01, *** p <.001.

Discussion

In Experiment 2, we replicated the results from Experiment 1 using stimuli that were less reliant on holistic processing. Consistent with Experiment 1, Experiment 2 found that in trials where participants had viewed the stimuli at study, recognition-ratings were lower when additional, non-diagnostic information was presented at test compared to trials in which non-diagnostic information was not present at test. Replicating this pattern with house and landscape stimuli means we can be confident that the pattern of results found in Experiment 1 was neither face-specific nor reliant on holistic processing. This provides support for Baranski and Petrusic's doubt-scaling model (1998), suggesting that confidence indexes not only the degree of match between studied and tested items but also, inversely, the amount of non-diagnostic information present when making a recognition judgment.

What is not yet clear, however, is whether this difference in confidence ratings is truly the result of a doubt-scaling mechanism, or rather a demonstration of transfer appropriate processing. Like Experiment 1, for landscape stimuli at least, we still see a reduction in confidence when participants viewed a full landscape at study followed by the corresponding partial landscape at test (cf. full landscape followed by the same full landscape), which cannot be explained solely by the doubt-scaling model (Baranski & Petrusic, 1998). The doubt-scaling model would only predict a decrease in recognition-ratings when non-diagnostic information is present, whereas transfer appropriate processing predicts a decrease in recognition-ratings associated with any differences between how the image appears at encoding compared to test. Interestingly, this effect was not found for house stimuli, despite the fact that houses seem to share more conceptual similarities with faces than landscapes do (e.g., symmetry, defined edges etc.). Thus, we would expect holistic processing to be more likely for houses than landscapes, given the nature of the images. In sum, although the results from Experiment 2 speak against an explanation for our results based on a holistic-processing mechanism, aspects of the data remain consistent with a transfer appropriate processing mechanism.

Experiment 3

Experiments 1 and 2 provide results consistent with Baranski and Petrusic's (1998) doubt-scaling model of confidence. However, based on our findings, we are currently unable to rule out to the contribution of a transfer appropriate processing mechanism. Thus, Experiment 3 tested the effects of non-diagnostic information on confidence in a perceptual discrimination paradigm, removing reliance on memory (and therefore the potential for TAP to affect confidence ratings). Similar to Experiments 1 and 2, we recorded confidence in the absence of a binary decision. As this task does not rely on recognition memory, we will simply refer to these ratings as "confidence ratings" (or confidence).

In this task, participants viewed a dynamic grid consisting of blue, orange, and (sometimes) white pixels (see Figure 8). The pixels were currently rearranging, but the total number of each colour pixel remained stable throughout the trial. Participants then indicated their confidence, from 0-100%, that the grid was "predominantly blue" or "predominantly orange" (see Figure 9 for a static image of the grid stimuli). In this paradigm, white pixels are used to operationalize non-diagnostic information (NDI). We included four levels of non-diagnostic information (0, 1/6, 1/3 and ½), allowing us to more thoroughly investigate the nature of the relationship between non-diagnostic information and confidence.

Although this paradigm allows a controlled manipulation of NDI, it also risks a confound: Simply increasing the amount of NDI involves a corresponding decrease in diagnostic information if the total amount of information (i.e., the total number of pixels in the grid) remains constant. For this reason, we also included a between groups manipulation of diagnostic information, operationalized through the use of different "grid-types". In the "stable" condition, the dynamic grid remained the same size regardless of the amount of non-diagnostic information present. This means that each increase in non-diagnostic information lead to a corresponding decrease in diagnostic information. The additive grid, however, increased in size to account for the additional non-diagnostic information. Thus, the total number of diagnostic pixels never decreases. Including this manipulation allowed us to comment on whether any relationship observed between non-diagnostic information and confidence is actually only due to a decrease in available diagnostic information rather than to an increase in non-diagnostic information.

Based on Baranski and Petrusic's (1998) doubt-scaling model and the results of Experiments 1 and 2, we predicted that confidence would decrease as the proportion of non-diagnostic information increased. Further, we expected this effect to persist regardless of whether there was a reduction in the total amount of diagnostic information.

Method

Design

We used a 2 (grid-type: stable or additive) × 4 (proportion of non-diagnostic information: 0, 1/6, 1/3, 1/2) × 2 (majority colour: blue vs. orange) mixed design, with grid-type as the between-subjects factor. The dependent variable was participants' recognition-rating at test.

Participants

80 participants[1] were recruited through Mechanical Turk, with one participant being excluded for consistently not responding to trials. They were reimbursed $3 USD for their participation. Participants with a history of epilepsy/related disorders were not able to participate, due to the "flashing" nature of the stimuli.

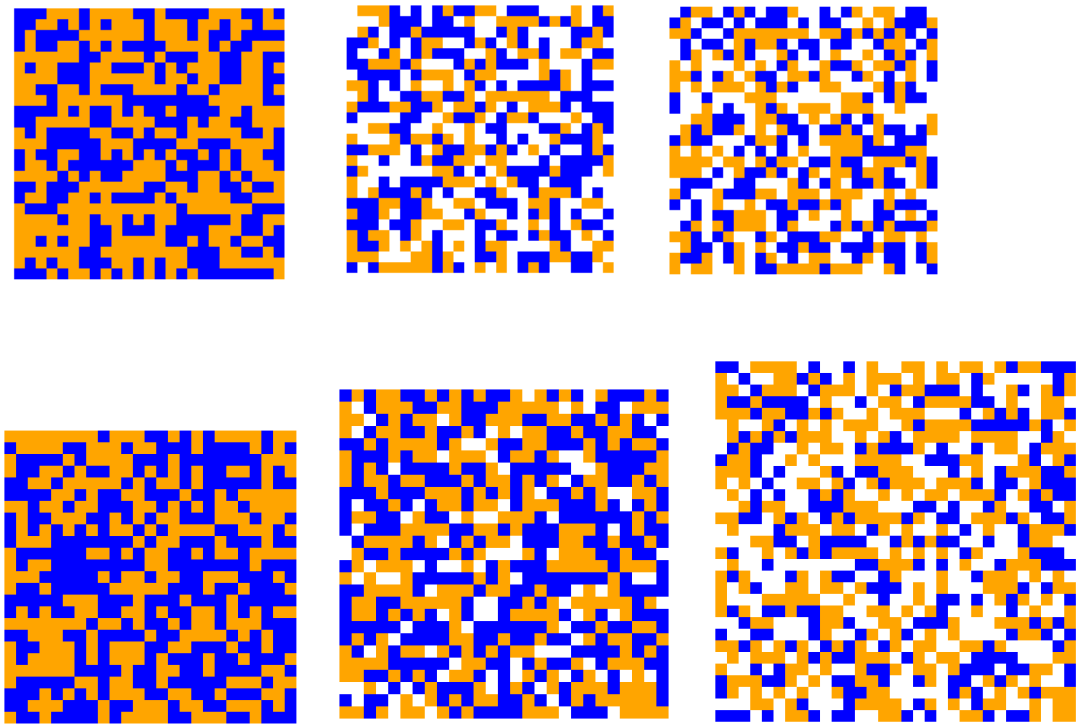
Stimuli

For each trial, participants viewed a dynamic grid consisting of blue, orange and sometimes white pixels (see Figure 8). Although the pixels in the grid were constantly rearranging, the proportion of blue (RGB = 0, 65, 255), orange (RGB = 255, 127, 0), and white (RGB = 128, 128, 128) pixels remained constant in each trial.

[1] Due to an error with the experimental software, data relating to age and sex were not collected.

Figure 8

Examples of Dynamic Grid Stimuli at Varying Levels of Non-Diagnostic Information (i.e., White Pixels). The First Row Represents the Stable Condition, Whereas the Second Row Represents the Additive Condition.



Note: Examples of the dynamic-grid stimulus at varying levels of non-diagnostic information. Individual pixels would randomly cycle through the available colour options (orange/blue or orange/blue/white) while keeping the proportion of each colour constant.

In the "stable" condition, the overall size of the grid remained stable (625 pixels). Thus, as the overall proportion of non-diagnostic pixels (i.e., white pixels) increased, the number of diagnostic pixels (i.e., blue and orange pixels) decreased accordingly. In the "additive" condition, the dynamic grid increased in size as the proportion of non-diagnostic pixels increased, meaning the number of diagnostic pixels remained constant regardless of the number of non-diagnostic pixels. When the grid contained no non-diagnostic information, it consisted of 625 total pixels. At 1/6 non-diagnostic information it consisted of 729 total

pixels, at 1/3 non-diagnostic information it consisted of 961 total pixels, and at ½ non-diagnostic information it consisted of 1250 total pixels.

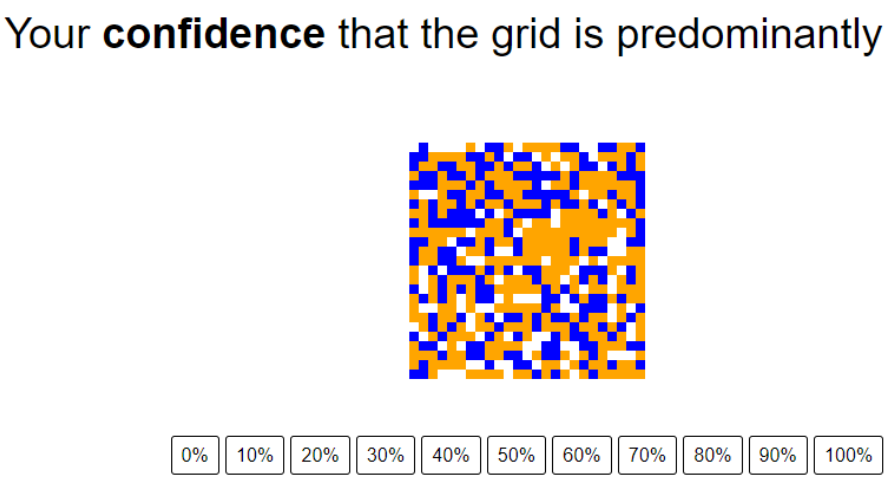
Procedure

The task was conducted via Mechanical Turk, with participants being informed that they must complete the task on a laptop or computer (rather than tablet or phone). The participants first completed a series of 40 training trials, separated into two blocks. In the first block, they were presented with a dynamic grid consisting of blue and orange pixels, and were asked to indicate their confidence as to whether the grid was primarily blue/orange from 0-100% (see Figure 9). After each trial, they were told whether the grid was in fact primarily blue or primarily orange. In the second practice block, white pixels were added to the grid to constitute non-diagnostic information. This was to allow participants to adjust to the task demands (practice block 1) before introducing the more complex stimuli (practice block 2).

Participants completed six experimental blocks, each consisting of 50 trials. Unlike the practice blocks, participants did not receive feedback for the experimental trials. Participants were encouraged to take rests between blocks as required. The task took approximately half an hour to complete.

Figure 9

Example of Trial and Stimulus



Note: Participants indicated their confidence by selecting the corresponding button. The target colour (i.e., blue or orange) differed between blocks of trials.

Results

Again, we conducted linear mixed effects models. Participant was included as a random effect to allow for random intercepts for individual participants. We set the reference point for comparison (i.e., the intercept) as 0% non-diagnostic information in the additive condition. Coefficients represent how confidence changed relative to this reference point, when non-diagnostic information is included (at either 1/6, 1/3 or 1/2) and when the grid type was stable (cf. additive).

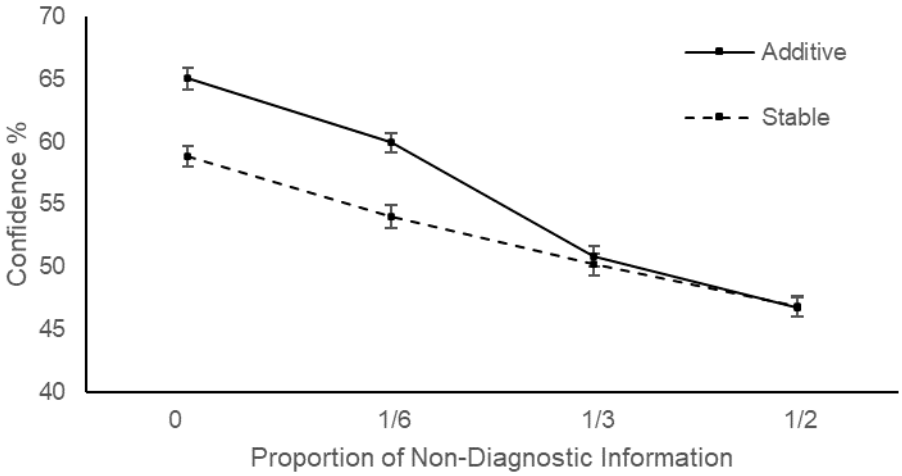
A model including both grid style and proportion of non-diagnostic information showed that confidence decreased in a monotonic fashion as the proportion of non-diagnostic information increased (see Figure 10). Each increase in non-diagnostic information produced a significant reduction in confidence (see Table 7).

While there was no main effect of grid-type, there does appear to be an interaction between grid-type and the proportion of non-diagnostic information (see Table 5).

Specifically, the reduction in confidence for the 1/3 and ½ trials was larger for the stable grid compared to the additive grid. This is demonstrated in Figure 10, as well as by the coefficients in Table 5. These results together suggest that while non-diagnostic information may be the primary force driving the reduction in confidence, that a reduction in diagnostic information (caused by an increase in non-diagnostic information in the stable condition) may amplify the effects.

Figure 10

The Effect of Increasing Non-Diagnostic Information on Confidence for Stable and Additive Grid Types.



Note. Mean confidence at each level of non-diagnostic information. Y-axis has been truncated for easier interpretation of patters (c.f., 0-100% scale). Error bars represent estimated standard error, and are for demonstrative purposes only.

Table 7

Fixed effect coefficients for linear mixed-effects models predicting confidence as a function of non-diagnostic information and grid-type

Factors	Coefficient	*SE*	*t*	95% CI	*p*
Intercept	59.63	2.22	26.81	[38.10, 44.87]	<.001***
White ratio 1/2	-12.68	1.19	-10.70	[19.55, 23.88]	<.001***
White ratio 1/3	-9.00	1.19	-7.54	[16.91, 21.22]	<.001***
White ratio 1/6	-4.89	1.20	-4.08	[-13.81, -10.30]	<.001***
Stable Grid	5.15	3.14	1.64	[-9.10, -2.97]	1.64
White ½:Stable	-4.89	1.68	-3.18	[15.17, 21.30]	0.001**
White 1/3:Stable	-4.91	1.68	-2.92	[0.57, 0.71]	0.003**
White 1/6:Stable	0.10	1.69	-0.06	[0.41, 0.52]	0.95

Note. * p <.05, ** p <.01, *** p <.001.

Discussion

Our findings provide strong support for Baranski and Petrusic's (1998) doubt-scaling model of confidence. Critically, confidence decreased systematically as the amount of non-diagnostic information present at the time of the decision increased. This finding generally persisted regardless of whether the total amount of diagnostic information decreased accordingly or remained stable, although the effect was largest for the 1/3 and ½ non-diagnostic information trials in the stable condition. This suggests that while confidence may in-fact be inversely related to the amount of non-diagnostic information present test, that a reduction in diagnostic information may amplify such effects.

Experiment 3 returns three critical, but simple findings. First, there is a systematic, negative relationship between non-diagnostic information and confidence. Second, the fact that this effect is present in the stable grid condition, coupled with the lack of main effect for

grid-type, suggests that this relationship is not purely the result of a decrease in diagnostic information. Third, having no memory component removes the possibility of transfer appropriate processing having an effect.

General Discussion

Across three experiments, we found evidence supporting Baranski and Petrusic's doubt-scaling model of confidence. First, we consistently found that an increase in non-diagnostic information was associated with a corresponding decrease in confidence. Further, this effect was consistent across a variety of stimuli (i.e., faces, houses, landscapes and a dynamic grid) and experimental conditions (i.e., recognition memory vs perceptual discrimination). Second, this finding is not attributable to holistic processing, as it persisted when using stimuli that are not considered to rely heavily on holistic processing. Third, the reduction in confidence is not the result of a transfer appropriate processing mechanism (Morris, et al., 1977), as it persisted in Experiment 3, which did not contain a memory component. Finally, we can conclude that the reduction in confidence is not attributable to a reduction in diagnostic information. While a reduction in diagnostic information may amplify the effects (see Experiment 3 for elaboration), we still observe a monotonic reduction in confidence-ratings associated with an increase in non-diagnostic information when the amount of diagnostic information remains stable.

Theoretical Implications

First and foremost, these findings suggest that the doubt-scaling model of confidence (1998) provides a compelling theoretical account for recognition-ratings when the level of non-diagnostic information at test is manipulated at test. Although Experiment 3 did not involve a recognition memory component, we believe that when the results are interpreted in conjunction with those from Experiments 1 and 2 (which both involved participants completing a recognition memory task) that they paint a clear picture. Further, theories that

originated in the psychophysics literature have previously been generalized to account for confidence in recognition memory, such as Signal Detection Theory (Green & Swets, 1966) or accumulator models (Van Zandt, 2000).

More broadly, these findings suggest that existing theories of confidence need to consider not only how information that contributes to signal strength and discrimination factors into confidence ratings, but also how non-diagnostic information influences confidence. For a more comprehensive look at how non-diagnostic information influences confidence ratings, decision accuracy and reaction time in a perceptual task, see Kohl et al. (2022).

Applied Implications

There are a variety of applied domains in which non-diagnostic information may realistically be present when decisions are being made. For example, when a security screener is tasked with identifying a threat from a luggage x-ray, there will likely be additional, non-threat items within the bag. Given our interest in face-recognition (Experiment 1), one area of applied interest that seems particularly salient is that of eyewitness identification. Our findings suggest that a witness's confidence in an identification from a typical lineup would be comparatively low if the perpetrator had been wearing a face-mask while committing the crime. This provides further support, as well as theoretical justification, for Manley et al.'s (2019) paper, which showed that providing additional, non-diagnostic facial information at test reduced confidence and accuracy in a face recognition task. Further, Experiment 3 suggests that our findings expand beyond the realm of recognition memory and into confidence for perceptual tasks. Thus, the presence of non-diagnostic information may also decrease confidence in applied perceptual tasks, such as luggage or medical screening.

It has been found that essay length (regardless of content) is predictive of grades, with longer essays receiving higher grades (Diederich et al., 1961; Klein & Hart, 1968; Nold & Friedman, 1977). While our research cannot be directly applied to such a task, it does raise the question of whether a doubt-scaling mechanism could be involved. Specifically, if the presence of non-diagnostic information (i.e., additional, superfluous words) may be decreasing markers confidence in their decision regarding essay quality, leading them to be more reliant on cues and heuristics when assigning grades. Thus, markers may view longer essays more favorably. Clearly this speculation is tentative, but we believe it raises an interesting potential context for considering how the presence of diagnostic information may affect metacognition and reliance on cues and heuristics.

Limitations

It is important to recognize that any generalizations from this work to applied settings need to be made with appropriate caution. All of these tasks were conducted in a highly controlled paradigm, and as such have controlled many extraneous variables that are present in real-world decision scenarios (e.g., time between encoding and retrieval for eyewitness identification tasks, low prevalence effects in luggage/medical screening etc.). Further, we have only explored the effects of non-diagnostic information on recognition-ratings (and perceptual discrimination ratings) obtained in the absence of a categorical response. While we have reason to believe that such ratings would rely on similar processes to more typical, retrospective confidence judgements, we cannot definitively claim that our findings would generalize to tasks using a traditional recognition decision followed by a retrospective confidence judgement.

Conclusion

In the past, researchers have relied primarily on psychophysical models to explain the process by which confidence ratings are shaped for recognition memory tasks. The present

study provides support for one such theory: Baranski and Petrusic's doubt-scaling model of confidence (1998). Unlike other psychophysical models, the doubt-scaling model posits that as the amount of additional, non-diagnostic information present at the time a recognition judgement is made increases, a person's confidence in their decision will decrease proportionally. This understanding contributes not only to our fundamental understanding of metacognitive processes, but also has potential real-world connotations for scenarios in which non-diagnostic information while confidence is being assessed (e.g., eyewitness identification, luggage screening etc.).

References

Angell, J. R. (1907). The province of functional psychology. *Psychological Review, 14*, 61–91. doi:10.1037/h0070817

Baranski, J. V., & Petrusik, W. M. (1998). Probing the Locus of Confidence Judgments: Experiments on the Time to Determine Confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 929-945. doi: 10.1037//0096-1523.24.3.929

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1601-1608. doi: 10.1037/a0031849

Bernbach, H. A. (1971). Strength theory and confidence ratings in recall. *Psychological Review, 78*(4), 338–340. doi:10.1037/h0031034

Bradfield, A.L., Wells, G.L., & Olson, E.A. (2002). The damaging effect of confirming feedback on the relation between eyewitness certainty and identification accuracy. *Journal of Applied Psychology, 87*, 112-120. doi:10.1037/0021-9010.87.1.112

Brewer, N., Burke, A. (2002). Effects of Testimonial Inconsistencies and Eyewitness Confidence on Mock-Juror Judgments. *Law Human Behaviour 26*, 353–364 (2002). doi:10.1023/A:1015380522722

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a

lineup using confidence judgments under deadline pressure. *Psychological Science, 23*, 1208–1214. doi:10 \.1177/0956797612441217

Brewer, N., Weber, N., & Guerin, N. (2020). Police lineups of the future? *American Psychologist, 75*(1), 76–91. doi:10.1037/amp0000465

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, functional size and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11-30. doi: 10.1037/1076-898X.12.1.11

Brewer, N., & Williams, K. D. (2005). *Psychology and Law, An Empirical Perspective*. New York: Guilford Press.

Busey, T. A., Tunnicliff, J., Loftus G. R., & and Loftus, E., F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26-48. doi:10.3758/BF03210724

Carlson, C. A., Young, D. F., Weatherford, D. R., Carlson, M. A., Bednarz, J. E. & Jones, A. R. (2016). The Influence of Perpetrator Exposure Time and Weapon Presence/Timing on Eyewitness Confidence and Accuracy. *Applied Cognitive Psychology. 30*(6), 898-910. doi: 10.1002/acp.3275

Cutler, B.L., Penrod, S.D. & Dexter, H.R. (199). Juror sensitivity to eyewitness identification evidence. *Law Hum Behav* **14,** 185–191 (1990). doi:10.1007/BF01062972

Deffenbacher, K.A., & Loftus, E.F. (1982). Do jurors share a common understanding concerning eyewitness behavior? *Law and Human Behavior 6*(1), 15–30 (1982). doi:10.1007/BF01049310

Egan, J. P., Schulman, A. L., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31*(6), 768–773. doi:10.1121/1.1907783

Filliter, J. H., Glover, J. M., McMullen, P. A., Salmon, J. P., & Johnson, S. A. (2016). The DalHouses: 100 new photographs of houses with ratings of typicality, familiarity, and degree of similarity to faces. *Behavior Research Methods, 48*, 178-183. doi: 10.3758/s13428-015-0561-8

Fox J, Weisberg S (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Franks, J. J., Bilbrey, C. W., Lien, K. G., & McNamara, T. P. (2000). Transfer-appropriate processing (TAP). *Memory & Cognition, 28*(7), 1140-1151.

Gigerenzer, G., Hoffrage, U., & Kleinboelting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528. doi: 10.1037/0033-295X.98.4.506

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

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging, 5*(2), 215–227. doi:10.1037/0882-7974.5.2.215

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review, 18*(3), 186–201. doi:10.1037/h0074579

Horry, R., & Brewer, N. (2016). How target–lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145*(12), 1615–1634. doi:10.1037/xge0000227

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian Origins of Uncertainty in Judgement: A Sampling Model of Confidence in Sensory Discrimination. *Psychological Review, 104*(2), 344-366. doi: 10.1037/0033-295X.104.2.344

Juslin, P., Winman, A., & Olson, H. (2000). Naïve Empiricism and Dogmatism in

Confidence Research: Critical Examination of the Hard-Easy Effect. *Psychological Review, 107*(2), 384-396. doi: 10.1037//0033-295X.107.2.384.

Klein, S. P., & Hart, F. M., (1968). Chance and systematic factors affecting essay grades. *Journal of Educational Measurement, 5*(3), 197-206.

Kohl, T. A., Sauer, J. D., Palmer, M., Brooks, J., & Heathcote, A., (2022) *The Effects of Non-Diagnostic Information on Confidence and Decision Making*. Manuscript under review.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological review. 100*(4), 609-639. doi: 0.2307/1422236

Koriat, A. (1977). Monitoring One's Own Knowledge During Study: A Cue-Utilization Approach to Judgments of Learning. *Journal of Experimental Psychology: General, 126*(4), 349-370. doi: 10.1037/0096-3445.126.4.349

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490–517. doi:10.1037/0033-295X.103.3.490

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological Science, 25*(9), 1663-1673. doi:10.1177/0956797614541991

Macmillan, N. A., & Creelman, C. D. (1991). Detection theory: A user's guide. New York: Cambridge University Press.

Martinez, A. M., & Benavente, R. (1998). *The AR face database* (CVC Technical Report No. 24). Barcelona, Spain: Universitat Autonoma de Barcelona, Computer Vision Center.

Manley, K.D., Chan, J. C. K., & Wells, G. L. (2019). Do masked-face lineups facilitate eyewitness identification of a masked individual? *Journal of Experimental Psychology: Applied, 25*(3), 396-409. doi:10.1037/xap0000195.

Mansour, J. K. (2020). The confidence-accuracy relationship using scale versus other methods of assessing confidence. *Journal of Applied Research in Memory and Cognition, 9*(2), 215-231. doi:10.1016/j.jarmac.2020.01.003

Maurer, D., Le Grand R., & Mondloch, C. J. (2002). The many faces of configural processing. *TRENDS in Cognitive Sciences, 6*(6), 255-260. doi: 10.1016/S1364-6613(02)01903-4

Mickes, L. (2015). Receiver Operating Characteristic Analysis and Confidence-Accuracy Characteristic Analysis in Investigations of System Variables and Estimator Variables that Affect Eyewitness Memory. *Journal of Applied Research in Memory and Cognition.* doi: 10.1016/j.jarmac.2015.01.003

Morris, D. C., Bransford J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519-533. doi: 10.1016/S0022-5371(77)80016-9

National Academy of Sciences (NAS) (2014). Identifying the Culprit. Assessing eyewitness identification. Washington DC: The National Academic Press.

Neil v. Biggers, 409 U. S. 188 (1972).

Nold, E. W., & Friedman, S. W. (1977). An analysis of readers' responses to essays. *Research in the Teaching of English, 11*(2), 164-174.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*(1), 55-71. doi:10.1037/a0031602

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. Journal of Experimental Psychology: *Human Perception and Performance, 18*(4), 962-986. doi:10.1037/0096-1523.18.4.962

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgements. *Psychonomic Bulletin & Review, 10*, 177-183. doi:10.3758/BF03196482

Potter, R., & Brewer, N. (1999). Perceptions of witness behaviour-accuracy relationships held by police, lawyers and mock-jurors. *Psychiatry, Psychology and Law, 1*(1), 97-103. doi: 10.1080/13218719909524952

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin, 140*(5), 1281–1302. doi:10.1037/a0037004

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 185-208). Chichester: Wiley Blackwell.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple Confidence Estimates as Indices of Eyewitness Memory. *Journal of Experimental Psychology: General, 137*(3), 528-547. doi: 10.1037/a0012712

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347. doi:10.1007/s10979-009-9192-x

Sauer, J. D., Weber, N., & Brewer, N. (2012). Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness. *Psychonomic Bulletin & Review, 19*(3), 490-498. doi: 10.3758/s13423-012-0239-5

Sauer, J. D., Palmer, M. A., & Brewer, N. (2017). Mock-juror evaluations of traditional and ratings-based eyewitness identification evidence. *Law and Human Behavior, 41*(4), 375–384. doi:10.1037/lhb0000235

Sauerland, M., Raymaekers, L. H. C., Otgaar, H., Memon, A, Waltjen, T. T., Nivo, M … Smeets, T. (2016). Stress, stress-induced cortisol responses, and eyewitness identification performance. *Behavioral Sciences and the Law, 34*(4), 475-594. doi: 10.1002/bsl.2249

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. Journal of Mathematical Psychology, 32(2), 135-168. doi:10.1016/0022-2496(88)90043-0

Tanaka, J. W., & Gordon, I. (2011). Features, configuration, and holistic face processing. In G. Rhodes & J. Haxby (Eds.), *Oxford Handbook of Face Perception.* (pp. 177-195). Oxford University Press.

Technical Working Group: Eyewitness Evidence (1999*). Eyewitness evidence: A guide for law enforcement.* US Department of Justice, Office of Justice Programs, National Institute of Justice. NCJ 178240.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 582-600. doi: 10.1037/0278-7393.26.3.582

Vickers, D. (1970). Evidence for an Accumulator Model of Psychophysical Discrimination. *Ergonomics 13*(1), 37-58. doi:10.1080/00140137008931117

Weber, N., & Brewer, N. (2004). Confidence–Accuracy Calibration in Absolute and Relative Face Recognition Judgments. *Journal of Experimental Psychology: Applied. 10*(3), 156-172. doi:10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied, 14*(1), 50–60. doi:10.1037/1076-898X.14.1.50

Wickelgren, A. W., & Norman, D. A. (1966). Strength models and serial position in short-

term recognition memory. *Journal of Mathematical Psychology, 3*(2), 316-347. doi: 10.1016/0022-2496(66)90018-6

Windshitl, P. D., & Wells, G. L. (1996). Measuring Psychological Uncertainty: Verbal Versus Numeric Methods. *Journal of Experimental Psychology: Applied, 2*(4), 343-364. doi: 10.1037/1076-898X.2.4.343

Wixted, J.T., Wells, G.,L. (2017). The relationship between eyewitness confidence and identification accuracy: a new synthesis. Psychological Science in the *Public Interest, 18*(1), 10-65. doi:10.1177/1529100616686966

Chapter IV

The Effects of Non-Diagnostic Information on Confidence and Decision Making

The Effects of Non-Diagnostic Information on Confidence and Decision Making.

Amelia T. Kohl[1], James D. Sauer[1], Matthew A. Palmer[1], Jasmin Brooks[1], Andrew Heathcote[2]

[1] The School of Psychological Sciences

The University of Tasmania, Australia

[2] The School of Psychological Sciences

Newcastle University, Australia

This research was supported by funding from the Australian Research Council (grant DP200100655 to A. Heathcote, J. Sauer, M. Palmer et al.).

Manuscript under review at the Journal of Experimental Psychology: Learning, Memory and Cognition.

All raw data and supplementary materials for Experiments 1 and 2 are available at https://osf.io/exqba/ The experiments and analyses were not preregistered. Contact the corresponding author for a copy of the experimental and/or analyses code.

Corresponding Author:

Amelia Kohl, School of Psychological Sciences, University of Tasmania, Locked Bag 30, Hobart, Tasmania 7001, Australia

Email: Amelia.Kohl@Utas.edu.au

Abstract

Many decision making tasks are characterized by a combination of diagnostic and non-diagnostic information. Psychophysical models of responding and confidence typically focus on the contribution of diagnostic information (e.g., evidence associated with stimulus discriminability); largely ignoring the contribution of non-diagnostic information. Baranski and Petrusic's (1998) doubt-scaling model predicts a negative relationship between non-diagnostic information and confidence, and between non-diagnostic information and accuracy. In two perceptual tasks, we tested the effects of manipulating non-diagnostic information on confidence, accuracy, and reaction time (RT). In Experiment 1, participants viewed a dynamic grid consisting of flashing blue, orange and white pixels and indicated whether the stimulus was predominantly blue or orange (using a response scale ranging from low confidence blue to high confidence orange), with the white pixels constituting non-diagnostic information. Increasing non-diagnostic information reduced both confidence and accuracy, generally slowed RTs, and flattened of Conditional Accuracy Functions (indicating an increase in faster guesses). Experiment 2 replicated these results for a decision-only task, providing further support for the doubt-scaling model of confidence.

Keywords: confidence, decision making, doubt scaling, evidence accumulation.

We examine what we believe to be an empirically and theoretically interesting and important, but somewhat neglected, question: How are decision confidence, speed, and accuracy affected by non-diagnostic information (i.e., information that is not relevant to the choice and hence not determinative of accuracy)? Many theories of choice assume that only relevant evidence contributes to the decision process. This is true by construction when the evidence is unidimensional, as is assumed by the most widely applied theory of binary choice accounting for response probability, Signal Detection Theory (SDT; Bernbach, 1971; Egan et al., 1959; Green & Swets, 1966). It is also true of the most widely applied dynamic theory of binary choice, which also accounts for response time (RT), the Diffusion Decision Model (DDM; Ratcliff & McKoon, 2008). In both cases the input to the decision process is constructed from the evidence for one choice option minus the evidence for the other choice option, so the effect of any non-diagnostic information is effectively cancelled out.

However, another widely applied class of dynamic choice theories - accumulator models (Audley, 1960, Brown & Heathcote, 2008, Tillman et al., 2020, Usher et al., 2001, Vickers, 1970; Smith & Vickers, 1998: Vickers & Lee, 1998, Van Zandt et al., 2000) - have separate inputs corresponding to each response option, and could potentially be sensitive to non-diagnostic information. Baranski and Petrusic's (1998) doubt-scaling model comes from this class, and is one of the few theories quantitatively predicting accuracy and RT that attempts to directly address the effects of non-diagnostic information. Their model also addresses decision confidence, and it is with respect to confidence that there has been the most interest in the effects of non-diagnostic information. This interest is motivated by dissociations between confidence and accuracy - running counter to the generally robust positive confidence-accuracy relationship - that have been attributed to non-diagnostic information (e.g., Busey et al., 2000; Manley et al., 2019). However, empirical studies of confidence have rarely included direct quantitative manipulations of non-diagnostic

information. In this paper we study a perceptual choice task that affords such a direct manipulation and use our results to test predictions of the doubt-scaling model. Before reporting the results of two experiments, we first discuss why the effects of non-diagnostic information are of both theoretical and applied interest, describe the doubt-scaling model, and derive from it predictions about accuracy, RT and confidence.

The Doubt-Scaling Model of Confidence

The positive confidence-accuracy relationship has been influential in both theoretical and applied domains (Brewer & Wells, 2006; Gigerenzer et al., 1991; Juslin et al., 2000; Palmer et al., 2013; Sauer et al., 2010). For example, in eyewitness identification, confidence is often relied upon to assess the reliability of a decision when the correct response is unknown (Brewer & Wells, 2006; Wixted & Wells, 2017; National Academy of Sciences Report, 2014; Technical Working Group for Eyewitness Evidence, 1999). However, factors unrelated to accuracy may also shape confidence (Busey et al., 2000; Baranski & Petrusic, 1998; Van Zandt, 2000). For example, Busey et al. found that confidence in face recognition decisions increased, with no corresponding increase in accuracy, when the luminance of an image increased from study to test. Accuracy, however, was improved when luminance at encoding and test matched. Hence, non-diagnostic information can inflate confidence.

Although the doubt-scaling model is, to our knowledge, the only theory to explicitly account for the effect of non-diagnostic information on accuracy, RT and confidence, it has not, to our knowledge, been directly tested. Understanding the role of non-diagnostic information on confidence, accuracy and RT could be of great value in applied settings, where decision stimuli often contain non-diagnostic information. For example, consider an eyewitness identification test. During the commission of a crime, parts of a perpetrator's face may be concealed from or unobserved by an eyewitness. However, when the witness later views a lineup, the faces of the lineup members may be presented unobstructed. Thus, each

face will contain featural and configural information that is non-diagnostic because that information was not encoded during the initial event and therefore cannot contribute to genuine recognition. Manley et al. (2019) conducted a face recognition task involving a combination of full and partial faces (faces where only the eye area was visible, as might be the case if the perpetrator was wearing a ski-mask). Participants' confidence in their recognition decisions was lower for trials in which they studied a partial face but were tested with a full face, suggesting that decision confidence was reduced by the additional, non-diagnostic information present at test (i.e., parts of the face obscured at study but visible at test). Given the recent international surge of mask-wearing for health reasons, it is important to understand how non-diagnostic information affects recognition, and the confidence and RT associated with recognition. Non-diagnostic information may also affect applied perceptual discrimination tasks. For example, when border security agents compare passport images to real faces, some features are relatively stable and therefore likely to be diagnostic (e.g., shape of the face, distance between eyes), while others are easily changeable and therefore may prove non-diagnostic (e.g., colour/length of hair, lighting). Understanding how non-diagnostic information affects decision making processes may have substantial applied value.

The doubt-scaling model of confidence (Baranski & Petrusic, 1998) evolved from slow and fast guessing theory (Petrusic, 1992). As an extension of the "runs" model of binary choice (Audley, 1960), it assumes that on each time step evidence is dichotomized as either A>B (favouring choice A) or B>A (favouring choice B). This discrete evidence is tallied in corresponding accumulators until a response threshold is reached, triggering a decision. Slow and fast guessing theory suggests a third accumulation process: A=B (i.e., non-diagnostic evidence favouring neither choice option). If the A=B accumulator reaches its threshold first a guess response is triggered, reducing accuracy. The doubt-scaling model expands upon this account by making explicit predictions regarding the relationship between non-diagnostic

information and confidence; specifically, that confidence is inversely proportionate to the amount of information accumulated for A=B. Thus, the more non-diagnostic information accumulated, the less confident the responder will be.

The doubt-scaling model also makes two predictions regarding RT. First, the presence of non-diagnostic information slows overall RT. Accumulation in the runs model is competitive: if one type of evidence is tallied on a given time step the other evidence totals remain unchanged. Hence, when evidence increments are shared among accumulators the time for any one accumulator to reach threshold is slowed. A more fine-grained prediction is made with respect to the relationship between RT and accuracy. The speed and frequency of guessing responses increases as the rate of A=B increments increases.

We cannot observe guessing and non-guessing responses separately, but we can compare the speed of less accurate responses (which should include more guessing responses) and more accurate responses (which should include more non-guessing responses) using Conditional Accuracy Functions (CAFs; Thomas, 1974). CAFs allow us to plot accuracy as a function of RT, with RT first being ordered and divided into equal sized "bins". We do not fit the doubt-scaling model, and so cannot test any quantitative predictions with respect to the shape of the CAF (which is likely to be non-linear, e.g., Elliot et al, 2021). However, the model clearly predicts that as accuracy generally decreases (as a result of non-diagnostic information increasing), the accuracy of the faster bins will decrease relative to the accuracy of the slower bins, resulting in a flattening of the CAF curve.

Experiment 1

The choice stimuli in Experiment 1 were grids containing a combination of blue, orange, and white pixels whose arrangement changed dynamically. Participants provided a decision, and confidence rating for their decision, about whether there were more orange or

blue pixels, where each colour was more common equally often over trial. Hence, the white pixels provided non-diagnostic information.

We introduced two manipulations that test the generality of the doubt-scaling model's predictions, one of which provides a further test of the model. First, the more common colour constituted 55% of the diagnostic pixels in an easy-choice condition and 52% in a hard-choice condition. This difficulty factor was crossed with two ways of manipulating the amount of non-diagnostic information. In all cases there were four levels of non-diagnostic information, a control condition with no white pixels, and low, moderate and high non-diagnosticity conditions, where white pixels constituted, respectively, one sixth, one third or half of the total number of pixels. In the "additive" manipulation the number of coloured (i.e., diagnostic) pixels was kept constant as the number of white pixels increased, leading to an increase in total grid size. In the "stable" condition, grid size remained constant, and the total number of diagnostic pixels decreased while maintaining the blue-to-orange proportion. If the absolute amount of diagnostic or non-diagnostic information is important, the results for these two manipulations could differ. In contrast, the runs model assumes that all that matters is the relative amounts of the three different types of information, so the doubt-scaling model predicts no difference between the additive and stable conditions.

In summary, we expected that accuracy and confidence would be less, and RT slower, for the hard condition than the easy condition. The doubt-scaling model predicts that as non-diagnostic information increases, in both the hard and easy conditions, confidence and RT will decrease, that the accuracy of faster responses will increase relative to slower responses, and that the additive and stable conditions will not differ. If the absolute amount of non-diagnostic information is important, we would expect our manipulation of non-diagnostic information to have a larger effect in the additive condition. Alternatively, if the absolute

amount of diagnostic information is more influential, we would expect a larger effect in the stable condition.

Method

Open Science Practices

All raw data for both Experiments 1 and 2 is available at https://osf.io/exqba/ Contact the corresponding author for a copy of the experimental and/or analyses code.

Design

We used a 2 (grid-type: stable or additive) × 4 (proportion of non-diagnostic information: 0, 1/6, 1/3, ½) × 2 (difficulty: easy vs. hard) × 2 (majority colour: blue vs. orange) mixed design, with grid-type as the between-subjects factor. Dependent variables are decision confidence (low, moderate, and high), mean RT, and accuracy; measured by the equal-variance signal-detection theory discrimination (*d*') measure, and proportion correct as a function of RT (as used in the CAFs).

Participants

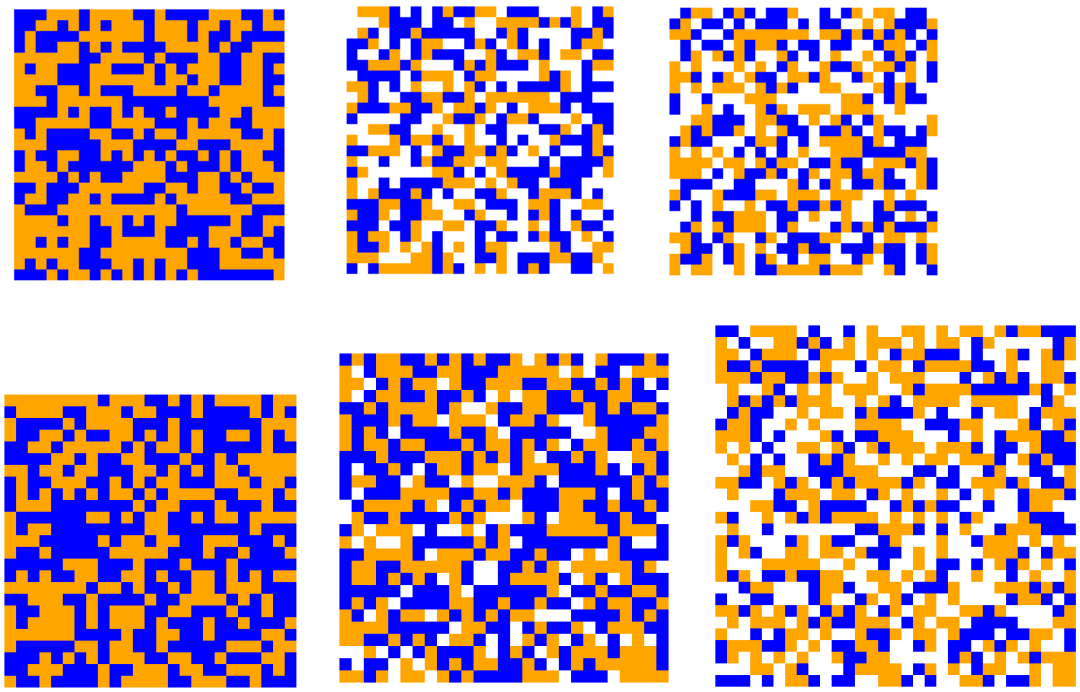
We randomly allocated 56 participants to the stable or additive condition, as to allow for a minimum of 20 participants per cell (see Simmons et al., 2011).Eight participants were excluded from analyses as their data showed truncated reaction time distributions due to the 5 sec response window (n= 5), or below 55% accuracy on 'easy' trials (n=3). An additional participant was excluded for incomplete data. This left 22 participants in the stable condition and 25 participants in the additive condition. First year psychology students were reimbursed with research credits and other participants received a $20 e-voucher. Participants were required to have normal or corrected-to-corrected normal vision, and were not eligible to participate if they suffered from epilepsy or related conditions.

Materials

Participants completed the task on in-lab desktop computers equipped with 3.30 GHz Intel i5-6600 processors, 16 GB RAM, and a Windows 7 enterprise operating system configured to minimize internal task-switching. The program was written and run using MATLAB (The MathWorks, R2016b). For each trial, participants viewed a dynamic grid consisting of blue (RGB = 0, 65, 255), orange (RGB = 255, 127, 0) and sometimes white (RGB = 128, 128, 128) pixels (see *Figure 1*). Although the colour of pixels in the grid changed constantly, the proportion of blue, orange, and white pixels remained constant. Table 1 provides a breakdown of how the coloured pixels varied between different levels of non-diagnostic information for each grid type.

Figure 1

Schematic representations of the dynamic-grid stimulus at varying levels of non-diagnostic information



Note: White pixels represent non-diagnostic information, and the total proportion of non-diagnostic information increases from left to right. The first row represents the stable condition, whereas the second row represents the additive condition.

Participants responded by moving their mouse from the start point (a circle on the screen, equidistant from six response options) to the relevant segment of the response arc (labelled low confidence blue, moderate confidence blue, high confidence blue, low confidence orange, moderate confidence orange and high confidence orange). This design allows reaction time data for multiple levels of confidence to be collected in a way that minimizes noise associated with motor responses (e.g., differences in motor time associated with the use of different fingers to indicate confidence using a keyboard). Participants clicked inside a circle in the middle of screen to begin each trial. Participants who responded too quickly (before .15 sec) were warned that they were too fast. Participants who took longer than 2 seconds to respond were warned that they were responding too slowly. If participants did not respond within 5 seconds, the trial ended, and they received an on-screen message saying they were too slow to respond.

In the stable condition, the overall grid size remained constant (see Table 1). Thus, as the overall proportion of non-diagnostic pixels (i.e., white pixels) increased, the number of diagnostic pixels (i.e., blue and orange pixels) decreased. In the additive condition, the dynamic grid increased in size as the proportion of non-diagnostic pixels increased, meaning the number of diagnostic pixels remained constant (529 diagnostic/coloured pixels). In easy trials, 55% of the diagnostic pixels consisted of the dominant color (i.e., correct response). In hard trials, 52% of the diagnostic pixels consisted of the dominant color (i.e., correct response). The remainder of the grid was filled with the incorrect color, and non-diagnostic pixels.

Procedure

Participants first completed three training blocks, with the first two blocks comprising of 20 trials and the third block of 40. In all three practice blocks, participants were provided with feedback indicating whether each response was correct. In the first practice block,

participants responded to a stimulus like that in Figure 1 (i.e., pixels where were either orange or blue; no white was included), by simply indicating whether the stimulus was predominantly orange or blue, with no confidence ratings required. The second practice block introduced the manipulation of non-diagnostic information (i.e., white pixels), and the third practice block introduced the 6 response categories (high confidence blue, moderate confidence blue, etc.). This approach was intended to help participants learn the demands of the task before starting experimental trials. Each experimental block comprised of 80 trials. Participants completed nine experimental blocks). Unlike the practice blocks, participants did not receive feedback for the experimental trials. Participants were encouraged to take rests between blocks as required. The task took approximately one hour to complete.

Analysis Methods

Few participants consistently used all three levels of confidence, with participants varying in the least-used level. To produce stable estimates for our analyses, we collapsed responses to two levels of confidence ("low" and "high"). For each participant, moderate responses were collapsed into either the high or low category based on upon which of these two options was used less frequently.

We used linear mixed-effect models assuming Gaussian error to analyze the logarithm of RT and generalized linear mixed-effect model with a probit link function to analyze the probability of high-confidence responses (Bates et al., 2015; Kuznetsova, 2017). Participant was set as a random factor, with grid-type (additive/stable), proportion of nondiagnostic information (0, 1/6, 1/3, ½), difficulty (hard/easy), and predominant stimulus color (orange or blue) included as fixed effects.

Due to response bias in the data—participants showed a bias towards orange stimuli at low proportions of non-diagnostic information, and a bias towards blue stimuli at high levels of non-diagnostic information–we analyzed accuracy using the SDT-based measure of

discrimination, *d´*, rather than raw accuracy scores. This allowed us to determine the effect of non-diagnostic information on participants' ability to discriminate between the correct and incorrect response independent of response bias. The discrimination analysis was accomplished using generalized linear mixed effect model with a probit link function on the proportion of blue responses, with *d´* corresponding to the difference between majority blue vs. majority orange stimuli, and effects on *d´* corresponding to interactions with the stimulus factor.

We constructed CAFs by dividing responses into quintile bins (separately for the stable and additive conditions, and then collapsed across both). For the first bin the accuracy of the fastest 20% of responses is plotted, for the second the accuracy for RTs between the 20th and 40th percentiles and so on up to the slowest 20% of responses for the 5th bin. The choice of number of bins is arbitrary; this relatively coarse division results in precise accuracy estimates for each bin as they are based on many responses. The same pattern of results, albeit with more variability, was found using more bins.

Results

Statements about significance are made with respect to a .05 criterion. Although we included the majority-colour factor in the RT and confidence ANOVAs we do not report tests of it as they are not germane to our hypotheses. Full ANOVA tables for confidence, accuracy and RT are provided in supplementary materials and important effects summarized below.

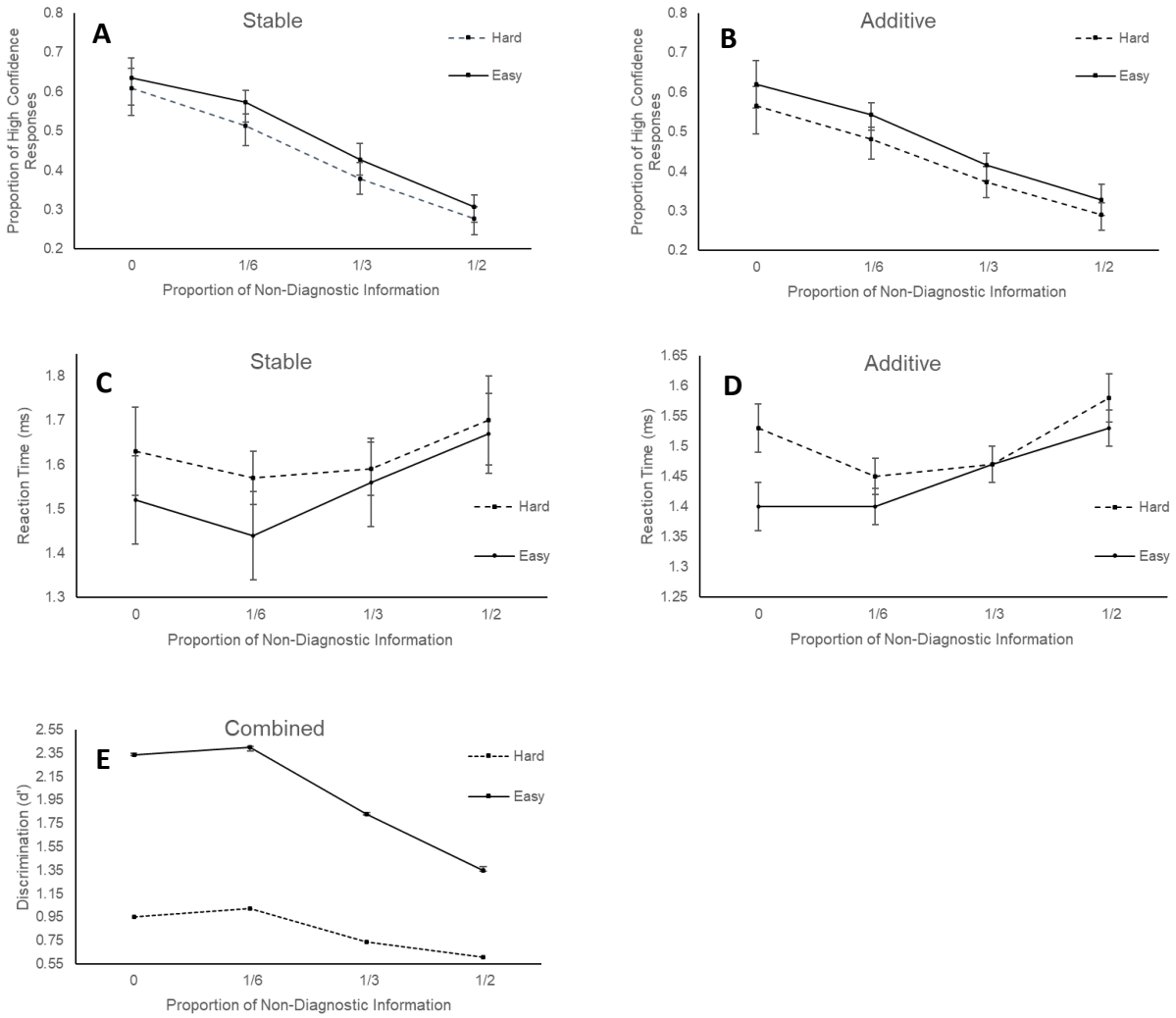
Confidence

Consistent with the doubt-scaling model, the proportion of high-confidence responses decreased as the amount of non-diagnostic information increased, $\chi^2(3) = 2102.32$, $p < .001$ (see Figure 2, Panels A and B). There was also a significant main effect of difficulty on confidence, with easy trials receiving a higher proportion of high-confidence responses than hard trials, $\chi^2(1) = 95.64$, $p < .001$, with no significant interaction between the two effects

(see Table 2). Although the main effect of grid type, and interactions with difficulty, were non-significant, grid type did interact with the proportion of non-diagnostic information, $\chi^2(1) = 24.25$, $p < .001$. However, the interaction effect was only small: non-diagnostic information exerted a slightly greater effect on confidence in the stable condition (mean confidence decreased from 62% at zero non-diagnostic information to 29% at half non-diagnostic information) than the additive condition (59% vs. 31%), collapsing across difficulty conditions.

Figure 2

The Effects of Non-Diagnostic Information on Confidence, Accuracy (*d*') and RT for Experiment 1



Note: Figures demonstrating the relationship between non-diagnostic information and confidence (panels A and B), RT (panels C and D), and accuracy (as indexed by d', panel E). Error bars represent the standard error.

Discrimination

Increasing the proportion of non-diagnostic information also affected discrimination, $\chi^2(3) = 322$, $p < .001$. As shown in Figure 2 (Panel E), discrimination was nearly identical at the two lowest levels, but decreased systematically thereafter. Difficulty had the expected strong main effect, $\chi^2(1) = 1251$, $p < .001$, and the amount of non-diagnostic information had a weaker effect for hard than easy trials, $\chi^2(1) = 66.8$, $p < .001$. The only effect including

grid type was a relatively weak interaction where the difficulty effect was larger in the stable than the additive condition, $\chi^2(1) = 4.83$, $p = .03$.

Reaction Times

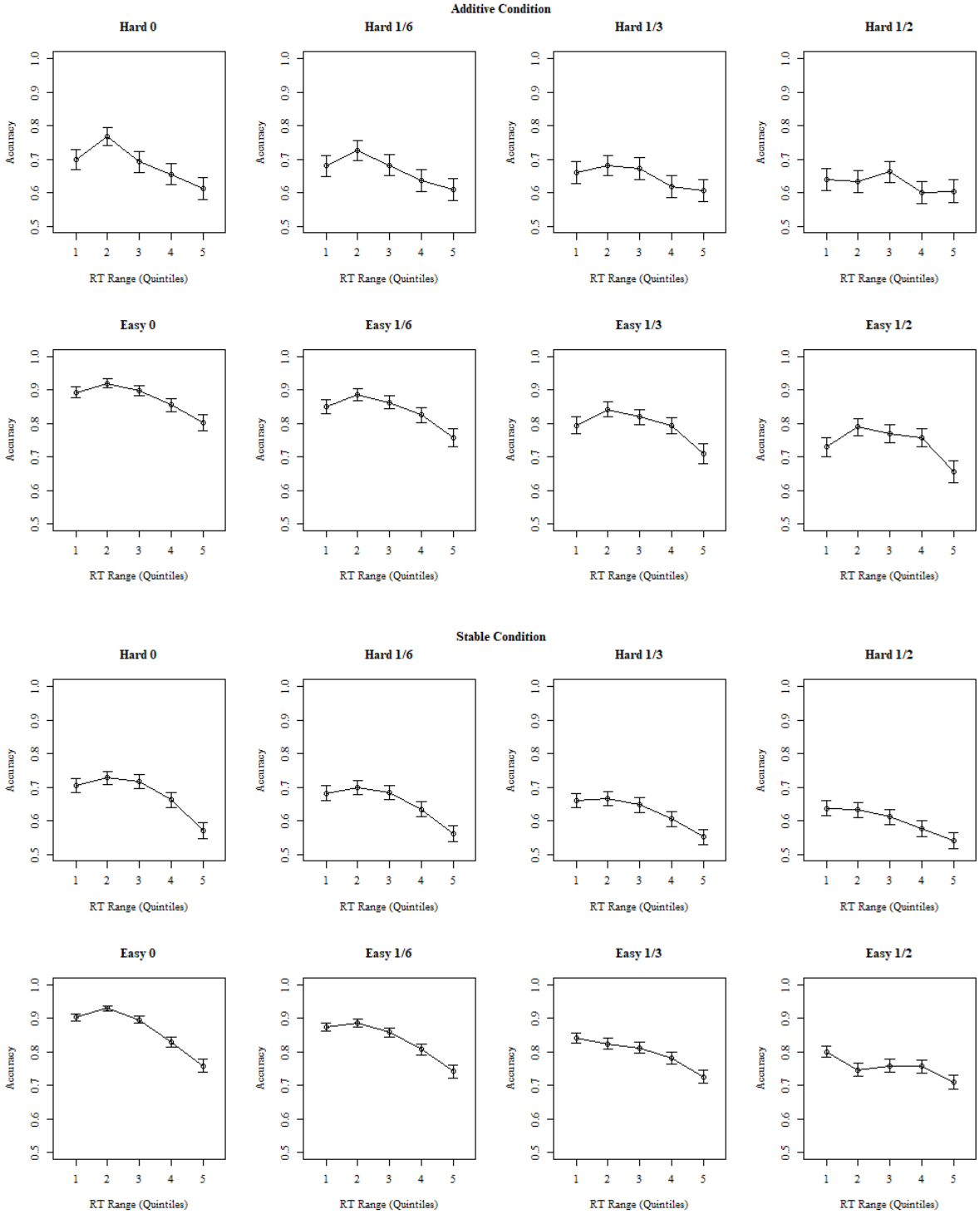
RT generally slowed as the proportion of non-diagnostic information increased, $F(3,33599) = 109.7$, $p < .001$ and was faster for easy than hard choices, $F(3,33599) = 218.2$, $p < .001$. However, as shown in Figure 2 (Panels C and D), the slowing was restricted to conditions where there was some non-diagnostic information, and the difficulty effect diminished as non-diagnostic information increased. The only significant effect of grid type was a significant interaction with both difficulty and non-diagnostic information, $F(3,33599) = 2.95$, $p = .03$, due to a larger difficulty effect for stable than additive for the low proportion of non-diagnostic information.

Conditional Accuracy Functions

Figure 3 shows a pattern consistent with the predictions of the doubt-scaling model: the overall level of the CAFs decreased as non-diagnostic information increased, and accuracy for slower bins increased relative to accuracy for faster bins, although the relative change was less marked for the hard additive condition.

Figure 3

Conditional Accuracy Functions for Experiment 1



Note: Conditional-accuracy functions of accuracy as a function of RT for hard and easy trials in both the additive and stable conditions. The first point on the x-axis represents the

fastest responses (below the 20th percentile of RT), the second point the 20th to 40th percentile etc. Error bars represent the standard error.

Discussion

Consistent with predictions based on the doubt-scaling model, the proportion of high confidence responses—and participants' ability to discriminate between correct and incorrect responses—decreased as non-diagnostic information increased. RT generally decreased as non-diagnostic information increased, and CAFs showed that the accuracy of fast responses generally decreased relative to that of slow responses. The grid type manipulation generally had little impact, with the exception of a larger difficulty effect on RT for stable than additive trials for the 1/6 proportion of non-diagnostic information, and a stronger effect of the proportion of non-diagnostic information on confidence in the stable condition than the additive condition. The latter result appears to indicate an effect of the absolute amount of diagnostic information in addition to an effect of non-diagnostic information. Although the effect is only small (a reduction of 33% in the stable condition compared to a reduction of 28% in the additive condition), it suggests a slight deviation from the predictions of the doubt-scaling model. We note that this inconsistency was not evident in either accuracy or RT (aside from the three-way interaction with difficulty and non-diagnostic information).

Experiment 2

Having participants consider confidence while making decisions slows the decision-making process (Baranski & Petrusic, 2001; 2003). To test whether the observed effects of non-diagnostic information on accuracy and RT generalize, Experiment 2 removed the confidence ratings. Our hypotheses with respect to accuracy, RT, and their combination in CAFs, remain the same.

Method

Design

We used a 4 (proportion of non-diagnostic information: 0, 1/6, 1/3, ½) × 2 (difficulty: easy or hard) within-subjects design. The outcome variables were accuracy (*d*') and RT (as indexed by CAFs and mean RT).

Participants

Twenty-one participants completed the experiment, with one being removed for truncated reaction times. Renumeration and exclusion criteria were the same as Experiment 1.

Procedure

The procedure and materials were identical to those of Experiment 1, with three exceptions. First, there were only two response options ("Orange" and "Blue"). Second, as grid-type did not moderate the effects of non-diagnostic information on accuracy or RT, we used only the stable manipulation. Third, participants completed only two practice blocks before beginning the experimental trials (20 trials without the presence of non-diagnostic information, 40 trials with the non-diagnostic manipulation), as confidence ratings were no longer relevant.

Analysis Methods

Data were analyzed using the same approach as Experiment 1.

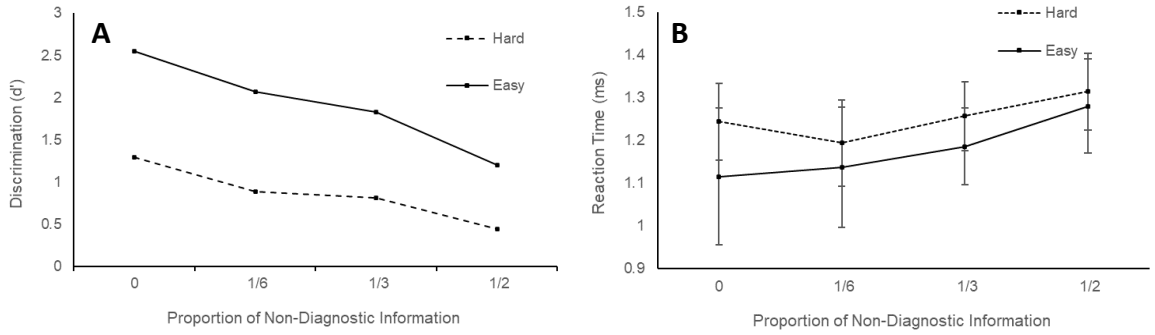
Results

Discrimination

As expected, *d´* decreased significantly as non-diagnostic information increased, $\chi^2(3) = 244$, $p < .001$. Again, there was a main effect of difficulty, with participants showing better discrimination for easy than hard trials, $\chi^2(1) = 448$, $p < .001$, and the difficulty effect reduced slightly with increased non-diagnostic information, $\chi^2(1) = 17$, $p < .001$ (see Figure 4, Panel A). Figure 4 also shows that, in contrast to Experiment 1, discrimination was highest in the control (no non-diagnostic information) condition.

Figure 4

The Effects of Non-Diagnostic Information on Accuracy (d') and RT for Experiment 2



Note: Figures demonstrating the relationship between non-diagnostic information and accuracy (as indexed by d', panel A) and RT (panel B). Error bars represent the standard error.

Reaction Times

As per Experiment 1, a linear mixed effects model on log RTs showed that RT generally increased as the proportion of non-diagnostic information increased, $F(3,14347) = 90.5$, $p < .001$. There was the expected overall slowing for difficult choices, $F(1,14347) = 135.6$, $p < .001$, which interacted with proportion of non-diagnostic information, $F(3,14347) = 7.7$, $p < .001$. The interaction was due to a weakening of the difficulty effect as non-diagnostic information increased, and the slowing in the control condition relative to the low non-diagnostic information condition seen in Experiment 1 was weakened (see Figure 4, Panel B).

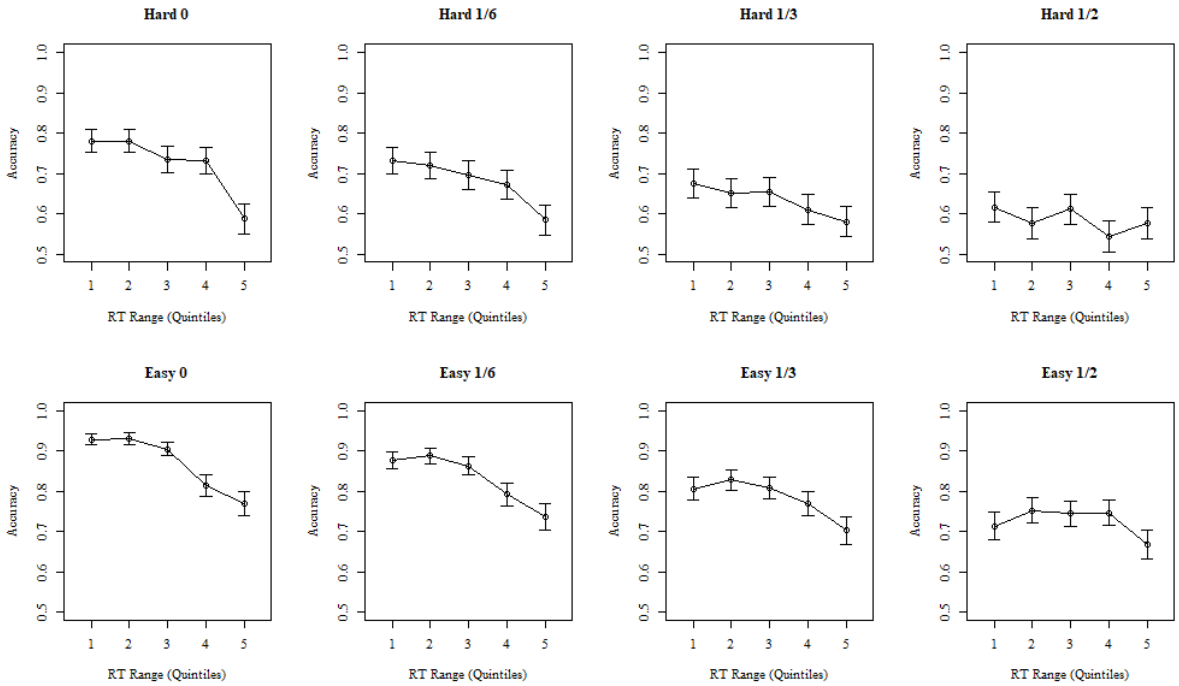
Conditional Accuracy Functions

Like Experiment 1, CAFs generally decreased when there was no non-diagnostic information and flattened as the amount of non-diagnostic information increases. Correspondingly, the interaction between RT range as factor and proportion of non-diagnostic information was significant, $\chi^2(1) = 41.9$, $p < .001$. In this case, the three-way

interaction with difficulty was also significant, $\chi^2(1) = 15.3$, $p = .004$; reflecting stronger flattening in the easy condition (see Figure 5).

Figure 5

Conditional Accuracy Functions for Experiment 2



Note: Conditional-accuracy functions of accuracy as a function of RT for hard and easy trials in both the additive and stable conditions. The first point on the x-axis represents the 20th percentile of RT, the second point the 20th to 40th percentile etc. Error bars represent the standard error.

Discussion

The results of Experiment 2 support the generality of the predictions made by the doubt-scaling model with respect to accuracy and RT. First, discrimination decreased significantly as non-diagnostic information increased, while RT increased as non-diagnostic information increased. The effect on these measures was more monotonic than in Experiment 1, although the control condition was still slower than the low non-diagnostic condition for

hard choices. Second, as non-diagnostic information increased, CAFs flattened, suggesting that participants' errors sped up as the proportion of non-diagnostic information increased.

General Discussion

Baranski and Petrusic's (1998) doubt-scaling model provides a framework for understanding the effect of non-diagnostic information (i.e., information that is not relevant to making a correct decision) on choice tasks. Although the model was proposed in the literature in 1998, it has not, to our knowledge, been directly tested in an experimental setting. Experiment 1 confirmed the model's prediction that confidence decreases monotonically as non-diagnostic information increases, and that this holds for both harder and easier decisions. Both our experiments confirmed the prediction that accuracy decreases, and RT increases, as non-diagnostic information increased. However, results deviated from the prediction that the control condition should show the highest accuracy and lowest RT in Experiment 1 (where confidence judgements were required), and regarding RT for hard decisions in Experiment 2 (without confidence).

We also tested two more fine-grained predictions of the doubt-scaling model. First, we predicted that as non-diagnostic information increased so would the accuracy of slow responses relative to the accuracy of fast responses. In both experiments we confirmed this prediction, consistent with the idea that the presence of more non-diagnostic information causes faster and more frequent guessing responses. The second prediction concerns the effect of non-diagnostic information being determined by only the proportion rather than the absolute amounts of diagnostic and non-diagnostic information. This prediction held for accuracy and RT, but not for the confidence judgements in Experiment 1, where an increase in the proportion of non-diagnostic information had a stronger effect when it was accompanied by a decrease in the absolute amount of diagnostic information.

Apart from these two relatively minor deviations from predictions, our results provide clear support for the doubt-scaling model, at least in the context of the perceptual decisions studied in the present experiments. We now discuss the applied and theoretical implications of our results. Although the present results are generally consistent with the predictions of Baranski and Petrusic's (1998) doubt-scaling model, those predictions are not necessarily unique to that model, and the deviations from its predictions suggest some avenues for further theoretical investigation. First, the effect of the absolute amount of diagnostic information is consistent with the small but reliable effects of absolute stimulus magnitude found in paradigms ranging from simple perceptual choices (Teodorescu at al., 2016) to value judgements (Miletić et al., 2021). Modern descendants of Audley's (1960) runs model, such as McClelleand and Usher's (2001) Leaky Competitive Accumulator model and van Ravenzwaaij et al.'s (2020) Advantage Linear Ballistic Accumulator model, produce small magnitude effects as the relative amount of information for each choice increases. In future work, it would be interesting to incorporate these models into the doubt-scaling framework.

A second potential theoretical extension could be considered with respect to the doubt-scaling model's guessing mechanism. Hawkins and Heathcote (2021) used a guessing process to provide a broad and integrative account of the effect of the passage of time on decisions. Like the doubt scaling model, their Timed Racing Diffusion Model (TRDM) produces guesses when a third accumulator beats both accumulators that accrue diagnostic information about a binary choice. In their case, the guessing accumulator is driven by a constant input and so provides a measure of the passage of time. It would be interesting to investigate whether the rate of this accumulator is modulated by the presence of non-diagnostic information, further broadening the explanatory reach of the TRDM.

Our work also raises questions regarding the mechanisms that lead to higher accuracy rates for recognition tasks when lineup members are presented simultaneously vs.

sequentially. It has been suggested that simultaneous presentation allows participants to discount features that are shared by the lineup members and hence non-diagnostic (Wixted & Mickes, 2014; Wixted et al., 2018). Although discounting of common features seems likely to occur at least to some degree with faces and other complex visual stimuli (see also Heathcote et al, 2009; Tulving, 1981), our findings suggest that the presence of non-diagnostic information could have a residual effect in terms of decreased confidence and accuracy. Hence, further work examining the role of non-diagnostic information on decision making using a complex recognition task (cf. a basic perceptual task) is needed to clarify the generalisability of our results.

Finally, although addressing only simple perceptual decisions, our findings have potential implications in real-world contexts. For example, if an eyewitness saw a masked criminal commit a crime, they may be less confident identifying the perpetrator from a lineup if presented with their un-masked face (e.g., Manley et al., 2019). The current research provides some initial insight into the effects of non-diagnostic information on confidence and accuracy that may be utilized in applied scenarios to better understand and evaluate decision making. It would be useful, however, to extend our investigation to the sorts of complex perceptual decisions, and recognition memory decisions, relevant to these applied contexts.

References

Audley, R. J. (1960). A stochastic model for individual choice behaviour. *Psychological Review, 67*(1), 1-15. doi:10.1037/h0046438

Baranski, J. V., & Petrusik, W. M. (1998). Probing the Locus of Confidence Judgments: Experiments on the Time to Determine Confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 929-945. doi:10.1037//0096-1523.24.3.929

Baranski, J. V., & Petrusik, W. M. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology/Revue Canadienne de psychologie experimentale, 55(*3), 195-206. doi:10.1037/h0087366

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

Bernbach, H. A. (1971). Strength theory and confidence ratings in recall. *Psychological Review, 78*(4), 338–340. doi:10.1037/h0031034

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, functional size and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11-30. doi:10.1037/1076-898X.12.1.11

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153–178. doi:10.1016/j.cogpsych.2007.12.002

Busey, T. A., Tunnicliff, J., Loftus G. R., & and Loftus, E., F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26-48. doi:10.3758/BF03210724

Egan, J. P., Schulman, A. L., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31*(6), 768–773. doi:10.1121/1.1907783

Elliott, D., Strickland, L., Loft, S. & Heathcote, A. (accepted 10/November/2021). Integrated responding improves prospective memory accuracy. *Psychonomic Bulletin & Review*.

Festinger, L. (1943). Studies in decision: I. Decision-time, relative frequency of judgement and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32*(4), 291-306. doi:10.1037/h0056685

Gigerenzer, G., Hoffrage, U., & Kleinboelting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528. doi:10.1037/0033-295X.98.4.506

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

Hawkins, G.E., & Heathcote, A. (2021). Racing against the clock: Evidence-based vs. time-based decisions. *Psychological Review, 128*, 222-263.

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

Juslin, P., Winman, A., & Olson, H. (2000). Naïve Empiricism and Dogmatism in Confidence Research: Critical Examination of the Hard-Easy Effect. *Psychological Review, 107*(2), 384-396. doi:10.1037//0033-295X.107.2.384.

Kuznetsova A., Brockhoff P.B., Christensen R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1-26, doi:10.18637/jss.v082.i13

Manley, K.D., Chan, J. C. K., & Wells, G. L. (2019). Do masked-face lineups facilitate

eyewitness identification of a masked individual? *Journal of Experimental Psychology: Applied, 25*(3), 396-409. doi:10.1037/xap0000195.

Miletić, S., Boag, R.J., Trutti, A. C., Stevenson, N., Forstmann, B.U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks, *eLife*. doi:10.7554/eLife.63055

National Academy of Sciences (NAS) (2014). Identifying the Culprit. Assessing eyewitness identification. Washington DC: The National Academic Press.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*(1), 55-71. doi:10.1037/a0031602

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. Journal of Experimental Psychology: *Human Perception and Performance, 18*(4), 962-986. doi:10.1037/0096-1523.18.4.962

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgements. *Psychonomic Bulletin & Review, 10*, 177-183. doi:10.3758/BF03196482

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922. doi:10.1162/neco.2008.12-06-420

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347. doi:10.1007/s10979-009-9192-x

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination.

Journal of Mathematical Psychology, 32(2), 135-168. doi:10.1016/0022-2496(88)90043-0

Teodorescu, A. R., Moran, R., & Usher, M. (2016). Absolutely relative or relatively absolute: violations of value invariance in human decision making. *Psychonomic Bulletin & Review, 23*(1), 22–38. doi:10.3758/s13423-015-0858-8

Tillman, G., Zandt, T. V., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review, 27*(5), 911–936. doi:10.3758/s13423-020-01719-6

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

Usher, M., & McClelland, L. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 109*(3), 550-592. doi:10.1037/0033-295X.108.3.550

Van Ravenzwaaij, D., Brown, S. D., Marley, A. J., & Heathcote, A. (2020). Accumulating advantages: A new conceptualization of rapid multiple choice. *Psychological Review, 127*, 186–215. doi:10.1037/rev0000166

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 582-600. doi:10.1037/0278-7393.26.3.582

Van Zandt, T. V., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review, 7*, 208–256. doi: 10.3758/BF03212980

Vickers, D. (1970). Evidence for an Accumulator Model of Psychophysical Discrimination. *Ergonomics 13*(1), 37-58. doi:10.1080/00140137008931117

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgements. I. properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences, 2*(3), 169-194. doi:10.1023/A:1022371901259

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. Acta Psychologica, 50(2), 179-197. doi:10.1016/0001-6918(82)90006-3

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*, 262–276. doi:10.1037/a0035940

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81-114. doi:10.1016/j.cogpsych.2018.06.001

Wixted, J.T., Wells, G.,L. (2017). The relationship between eyewitness confidence and identification accuracy: a new synthesis. Psychological Science in the *Public Interest, 18*(1), 10-65. doi:10.1177/1529100616686966

Chapter V

A Rating-Only Approach to Perceptual Discrimination

A Rating-Only Approach to Perceptual Discrimination

Amelia T. Kohl[1], Talira M. Kucina[1], James D. Sauer[1], Matthew A. Palmer[1]

[1] The School of Psychological Sciences

The University of Tasmania, Australia

Manuscript in preparation for submission to the Journal of Applied Research in Memory and Cognition.

This research was supported by funding from the Australian Research Council (grant DP200100655 to A. Heathcote, J. Sauer, M. Palmer et al.).

Corresponding Author:

Amelia Kohl, School of Psychological Sciences, University of Tasmania, Locked Bag 30, Hobart, Tasmania 7001, Australia

Email: Amelia.Kohl@Utas.edu.au

Abstract

In the recognition memory literature, confidence ratings have been used in the absence of a binary decision to discriminate previously studied from non-studied stimuli. Further, it has been suggested that such an approach may reduce the effects of biases that operate on criterion, therefore reducing a common source of decision error. In two experiments, we explored whether a rating-only approach to confidence (i.e., a confidence rating provided in the absence of a decision judgment) could be used in a perceptual task to effectively discriminate (1) whether a dynamic grid consisted of predominantly blue or predominantly orange pixels, and (2) target present from target absent trials in a visual search task. In Experiment 1 (N = 24), we confirmed that a rating-only approach could be used in a basic perceptual paradigm to indicate the likelihood of a dynamic grid consisting predominantly of blue or orange pixels. We found that participants were well calibrated in their responding, indicating that they were able to use the rating-only scale to convey the likelihood that the stimuli was predominantly blue or predominantly orange. Experiment 2 (N = 99) tested whether this could be extended to an applied visual search task: Searching luggage x-rays for a weapon. Specifically, we tested whether confidence ratings discriminated trials containing a weapon from those that did not, and whether using ratings could mitigate the low prevalence effect (where shifts in decision criterion mean rarely occurring targets often go undetected) by externally applying criteria to determine which responses were classified as target present (and absent). While we did not find strong evidence of low prevalence effects (and therefore could not comment on the utility of a rating-only approach to reduce them), we established how categorising the rating-only responses as either target present or absent impacts performance outcomes (and how this compared to outcomes when traditional binary decision judgments were used). Supporting our key aim, participants demonstrated they could use the rating-only scale to discriminate between target present and target absent trials.

Keywords: confidence, decision making, visual search, perceptual discrimination.

The theoretical frameworks typically used to explain decisions made in important real-world contexts commonly posit that a decision is based on some comparison of (a) evidence accumulated in favour of a particular response against (b) a threshold determining when the evidence is sufficient to warrant a particular response. The implications of such models are two-fold. First, errors may arise from noise in the accumulation process (reflecting noise in the stimuli or noise in the processing system), or from effects on response bias that make a certain response more (or less) likely, independent of the amount of evidence accumulated/stimulus discriminability. To use x-ray luggage screening as an example; following a security breech involving an aircraft, security screeners might have an increased awareness of the consequences associated with missing a weapon concealed in a bag. That is, the perceived level of threat would be heightened following such an event, and this expectation alone can impact criterion placement and, subsequently, decision making, irrespective of the objective evidence available in support of one response option over another. In this example, it would lead to a more liberal criterion placement, whereby the screener would respond positively (i.e., that there is a weapon in the bag) more often. This shift in threshold, associated with a bias against negative responding (i.e., that there is no weapon present), reduces the likelihood of misses (i.e., failing to identify a weapon in a bag), but increases the risk of false alarms (identifying a weapon when there is no weapon present).

Given the importance of accurate decision making in many real-word environments, mitigating the effects of biases that contribute to decision errors is critical. One way to attenuate these errors is to minimise adverse influences on response criteria by altering the way people make responses. For example, in recognition memory tasks, there is some indication that errors can be curtailed by avoiding categorical responding in favour of responses made on a continuous confidence scale. Ratings-based approaches to recognition can identify target stimuli among lures even without a categorical response. Researchers have

argued that perhaps this suggests that by avoiding the categorical response and simply asking for ratings, we might mitigate the effects of biasing factors on criterion placement, and the associated errors (e.g., Brewer et al., 2012; 2020; Cleary & Greene, 2000; Sauer et al., 2008). Similarly, this approach might reduce errors associated with shifts in response bias in applied perceptual discrimination/visual search tasks. As alluded to, response criteria can be affected by the perceived likelihood of a target being present, with criteria becoming more lenient as the perceived likelihood of a target increases. Further, the perceived cost of different errors can affect criterion placement, in some cases, placement will become more liberal if the cost of a miss error is high, for example, failing to detect a threat in the luggage screen scenario. Conversely, more conservative responding may arise where the outcome of a false alarm is more costly.

In this research, we explored whether participants would be able to use a rating-only response scale (i.e., asking for a confidence rating in the absence of a binary decision) to discriminate between predominantly blue or orange stimuli in a basic perceptual discrimination task (Experiment 1), and to discriminate target stimuli (i.e., stimuli containing a weapon) from lures an applied x-ray luggage screening task (Experiment 2).

Theoretical Bases of Confidence

For perceptual tasks, confidence judgements have traditionally been measured retrospectively, meaning confidence is measured after a decision has been made and reflects the likely accuracy of the decision (e.g., if deciding which of two lines is longer, participants would first indicate the line that they believe to be longer, before providing a confidence rating), or simultaneously with the decision judgement (e.g., if deciding which of two lines is longer, response options may be “Low confidence Line 1”, “Moderate confidence Line 1”, “High confidence Line 1”, “Low confidence Line 2” etc.). It is often posited that confidence reflects the amount of evidence accumulated in favour of the chosen option, either in regards

to how much evidence has accumulated compared to the decision criterion (with confidence increasing in line with the extent to which evidence in favour of the chosen response exceeds the response criterion; e.g., SDT based models, Bernbach, 1971; Egan et al., 1959; Green & Swets, 1966), or in terms of the amount of evidence in favour of the chosen option compared to the alternative response option (with confidence increasing as the amount of evidence in favour of the chosen response compared to the non-chosen response increases; e.g., accumulator based models coupled a balance of evidence hypothesis; Smith & Vickers, 1998; Van Zandt, 2000; Vickers, 1970; Vickers & Lee, 1998).

As noted, non-evidential factors can lead to criteria becoming more conservative or more liberal. For example, the low prevalence effect (LPE; Wolfe et al., 2005) refers to the phenomena by which participants become more conservative in visual search tasks when target prevalence is low (i.e., when 10% or less of trials contain a target). Adopting a conservative approach under these circumstances can lead to more efficient decision making; the low likelihood of a target being present means that the most efficient approach is to generally opt for a response of "not present", allowing for the participant to move onto the next trial more quickly. The tendency toward conservative responding leads to an increase in misses (i.e., failures to detect target stimuli when present). The LPE commonly emerges in applied scenarios in which target prevalence is low, such as luggage screening (Wolfe et al., 2007). One method that has been suggested in the literature to reduce the effects of biases on responding is to use a rating-only approach (Sauer et al., 2012). Unlike traditional binary decisions, in which participants are asked to make an explicit yes/no or old/new judgement, a rating-only scale asks participants to provide a confidence rating in the absence of a binary decision (e.g., "how confident are you that Line A is longer than Line B? 0%, 10%, 20% … etc.").

A Ratings-Based Approach

In the psychophysics literature, using rating scales (cf. a categorical response) is not a new approach. Historically, researchers using signal detection-based approaches have often used ratings scales (e.g., a 6-point scale ranging from “sure old” to “sure new”, as used by Ratcliff et al., 1994) as an alternative for binary yes/no or old/new judgements. While this approach demonstrates the utility of a rating-scale, it still includes a dichotomous decision judgement (e.g., “old” vs “new”). In the present research, our interest specifically lies in rating scales that do not explicitly infer a decision; a concept that has been tested within the context of recognition memory. It has been demonstrated that participants are able to use a rating-only confidence scale to discriminate between studied and novel faces in face recognition tasks, as well as to identify culprits among foils in eyewitness identification tasks (e.g., Brewer et al., 2012, 2020; Sauer et al., 2008, 2012). Further, compared to binary decisions, confidence ratings made in the absence of explicit decisions have been shown to be less vulnerable to factors such as retention interval and distinctiveness, that have been demonstrated to impair discriminability, and to provide additional diagnostic evidence beyond that of a binary decision (Brewer et al., 2012; Cleary & Greene, 2000; Sauer et al., 2008; 2012; Weber & Varga, 2012).

One argument in favour of exploring a rating-only approach is the potential for reducing the effects of bias on decision making in perceptual discrimination tasks. As discussed, a decision is thought to be made when evidence accumulation reaches a predetermined threshold, and that where a threshold is set can be influenced by non-evidential biases (e.g., prior beliefs, LPE etc.). Confidence, on the other hand, is believed to be reflective of the total amount of evidence accumulated in favour of the chosen response, and is triggered when a decision has been reached. By focussing solely on confidence, and removing the explicit decision component, it may be that we are able to capture a more nuanced account of evidence accumulation, minimising the effects of bias on criteria (Brewer

et al., 2020). For example, a liberal criterion shift may result in a categorical visual search decision shifting from “target absent” to “target present”. When using a rating-only approach, however, this may be represented as a shift from 40% confidence to 50% confidence. While shifting a categorical response from target absent to target present (or vice-versa) completely changes the decision outcome (and can lead to serious consequences in applied contexts), it is less likely that the corresponding shift in confidence would have such a definite effect. Thus, bias-related effects on responding, and the errors they cause, may be mitigated.

The rating-only approach represents a simple, easily implemented alternative to the traditional responding, and may help attenuate bias-related decision errors. Beyond the fact that removing the requirement for an explicit, categorical response could allow for a more direct index of the evidence upon which decisions are made, it also takes criteria-setting out of the hands of the decision maker. Although this may appear a counter-intuitive goal, work demonstrating the utility of ratings-based approaches for person recognition and word recognition demonstrates that individuals often set sub-optimal criteria, and fail to make the best use of the evidence (i.e., memorial information) available to them. Instead, an independent third-party, or even a computer algorithm, could impose criteria upon the confidence ratings. To use the LPE in luggage x-ray screening as an example, low target prevalence leads to luggage screeners becoming more conservative in their responding, leading to an increase in miss errors. In this context, a third-party evaluator or computer algorithm would be able to account for the low target prevalence and therefore set a more lenient criterion regarding what constitutes an “identification” (e.g., a confidence rating of 30% or above may warrant inspection of a piece of luggage, even though that same level of underlying evidence may not be sufficient to return a positive categorical response). Further, such criteria would be able to vary between luggage screeners, adjusting for those who are dispositionally more liberal (or conservative) with their responding (Brewer et al., 2020; see

also Brewer et al., 2012; Sauer et al., 2008, 2012, for discussion of individual classification algorithms).

This research endeavoured to answer two questions. First, can participants utilise a 0-100% rating-only confidence scale to indicate whether a dynamic grid stimulus consists predominantly of a target colour (e.g., "Indicate your confidence that the flashing grid is predominantly ORANGE"; see Experiment 1)? Second, if so, can a rating-only approach be used effectively in a complex, applied perceptual discrimination task to indicate the likelihood that a target is present in a visual array (e.g., a luggage screening task; see Experiment 2)?

Experiment 1

Given the utility of rating-only approaches for recognition memory tasks, we tested whether participants could use a similar approach to effectively discriminate between trial types in a basic perceptual task. Specifically, Experiment 1 tested whether participants could use ratings (in the absence of categorical decision) to discriminate whether a dynamic grid was predominantly orange or predominantly blue (see Figure 1). We also included a manipulation of task difficulty - manipulating the proportion of the dominant colour - to observe whether these ratings were susceptible to the "hard-easy" effect; the phenomena by which participants overestimate their probability of success for a "difficult" task, while underestimating their probability of success for an "easy" task (Gigerenzer et al., 1991). This differs somewhat from the aforementioned models of confidence, which simply suggest that confidence should be lower for "hard" trials (i.e., a trial with less evidence in favour of the correct option) than "easy" trials (i.e., a trial with more evidence in favour of the correct response option). In the "hard" condition, 52% of cells were filled with the target colour (with the remainder being filled with the alternative colour), whereas in the "easy" condition,

54% of the cells were filled with the target colour (with the remainder being filled with the alternative colour).

The hard-easy effect was originally observed in tasks for which confidence was used as an index of likely accuracy (Gigerenzer et al., 1991). The hard-easy effect can be thought of as indicating that participants fail to lower their confidence sufficiently to reflect the reduction in performance. For a task such as ours, participants are not making an explicit decision judgement, and confidence is no longer an estimation of accuracy of that judgment. Instead, confidence indexes the evidence in favour of one response option. Thus, the extent to which these ratings exhibit the hard-easy effect (i.e., increased overconfidence under difficult conditions) is unclear. This raises the question of whether a rating-only approach may be more resilient to the hard-easy effect, or whether participants would still fail to adjust confidence sufficiently to reflect the difficulty of the task. Further, we were interested in whether we observe the basic patterns of confidence associated with difficulty (i.e., higher confidence for easy trials, lower confidence for hard trials), and whether ratings retain their utility in both hard and easy conditions.

Figure 1

Example of Dynamic Grid Stimuli

We hypothesised that participants would be able to use the rating-only confidence scale to discriminate between trial types (i.e., that participants' confidence that a stimulus was predominantly blue/orange would increase systematically with the likelihood that the

stimulus was blue/orange), as indexed by the Adjusted Normalised Discrimination Index (ANDI; Yaniv, Yates & Smith, 1991).

Further, we expected to see a systematic increase in confidence (as measured by a 6-point probabilistic confidence scale, e.g., 20%, 40% etc.) associated with a corresponding increase in the objective likelihood that the stimulus was the same as the target colour (i.e., blue or orange). We assessed this using calibration curves and associated indices of calibration (C) and over/underconfidence (OU).

Finally, we hypothesised that discrimination (as indexed by ANDI) would be higher for "easy" trials than "hard" trials. Based on previous findings (Gigerenzer et al., 1991; Baranski & Petrusic, 1997), we also expected to see increased overconfidence (or reduced under confidence) for "hard" compared to "easy" trials (as indexed by O/U statistics; with positive values indicating overconfidence, and negative values indicating underconfidence).

Method

Design

We used a 2 (difficulty: hard or easy) × 2 (focus colour: blue or orange) mixed design, with focus colour being the between-groups factor (i.e., participants were asked either to rate their confidence the stimulus was predominantly blue or predominantly orange). The dependent variable was participants' confidence ratings. Difficulty was manipulated within blocks of trials, with each block having an equal proportion of hard and easy trials.

Participants

Twenty-four participants (aged 21-52 years old) were randomly allocated into one of the "focus colour" conditions. Participants were required to have normal-to-corrected normal vision and, due to the "flashing" nature of the stimuli, were not eligible to participate if they suffered from epilepsy or related conditions. First year psychology students received research participation credit and other participants were reimbursed with a $15 gift-voucher.

Materials and Procedure

The experiment was completed on desktop computers equipped with 3.30 GHz Intel i5-6600 processors, 16 GB RAM, and a Windows 7 enterprise operating system configured to minimise internal task-switching. The program was written and run using MATLAB (The MathWorks, R2016b). Each participant completed 10 blocks of 50 trials (and one block of training trials), within which they saw an equal number of hard and easy trials, and an equal number of dominant-blue and dominant-orange trials.

For each trial, participants indicated their confidence as to whether a dynamic grid was "primarily blue" or "primarily orange" by moving their cursor over the edge of an arc-shaped confidence scale. The confidence scale consisted of six numerical responses, 0%, 20%, 40%, 60%, 80% and 100%. The dynamic grid consisted of blue (RGB = 0, 65, 255) and orange (RGB = 255, 127, 0) 8 x 8 pixel cells (see *Figure 1)*. Participants received warnings telling them to respond faster if they were slower than 2 seconds, or slower if they were faster than 15 milliseconds.

Results and Discussion

Calibration and discrimination typically refer to the extent to which retrospective confidence (i.e., confidence after a decision) maps onto the likely accuracy of a decision, and discriminates correct from incorrect decisions (Keren, 1991; Yaniv et al., 1991). Here, however, calibration and discrimination refer to the extent to which confidence reflects the likelihood that the stimulus is the target colour (e.g., if the stimulus is predominantly blue, and the instruction read "Indicate your confidence that the square is primarily BLUE"), and discriminates between cases where the stimulus is and is not the target colour. For example, if we assume ideal calibration, we would expect the stimulus to be predominantly made up of the target colour in 80% of trials in which participants reported 80% confidence, for 60% of the trials in which participants reported 60% confidence, and so on.

Table 1

Mean ANDI Scores, C Statistics and O/U Statistics Across Difficulty for Experiment 1

Difficulty	Easy trials		Hard trials		Combined	
	M (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
ANDI	.42 (.24)	[.32, .52]	.16 (.12)	[.11, .21]	.29 (.17)	[.19, .21]
C	.25 (.08)	[.21, .28]	.18 (.05)	[.16, .20]	.20 (.09)	[.16, 24]
O/U	.02 (.07)	[-.05, .01]	-.02 (-.02)	[-.05, .02]	-.02(.07)	[-.05, .01]

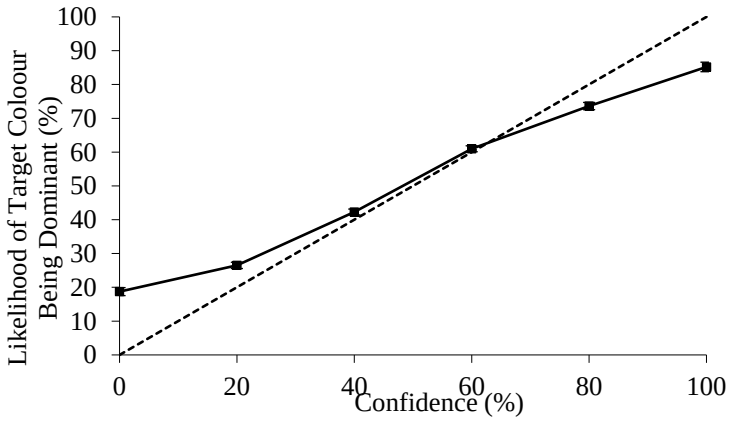
Note. CI = confidence interval.

Of particular interest was whether participants would be able to use these ratings to discriminate between trials in which the stimulus does consist predominantly of the target colour, compared to trials in which it does not consist predominantly of the target colour. This is demonstrated by the fact that 95% confidence intervals for the combined ANDI score do not overlap 0 (see Table 1). This remained true for both hard and easy trials.

However, consistent with our third hypothesis, participants showed better discrimination (i.e., increased ANDI scores) in the easy trials, $t(46) = 4.79$, 95% CI [.15, .36], $p < .001$.

Figure 2

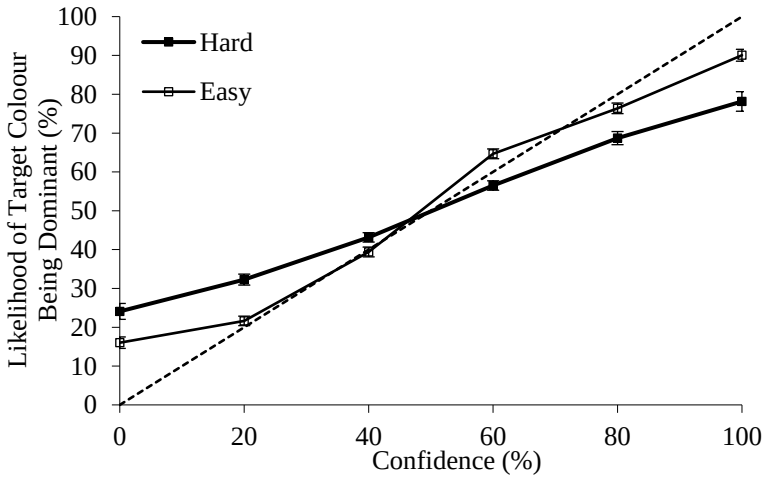
Likelihood of Target Colour Being Dominant as a Function of Confidence



Note. Calibration curve plotting the likelihood of the target colour being dominant as a function of variations in confidence. The dotted line demonstrates perfect calibration. Error bars represent the standard error.

Figure 3

Likelihood of Target Colour Being Dominant as a Function of Confidence, Separated for Hard and Easy Trials



Note. Calibration curve plotting the likelihood of the target colour being dominant as a function of variations in confidence for hard and easy trials. The dotted line demonstrates perfect calibration. Error bars represent the standard error.

Figures 2 and 3 show a systematic, positive relationship between ratings and likelihood of the stimulus consisting predominantly of the target colour, both collapsed across difficulty levels (Figure 2), and for each difficulty level (Figure 3). Thus, despite some underconfidence at lower levels of the scale and overconfidence in the top half of the scale, participants were able to use ratings to indicate trial type.

The O/U statistics provided in Table 1 show no evidence of a general increase in overconfidence in the hard condition, as the 95% confidence intervals overlap zero for both the hard trials and the easy trials. However, a look at the curves in Figure 3 suggests increased overconfidence in the upper half of the confidence scale for the hard condition compared to the easy condition. This was offset by increased underconfidence at lower levels of confidence, possibly resulting the similar "overall" levels of overconfidence evident in the summary O/U statistics. Thus, there is some evidence that ratings might show hard-easy effects, particularly in the upper half of the scale.

Overall, results supported our hypotheses. Participants could use ratings to discriminate between stimuli that did/did not consist predominantly of the target colour, and the probability of a stimulus consisting predominantly of the target colour increased systematically with confidence for both hard and easy trials (despite some evidence for increased overconfidence in hard trials). Thus, a rating-only scale was effective for a basic perceptual discrimination task. These findings provide a foundation for Experiment 2, which examines the utility of these ratings in a more complex, applied perceptual task, and tests whether using ratings (cf. categorical responses) can attenuate bias-related errors in a low target prevalence domain.

Experiment 2

Experiment 2 tested the utility of the rating-only approach in a naturalistic decision scenario: A luggage screening task. In this context, target prevalence can impact criterion

placement, consequently affecting the decisions people make such that rare targets are often missed because responding has become more conservative. The rating-only approach may reduce these effects by reducing the impact of criteria shifting to become more conservative. To test this possibility, we manipulated the rate at which target items (i.e., a weapon) appeared in a luggage screening task and examined how this affected task performance using a standard categorical response method and a rating-only approach. Our first aim was to ascertain whether the rating-only measure could be used effectively to discriminate target present from target absent trials. We also investigated how performance using the rating-only scale varied dependent upon the external criterion we applied (i.e., the magnitude of rating required to indicate the presence of a target). Additionally, we examined how the rating-only approach would fare in a low prevalence visual search task, and whether a ratings-based approach could discriminate between target present and target absent trials under low prevalence conditions that typically increase miss errors for binary responses.

We hypothesised, based on the results of Experiment 1 and previous work in the recognition memory domain, that participants would be able to use the rating-only scale to detect targets as evaluated by calibration and discrimination measures. In particular, we expected participants to be able to use the scale to discriminate between images that did and did not contain a target (assessed via ANDI). We also predicted that (a) miss errors would be higher in low prevalence trials compared to high prevalence conditions, and (b) this effect would be at least somewhat attenuated in the rating-only condition, such that fewer targets were missed.

To test the latter predictions, we classified rating-only responses as “correct” or “incorrect” based on a split between each confidence level. In this way, the data can be thought of as almost equivalent to binary responses, with one exception: the individual has not imposed their own criterion on what they consider enough evidence to warrant a target

present response. As an example, we can consider all responses ≥ 60% as indicating the presence of a target, and therefore as hits for target present trials and false alarms for target absent trials. Thus, all responses ≤ 40% would be deemed misses for target present images and correct rejections for target absent images. We then repeated this process splitting the data at each confidence level (further detail on this matter below). As a result of this classification process, we could assess participant responses in terms of accuracy and by using signal detection measures. Overall, we were able to test (a) whether rating-only responses could discriminate target present from target absent images and (b) how ratings fared in terms of classification relative to traditional binary responses.

Method

Design

We used a 2 (target prevalence: high [50%] or low [3.3%]) × 2 (order: high prevalence first or low prevalence first) × 2 (response-type: rating-only or binary) × 2 (target presence: target absent or target present) mixed design. Target presence and prevalence were manipulated within-subjects; whilst response-type and order were between-subjects variables.

Participants

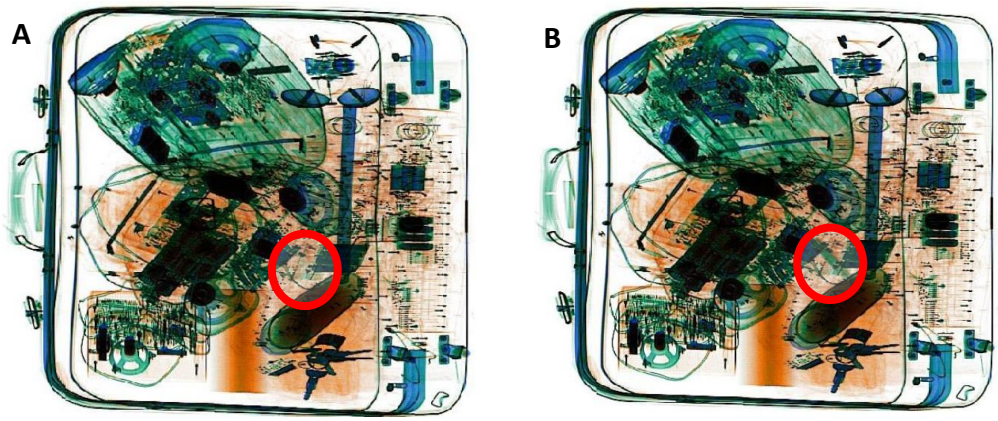
A total of 99 participants (aged 18-75 years old) were recruited for this study. All participants were required to have normal (or corrected-to-normal) colour vision. We removed eight participants (four from each response-type) who failed the attention check (i.e., providing a specific key response when prompted). Another participant was excluded from the binary condition due to complete invariant responding across multiple blocks, resulting in a final sample size of 90. The sample included both first-year psychology students and members of general population, who received course credit or $30 gift voucher as compensation for their time.

Materials

A total of 576 security-like images of luggage were taken from previous research (McCarley et al., 2004). These were target absent and target present coloured x-ray images displayed on a white background (see Figure 4). Some target absent images were inverted and/or rotated so that a sufficient number of trials could be implemented in the experimental phase. However, no target present images were repeated on target absent trials and images presented in a practice block were not used as experimental images. The images varied in difficulty as well as the degree of visual clutter they contained. These factors remained comparable across all experimental blocks. The target item was a knife, which was the same in each image and approximately 100 pixels in length. It was placed at randomly selected locations and at a randomly chosen orientation of 0°, 45°, 90°, 135°, 180°, 225°, 270°, or 315° in the frontoparallel plane.

Figure 4

Examples of Stimuli Used for Experiment 2



Note. Example of target absent (A) and target present (B) stimuli. The red circle is for illustrative purposes to signify target presence vs. absence.

Procedure

This experiment was conducted online, with participants using their own device. We limited participation to those with either a laptop or desktop computer to ensure stimuli were correctly displayed. The study was run via the online platform Pavlovia (pavlovia.org) and

participants received initial information regarding the experiment along with a URL to the study. After providing informed consent, participants completed a practice block of 20 trials at 50% prevalence to familiarise themselves with the task. This was followed by the experimental phase which consisted of 9 blocks of 64 trials (576 images in total). Eight of the experimental blocks were low prevalence (3.3% approx.) where the target appeared on 2-3 trials per block. The remaining block was high prevalence where target prevalence was set at 50%. Participants were randomly assigned to either start or finish with the high prevalence block. There were enforced breaks, one of which separated the high and low prevalence conditions to limit the transfer of criterion effects, and optional breaks, which occurred between each block.

On each trial, participants viewed a single item of luggage and searched for a knife. Those in the rating-only condition responded to the prompt: *Confidence knife is present*. Confidence was presented in 20% increments from 0 to 100% and participants were required to press the key corresponding to their level of confidence for that trial. In line with the scale used in Experiment 1, there were no other labels indicating how the scale should be interpreted. This was to avoid participants engaging in categorical decision making. For the binary condition, participants responded to the prompt: *Is there a knife present?* via key press to indicate yes or no. Following this they provided their confidence in their decision. This was indicated in a similar way to the rating-only condition with 20% increments, however, the labels on the anchor points displayed *not at all confident* (at 0%) and *completely confident* (at 100%). Total completion time for the experiment was approximately two hours.

Results and Discussion

When considering our treatment of the data, there are two points to note. First, retrospective confidence in binary condition and confidence ratings in the rating-only condition have generally been analysed separately as the confidence responses required for

each condition are not always directly comparable. Confidence in the binary condition represents confidence in the decision made, while rating-only confidence corresponds to confidence in target presence. Second, for the binary condition, responses have been separated into positive (i.e., target present responses) and negative (i.e., target absent responses) decisions where appropriate. Note that some data were skewed (e.g., C statistic and ANDI across both conditions), however, application of a square root transformation often did not eliminate skew. Analyses based on raw and transformed data produced equivalent results. Thus, we report the results based on raw data for ease of interpretation.

Before evaluating participants' ability to effectively use the rating-only scale to discriminate target absent from target present searches, we established that participants were able to complete the task by considering objective accuracy in the binary condition. Accuracy in the high prevalence binary condition for target present and target absent images was 83.2% and 83.6%, respectively, indicating that detection of the target knife was possible. For experimental purposes, this level of accuracy is acceptable – performance is at neither floor nor ceiling meaning there is space to investigate and improve task accuracy.

Calibration and Discrimination

As in Experiment 1, the capacity to use the rating-only confidence scale has been assessed with calibration and discrimination statistics. Again, we used ANDI (Yaniv et al., 1991) to assess whether ratings effectively discriminated target present from target absent images. For both the low and high prevalence conditions, the 95% confidence intervals around mean ANDI scores did not overlap with zero (see Table 2). Thus, as hypothesised, ratings could be used to discriminate target present from absent trials, even under low prevalence conditions. We also compared mean ANDI scores for the high and low prevalence trials. A paired samples *t*-test showed significantly better discrimination in the high prevalence compared to low prevalence condition, $t(42) = 3.89$, 95% $CI_{difference}$ [.06, .18], $p <$

.001. In sum, discrimination performance was superior in the high prevalence condition, however, participants could still clearly use the scale to differentiate between target absent and target present images in low prevalence searches.

We also examined the extent to which retrospective confidence discriminated correct from incorrect responses in the binary response condition. For positive responses (i.e., indicating a target was present), discrimination was similar to that of the rating-only condition (see Table 2). However, for negative responses (i.e., responding target absent), mean ANDI scores were close to zero, showing that confidence did little to discriminate between correct and incorrect responses. We note that accuracy for negative responses was generally high (i.e., over 90% correct across confidence levels) in the low prevalence condition (because a negative response was overwhelmingly likely to be correct in those conditions), and this possibly constrained any potential relationship between confidence and accuracy. However, this asymmetry in ANDI between positive and negative decisions has commonly been observed in other tasks such as face recognition and eyewitness identification (e.g., Brewer & Wells, 2006; Weber & Brewer, 2004). Additionally, negative responding in the high prevalence condition was associated with a slightly lower degree of accuracy, with accuracy sitting around 80% for all confidence levels besides 0% confidence which was at 60% accuracy.

Table 2

Mean ANDI Scores Across Prevalence Blocks for Ratings-Only and Binary (Positive and Negative Responses) Conditions

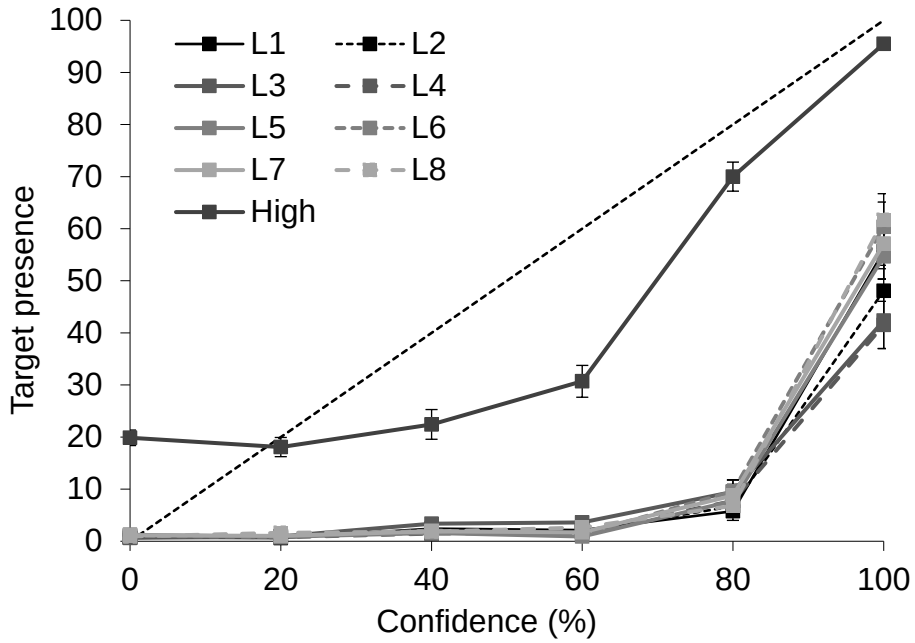
	High prevalence			Low prevalence		
Condition	*n*	*M* (*SD*)	95% CI	*n*	*M* (*SD*)	95% CI
Rating-only	43	.56 (.23)	[.49, .64]	43	.44 (.25)	[.37, .52]
Binary positive	40	.50 (.29)	[.41, .60]	47	.49 (.20)	[.43, .54]
Binary negative	42	.05 (.09)	[.02, .07]	44	.03 (.11)	[.00, .07]

Note. CI = confidence interval.

We also plotted calibration curves to assess the degree of correspondence between participants' subjective judgments (i.e., confidence ratings) and the objective probability of target presence. Overall, as illustrated in Figure 5, rating-only responses in low prevalence blocks showed marked overconfidence at all levels of confidence 80% and below (besides 0% which reflected perfect calibration), with this overconfidence decreasing (but still evident) at 100%. This highlights an important, though perhaps unsurprising, finding: the objective likelihood of a target being present noticeably increased when participants provided a confidence rating of 100%. Thus, a rating of 100% offers at least somewhat meaningful information about the likely presence of a target. Many positive responses will turn out to be incorrect as there are so few targets in low prevalence search. If participants were able to use the scale adequately, we would expect this to reduce at the highest level of confidence since, ideally, high levels of confidence would be reserved for target present trials. For the high prevalence block, responses in the rating-only condition were generally well-calibrated, despite some overconfidence (and underconfidence at 0%).

Figure 5

Calibration Curves for Rating-Only Responses



Note. Calibration curves across prevalence blocks at each level of confidence for rating-only responses. The dotted line represents perfect calibration. L1-L8 represent the eight low prevalence blocks. Error bars represent standard errors.

To augment consideration of these curves, we also calculated C and OU statistics (Table 3). A mixed ANOVA evaluating calibration (C statistic) across order and prevalence indicated that the main effect of prevalence was significant, $F(1,41) = 15.86$, $p < .001$, indicating better calibration in the high prevalence block compared to the low prevalence condition. Consistent with the interpretation of Figure 5, confidence ratings better reflected target presence in the high prevalence, compared to low prevalence, condition. The order of prevalence presentation did not produce significant differences in calibration and the order × prevalence interaction also was not significant (Fs < 2, ps > .2). Over and underconfidence (measured using the O/U statistic; see Table 4) was also analysed using a mixed ANOVA, where both the prevalence main effect, $F(1,41) = 95.11$, $p < .001$, and order × prevalence interaction were significant, $F(1,41) = 5.35$, $p = .026$. The analysis revealed no significant

main effect of order ($F < 1$, $p > .6$). In both order conditions, participants maintained roughly an equal degree of overconfidence across the low prevalence conditions. The main difference being that those beginning with the high prevalence block were slightly overconfident, whilst those finishing with this block had a mean score of almost zero (i.e., no over/underconfidence).

Table 3

Mean C Statistic Across Prevalence Blocks for Binary (Positive and Negative Responses) and Rating-Only Responses

	High prevalence		Low prevalence	
Condition	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Rating-only				
High-low (*n* = 22)	.04 (.03)	[.02, .06]	.09 (.09)	[.06, .13]
Low-high (*n* = 21)	.06 (.05)	[.05, .08]	.11 (.09)	[.07, .15]
Binary positive				
High-low (*n* = 24)	.03 (.02)	[.02, .05]	.13 (.08)	[.09, .17]
Low-high (*n* = 23)	.04 (.04)	[.02, .05]	.16 (.12)	[.12, .20]
Binary negative				
High-low (*n* = 24)	.13 (.09)	[.09, .18]	.21 (.15)	[.15, .28]
Low-high (*n* = 23)	.15 (.13)	[.10, .20]	.27 (.18)	[.20, .34]

Note. CI = confidence interval.

Table 4

Mean O/U Statistic Across Prevalence Blocks for Binary (Positive and Negative Responses) and Rating-Only Responses

	High prevalence		Low prevalence	
Condition	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Rating-only				
High-low (*n* = 22)	.05 (.10)	[-.01, .11]	.19 (.14)	[.13, .25]
Low-high (*n* = 21)	-.02 (.17)	[-.08, .04]	.22 (.14)	[.16, .28]
Binary positive				
High-low (*n* = 24)	-.01 (.09)	[-.06, .03]	.13 (.08)	[.09, .17]
Low-high (*n* = 23)	-.04 (.12)	[-.09, .01]	.16 (.18)	[.12, .20]
Binary negative				
High-low (*n* = 24)	-.26 (.15)	[-.33, -.19]	-.40 (.15)	[-.46, -.33]
Low-high (*n* = 23)	-.28 (.19)	[-.35, -.21]	-.46 (.18)	[-.53, -.39]

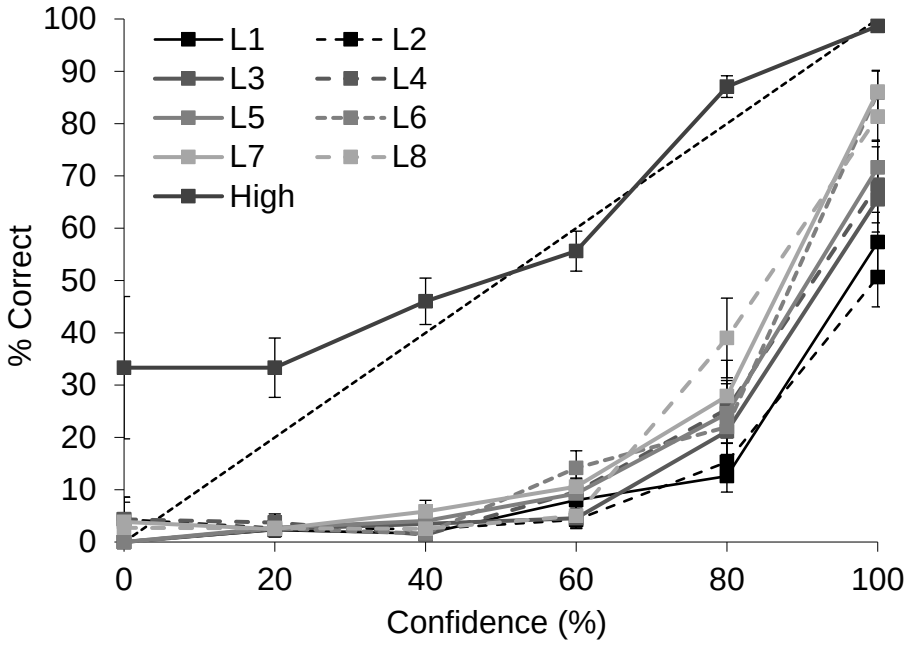
Note. CI = confidence interval.

For positive responses in the binary condition, Figure 6 shows a systematic, positive relationship between confidence and accuracy emerged (despite some underconfidence at the lower end of the confidence scale). Similar to the rating-only condition, extreme overconfidence was present for the most part in the low prevalence conditions, with calibration starting to improve at the highest confidence levels (see Tables 3 and 4 for C and O/U descriptive statistics). A mixed ANOVA assessing the effect of prevalence and order revealed the main effect of prevalence was significant, $F(1,45) = 52.41$, $p < .001$, providing evidence in support of participants being better calibrated in the high prevalence condition. That is, their subjective confidence ratings were better aligned with objective performance in

comparison to the low prevalence conditions. There was no significant main effect of order and the order × prevalence interaction also was not significant (Fs < 2, ps > .2). The same analysis applied to the O/U statistic revealed a significant main effect of prevalence, $F(1,45) = 52.41$, $p < .001$, whereby participants were overconfident in the low prevalence conditions compared to the high prevalence condition, where they were only marginally underconfident. Both the main effect of order and the interaction were not significant (Fs < 4, ps > .08).

Figure 6

Calibration Curves for Positive Binary Responses



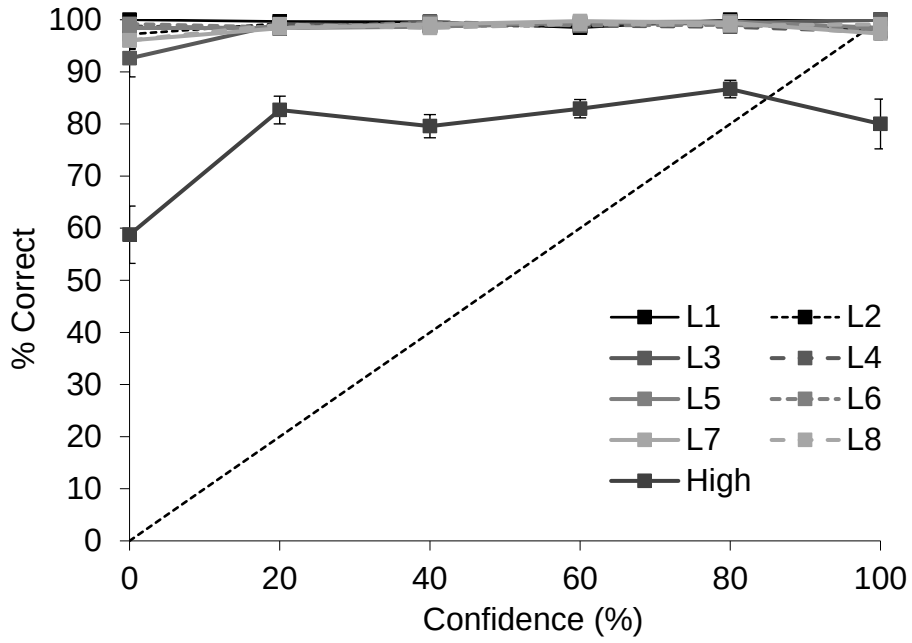
Note. Calibration curves across prevalence blocks at each level of confidence for positive binary responses. The dotted line represents perfect calibration. L1-L8 represent the eight low prevalence blocks. Error bars represent standard errors.

For negative responses, participants were generally poorly calibrated in the low prevalence and high prevalence blocks, though underconfidence was more pronounced for low prevalence trials (see Figure 7). A mixed ANOVA on the C statistic indicated the prevalence main effect was significant, $F(8,360) = 9.23$, $p < .001$, such that calibration was superior in the high prevalence condition (possibly because accuracy was near ceiling in the

low prevalence condition, leaving little opportunity for systematic covariation between confidence and accuracy). However, as evident in Figure 7, neither the high nor low prevalence conditions showed evidence of a systematic relationship between confidence and accuracy for negative responses. Neither the order main effect nor the order × prevalence interaction were statistically significant (Fs < 2, ps > .2). Moreover, in terms of over/underconfidence, a mixed ANOVA revealed a significant main effect of prevalence, $F(1,45) = 50.92$, $p < .001$. This provides support for the idea that participants were less underconfident in the high prevalence compared to low prevalence conditions. There was no significant main effect of order or interaction (Fs < 1, ps > .3).

Figure 7

Calibration Curves for Negative Binary Responses



Note. Calibration curves across prevalence blocks at each level of confidence for negative binary responses. The dotted line represents perfect calibration. L1-L8 represent the eight low blocks. Error bars represent standard errors.

Adjusting the Criterion for Rating-Only Responses

We established that the rating-only scale could be used to discriminate target present from target absent images but that it resulted in poor calibration in the low prevalence conditions. Next, we examined whether it was possible to use the rating-only approach to detect targets in conditions that might otherwise lead to increased misses when reliant upon binary judgments. In this way, the rating-only procedure may be advantageous as it allows external decision makers to adjust the criterion for what is classified as a positive (i.e., target present) response. In this section we detail how altering the criteria that determined when a rating was taken as indicating the presence of a target could be implemented to reduce conservative responding and how performance then compared to that of the binary condition.

Responses were analysed according to the two parameters critical in signal detection theory (Green & Swets, 1966): discriminability (i.e., ability to discriminate target present from target absent trials) and response criterion (i.e., the degree of evidence needed to provide a "present" response as distinct from an "absent" response). The former can be measured by *d*', which calculates the distance between the signal (target present) and noise (target absent) means (using standard deviation units), whilst the latter has been assessed using *c*, the distance between an ideal observer (minimisation of misses and false alarms) and the actual threshold of participants.

We begin by analysing the binary data to determine whether low prevalence conditions were associated with increased miss errors (i.e., consistent with the typical LPE findings). We excluded four participants (two from each order condition) as outliers (as indicated by *z*-scores > ±2.58). To compare high and low prevalence conditions to determine whether the low conditions were more conservative overall, we conducted a mixed ANOVA on mean *c* values. The main effects of prevalence and order, along with the prevalence × order interaction were not significant (Fs < 4, ps > .05). While, numerically, criterion was slightly more conservative in the low prevalence condition (M = 0.36, SD = 0.33) compared

to the high prevalence condition (M = 0.05, SD = 0.42), the lack of significant findings suggest the LPE did not emerge as strongly as we would expect. In terms of discriminability, the main effect of order and the prevalence × order interaction were not significant (Fs < 4, ps > .06). However, the prevalence main effect was significant, $F(1,41) = 5.31$, $p = .026$. Participants showed superior discrimination in the high prevalence trials (M = 2.2, SD = 0.7, 95% CI [1.98,2.39]) compared to the low prevalence trials (M = 2.0, SD = 0.7, 95% CI [1.77,2.23]).

Typically, signal detection measures use accuracy data (i.e., hits, misses, false alarms, and correct rejections). As the rating-only approach omits the explicit decision component, to allow an analysis of this condition, we adapted the approaches to classification used in research with similar rating-only scales (e.g., Brewer et al., 2012; Sauer et al., 2008, 2012). For present purposes, we categorised confidence responses as indicating "target present" for responses equal to or above each confidence level and responses as "target absent" for confidence ratings below the given confidence level to determine how the different classifications impacted performance outcomes. For example, we first categorised confidence ratings of 60% and above as though the participant had responded "target present", and responses 40% and below as though they had responded "target absent". Thus, confidence ≤ 40% was considered "correct" for target absent trials and incorrect for target present trials, while confidence ≥ 60% was "incorrect" for target absent trials and "correct" for target present trials. We repeated this process at each of the confidence levels.

Here we report the results from an analysis of the 20%-40% confidence split to see the effects of a less conservative classification when comparing the prevalence conditions. One participant was considered an outlier (z-score > 2.58) on the criterion measure and removed from these analyses. We compared c for the high and low prevalence conditions across order. A mixed ANOVA found no significant difference between the low prevalence

(M = -0.2, SD = 0.5) and high prevalence blocks (M = -0.3, SD = 0.7), $F(1,40) = 1.69$, p = .201. The order main effect and prevalence × order interaction were both not significant (Fs < 3, p > .1). Interestingly, with this confidence classification, the overall responding for low prevalence blocks was slightly liberal and does not show the typical LPE whereby responding is much more conservative in comparison to high prevalence scenarios. This provides some evidence in support of the ability of the rating-only scale to produce less conservative responding.

While of lower importance, for completeness we also compared discriminability across prevalence order and prevalence conditions. A mixed ANOVA revealed both main effects and the interaction were not significant (Fs < 3, p > .1). This finding suggests participants were able to discriminate between target absent and present trials to a similar degree for both prevalence conditions and orders.

Task Accuracy

As highlighted, a key feature of the rating-only scale is the ability to externally shift criterion placement. Tables 5 and 6 show how the outcome of the visual search task varies as a result of the differing classifications for low and high prevalence, respectively. Logically, as the confidence split moves closer to zero, the miss rate reduces and the false alarm rate increases. For low prevalence search in the binary condition, the miss rate was .19, and the false alarm rate was .17. The former maps directly to the confidence rating of 40% and above being classified as target present responses, and the 60% rating and above being considered as target present closely resembles the latter. For high prevalence search in the binary condition, the miss rate was .17, and the false alarm rate was .16. In the rating-only group, again, the categorisation of 40% and above as target present most closely resembles the former, while the latter is closest to the 60% and above rating.

Table 5

Proportion and Number of Hits, Misses, False Alarms, and Correct Rejections at Each Classification Level for Combined Low Prevalence Search

Confidence level	Hits (*n*)	Misses (*n*)	False alarms (*n*)	Correct rejections (*n*)
100%	.56 (423)	.44 (333)	.02 (389)	.98 (20359)
80%	.70 (527)	.30 (229)	.08 (1624)	.92 (19124)
60%	.75 (567)	.25 (189)	.17 (3550)	.83 (17198)
40%	.81 (616)	.19 (140)	.28 (5911)	.72 (14837)
20%	.89 (674)	.11 (82)	.58 (12017)	.42 (8731)
0%	1 (756)	0 (0)	1 (20748)	0 (0)

Note. Confidence values indicate the level considered as "positive" responses. For example, 100% reflects responses of 100% being classified as "target present" (and ratings below this as being "target absent"), 80% reflects responses at or above 80% being classified as target present, and so on until 0%, which reflects all responses being considered target present.

Table 6

Proportion and Number of Hits, Misses, False Alarms, and Correct Rejections at Each Classification Level for High Prevalence Search

Confidence level	Hits (*n*)	Misses (*n*)	False alarms (*n*)	Correct rejections (*n*)
100%	.61 (820)	.39 (524)	.03 (40)	.97 (1304)
80%	.75 (1008)	.25 (336)	.09 (121)	.91 (1223)
60%	.80 (1078)	.20 (266)	.21 (279)	.79 (1065)
40%	.84 (1126)	.16 (218)	.33 (444)	.67 (900)
20%	.90 (1205)	.10 (139)	.59 (798)	.41 (546)
0%	1 (1344)	0 (0)	1 (1344)	0 (0)

Note. Confidence values indicate the level considered as "positive" responses. For example, 100% reflects responses of 100% being classified as "target present" (and ratings below this as being "target absent"), 80% reflects responses at or above 80% being classified as target present, and so on until 0%, which reflects all responses being considered target present.

To explore the flexibility provided by the rating-only approach, we assessed accuracy in the low prevalence condition at the 20-40% split and 40-60% split as these aligned most closely with the binary outcomes. We begin with analysis of the 20-40% confidence split (i.e., responses ≥ 40% were categorised as false alarms in target absent searches, and ratings ≤ 20% were categorised as misses in target present searches). In all conditions, we excluded the same participants as for *c* and *d*', and two additional outliers from the low-high order condition for the binary group (for the analysis of miss rate). In analysing the effect of prevalence on miss rates across response-type and prevalence order, a three-way mixed ANOVA revealed the main effects of prevalence, order, and response-type, while all two-way and three-way interactions were not significant (Fs < 4, ps > .05). This indicated that the proportion of misses did not vary significantly across high and low prevalence searches or in

relation to response-type or prevalence order. The same analysis but with the data split at 40-60% confidence, revealed a significant main effect of prevalence, $F(1,79) = 6.94$, $p = .010$ and order, $F(1,79) = 4.00$, $p = .049$. The response-type main effect and all interactions were non-significant. Thus, there were no significant effects of response-type at either confidence division when it came to miss errors, suggesting that although the proportion of misses varied numerically across the confidence splits, there was no significant difference when compared to the binary condition, in either case.

For false alarms, we begin with analysis of the 20-40% confidence split and binary data. The outcome of the three-way mixed ANOVA comparing the prevalence conditions across prevalence order and response-type revealed the response-type main effect was significant, $F(1,81) = 11.78$, $p < .001$, such that there were more false alarms in the rating-only compared to the binary condition. Following the application of Bonferroni corrections for multiple comparisons, the prevalence × order interaction no longer revealed significant differences. Additionally, the main effects of prevalence and order and the remaining interactions were not significant (Fs < 4, ps > .08). When considered alongside the miss rate analysis, the only significant difference between responding in the rating-only condition at the 20-40% confidence split, and the binary condition was that false alarms were less likely to occur in the binary condition.

Next, we conducted the same three-way ANOVA on false alarm errors at the 40-60% confidence split and the binary data. 0There were no significant main effects or interactions (Fs < 4, ps > .08), suggesting that the proportion of false alarms did not vary significantly in terms of response-type and prevalence order, or across high and low prevalence searches. Taken together, the false alarm and miss error findings indicated that there were no significant differences concerning this confidence split and the binary data. The only significant variation was in regard to prevalence order and prevalence in terms of miss errors.

In sum, the results of the accuracy analyses demonstrate how outcomes can vary dependent upon the criterion employed. We saw that the more liberal criterion of 20-40% produced significantly more false alarms compared to the traditional binary measure, but that this difference was no longer present at the 40-60% confidence split. The ability to adjust criteria to systematically reduce the risk of certain error types depending on contextual demands may be useful in applied settings. However, interestingly, in the present experiment there was little evidence that ratings offered improved overall performance relative to the binary condition.

General Discussion

The current paper had two overall aims. First, to determine whether a rating-only confidence scale could be used by participants to effectively discriminate stimulus types in a basic perceptual discrimination task (i.e., pertaining to the likelihood that a dynamic grid was primarily blue/orange). Second, to determine whether a rating-only approach could be used in a low prevalence luggage screening task and whether such an approach might mitigate errors associated with the LPE (Wolfe et al., 2005). Experiment 1 provided clear support for the utility of a rating-only approach for a basic perceptual task, but the utility of such an approach for reducing the LPE was less clear cut.

Experiment 1 required participants to indicate their confidence (with one 6 response options ranging from 0% to 100%) as to whether a dynamic grid consisting of blue and orange pixels was predominantly blue, or predominantly orange. To determine that participants were able to use the scale to effectively communicate their confidence, we expected to observe three key patterns. First, that participants would be able to discriminate between trials in which the stimuli did consist predominantly of the target colour vs trials in which the stimuli did not consist predominantly of the target colour. Second, that as participants confidence increased, the likelihood of the stimuli consisting predominantly of

the target colour would increase in turn. Finally, that discrimination would be higher for "easy" trials than "hard" trials.

The results of Experiment 1 supported all three of these key points. Discrimination values (ANDI) were higher for trials in which the target did consist predominantly of the target colour compared to trials in which it did not, demonstrating that participants could use the confidence scale to discriminate between correct and incorrect responses. Calibration curves showed that participants were well calibrated with their responses, meaning that the likelihood of the statement being correct increased as confidence increased. Finally, participants showed better discrimination in easy trials than hard trials, suggesting that the scale was sensitive to factors that should decrease confidence (i.e., difficulty). While we did not see an overall trend of overconfidence in the hard condition, we did see some evidence of the hard-easy effect, as evidenced by increased overconfidence in the upper half of the confidence scale for the hard condition compared to the easy condition. These results together provide strong evidence for the utility of rating-only confidence scales as an informative measurement tool for perceptual tasks.

Experiment 2 extended upon Experiment 1 to evaluate the use of the rating-only scale in a more complex, applied scenario. More specifically, a luggage screening task. Our primary interest lay in whether participants could use the rating-only scale to discriminate between target present and target absent trials. A secondary interest lay in whether a rating-only approach would reduce a common source of bias associated with low prevalence visual search tasks – the tendency for participants to respond more conservatively when target prevalence is low, compared to when target prevalence is high (i.e., the LPE; Wolfe et al., 2005). We considered the binary condition when determining the emergence of the LPE and while we did observe superior discrimination (*d*') for high prevalence trials compared to low prevalence trials, along with numerically more conservative criterion (*c*) values, we did not

find participants to be significantly more conservative in the low prevalence condition. Miss errors were also numerically higher, though not significantly, in the low prevalence condition. Moreover, the literature includes research noting either higher miss rates in low prevalence conditions than we did (e.g., Peltier & Becker, 2020; Wolfe et al., 2007) or lower miss rates in high prevalence conditions (e.g., Thomson & Goodhew, 2021; Wolfe et al., 2005). These discrepancies between our data and other findings in the literature limit the extent to which we can draw strong conclusions as to the utility of a rating-only response scale as a method of reducing common sources of bias such as those arising in low prevalence visual search.

Nevertheless, several aspects of the results provide reason to be optimistic that a ratings-only approach may prove useful for tasks involving low target-prevalence. First, we found evidence in support of our key hypothesis: that participants would be able to use the rating-only scale to effectively discriminate between target present and target absent trials, as evidenced by the reported discrimination measures (both ANDI and *d*'). Specifically, ANDI scores were significantly greater than zero for both high and low prevalence conditions, and *d*' values indicated good discriminability that did not differ significantly between the prevalence groups. Another point of interest centres around the advantage of the rating-only scale when evaluated by ANDI scores – as the binary judgment is eliminated, the scale does not generate the same asymmetry in discrimination accuracy between positive and negative responses. For binary responses, ANDI scores indicated good discrimination for positive responses, however, discrimination was quite poor (i.e., close to zero) for negative responses. This finding is consistent with other domains in the literature including face recognition (Weber & Brewer, 2004) and eyewitness identification (Brewer & Wells, 2006; Palmer et al., 2012; Sauer et al., 2010). Thus, in both prevalence conditions, confidence ratings for negative binary responses did little to discriminate correct from incorrect responses. In contrast, with

the rating-only scale (with binary judgment eliminated), ANDI scores indicated that discrimination across all ratings-only responses was comparable to that for positive binary responses. Hence, the ratings-only scale avoids the problem of poor discrimination for negative responses.

Finally, one of the key features of a rating-only scale is that criterion can be imposed independently from the decision-maker to optimise decision outcomes. Thus, we interpreted our data at both a 40-60% confidence split (with ratings of 40% and below classified as "target absent", and 60% and above classified as "target present") as well as a 20-40% split (with ratings of 20% and below classified as "target absent", and 40% and above classified as "target present"). To give an example: if a participant were to report 40% confidence in a target present trial, this would be classed as incorrect for the 40-60% split but correct for the 20-40% split. Shifting the decision criteria affected responses in multiple ways. First, adopting a more liberal decision criterion (i.e., dropping the threshold for a target present response from 60% to 40%) led to an increase in the number of hits, as well as an increase in the number of false alarms. Logically, this also produced a decrease in misses and correct rejections. Second, responding was more liberal for low prevalence trials. These findings demonstrate that controlling the ability to impose criteria can affect performance outcomes. This was further illustrated by considering these confidence splits in comparison to the binary outcomes. Take accuracy in low prevalence searches as an example where hits for the rating-only condition at the 20-40% split were equivalent to the binary condition. However, the 40-60% split more closely reflected false alarms. With binary judgments, the decision is static, while for rating-only decisions, the decision criteria can be altered depending upon the task at hand.

In principle, the ability to impose a more conservative or liberal criterion has clear applied value. In a task such as luggage screening, in which misses can have serious

consequences, the ability to impose a more liberal criterion should increase the number of threats detected. Conversely, for tasks in which failing to identify a target is less costly, a more conservative criterion could be imposed to reduce the amount of time spent investigating possible targets. Such an approach also opens the door for individual criterion to be calibrated specifically for each decision maker, optimising their performance to meet the needs of a certain task. However, it must be acknowledged that in the present experiment, overall accuracy was superior in the binary control condition compared to the rating-only condition.

To conclude, the findings outlined in both Experiment 1 and 2 demonstrate that a rating-only confidence approach can be used to communicate the likelihood of (a) a stimulus consisting predominantly of the target colour, and (b) a luggage x-ray image containing a threat. Such an approach could be of particular use in applied scenarios in which the costs of incorrect decisions can be particularly high (where a more liberal criterion could be set for deleterious miss errors for example), or where the cost of miss errors is low compared to time needed to investigate (where a more conservative criterion would be set). We cannot comment on whether our rating-only approach can reduce biased responding associated with LPE, as our manipulation did not strongly produce the patterns of responding usually associated with the LPE (Wolfe et al., 2005). Thus, further research is needed to draw conclusions regarding the utility of our scale for such a purpose.

References

Baranski, J. V., & Petrusic, W. M. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *The American Journal of Psychology, 110*(3), 543–572. doi: 10.2307/1423410

Bernbach, H. A. (1971). Strength theory and confidence ratings in recall. *Psychological Review, 78*(4), 338–340. doi: 10.1037/h0031034

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*(10), 1208–1214. doi: 10.1177/0956797612441217

Brewer, N., Weber, N., & Guerin, N. (2020). Police lineups of the future? *American Psychologist, 75*(1), 76–91. doi:10.1037/amp0000465

Cleary, A. M., & Greene, R. L. (2000). Recognition without identification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*(4), 1063–1069. doi: 10.1037/0278-7393.26.4.1063

Egan, J. P., Schulman, A. L., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31*(6), 768–773. doi: 10.1121/1.1907783

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506–528. doi: 10.1037/0033-295X.98.4.506

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

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*(3), 217–273. doi: 10.1016/0001-6918(91)90036-Y

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15*(5), 302–306. doi: 10.1111/j.0956-7976.2004.00673.x

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*(1), 55-71. doi:10.1037/a0031602

Peltier, C., & Becker, M. W. (2020). Individual differences predict low prevalence visual search performance and sources of errors: An eye-tracking study. *Journal of Experimental Psychology: Applied, 26*(4), 646–658. doi: 10.1037/xap0000273

Ratcliff, R., McKoon, G., & Tindal, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 763–785. doi: 10.1037/0278-7393.20.4.763

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137*(3), 528–547. doi: 10.1037/a0012712

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347. doi:10.1007/s10979-009-9192-x

Sauer, J. D., Weber, N., & Brewer, N. (2012). Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness. *Psychonomic Bulletin & Review, 19*(3), 490–498. doi: 10.3758/s13423-012-0239-5

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination.

Journal of Mathematical Psychology, 32(2), 135-168. doi:10.1016/0022-2496(88)90043-0

Thomson, K. J., & Goodhew, S. C. (2021). The relationship between the subjective experience of real-world cognitive failures and objective target-detection performance in visual search. *Cognition, 217*, 104914. doi: 10.1016/j.cognition.2021.104914

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 582–600. doi: 10.1037/0278-7393.26.3.582

Vickers, D. (1970) Evidence for an Accumulator Model of Psychophysical Discrimination. *Ergonomics 13*(1), 37-58. doi:10.1080/00140137008931117

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgements. I. properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences, 2*(3), 169-194. doi:10.1023/A:1022371901259

Weber, N., & Varga, M. (2012). Can a modified lineup procedure improve the usefulness of confidence? *Journal of Applied Research in Memory and Cognition, 1*(3), 152-157. doi: 10.1016/j.jarmac.2012.06.007

Weber, N., & Brewer, N. (2004). Confidence–Accuracy Calibration in Absolute and Relative Face Recognition Judgments. *Journal of Experimental Psychology: Applied. 10*(3), 156-172. doi:10.1037/1076-898X.10.3.156

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*(7041), 439–440. doi: 10.1038/435439a

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General, 136*(4), 623–638. doi: 10.1037/0096-3445.136.4.623

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*(3), 611–617. doi: 10.1037/0033-2909.110.3.611

Chapter VI

General Discussion

General Discussion

Recognition-ratings (i.e., confidence judgements made in the absence of a binary decision) have been shown to be a useful tool for indexing the evidence underlying decision making in recognition memory tasks. Not only have such ratings proven to be an accurate method of communicating the likelihood that an image has been seen before (Brewer et al., 2012; 2020; Cleary & Greene, 2000; Sauer et al., 2008; 2012), they have also been used to communicate uncertainty in perceptual tasks (e.g., Ratcliff, McKoon, & Tindall, 1994). However, there has been little direct investigation of the mechanisms that shape such ratings.

The work outlined in this thesis aimed to address the following; First, we investigated which theoretical mechanisms best accounted for patterns of responding when participants used recognition-ratings (as opposed to traditional measures of confidence). We were initially also interested in the best approach for measuring such ratings, but, given very similar patterns of results across rating scale types in early experiments, we instead turned our attention to (a) more targeted tests of what appeared to be the best theoretical mechanism and (b) the application of these ratings to different basic and applied task types.

Theories of Confidence

The manuscript presented in Chapter 3 reports three experiments comparing two dominant models of confidence processing by examining whether recognition-ratings shifted when non-diagnostic information was present during the test phase of a recognition memory task (Experiments 1 and 2), and determining how non-diagnostic information influenced similar ratings in a perceptual discrimination task (Experiment 3). While inferential models of metacognition (e.g., the accessibility model; Koriat 1995; the cue utilization model; Koriat, 1997) would suggest that the presence of non-diagnostic information at test should inflate confidence ratings, psychophysical models of confidence (e.g., Signal Detection Theory; Green & Swets, 1966; or accumulator models coupled with a balance of evidence hypothesis;

Smith & Vickers, 1998: Vickers & Lee, 1998) suggest that confidence should remain stable in the presence of non-diagnostic information, as long as the amount of *diagnostic* evidence does not decrease. One psychophysical theory that deviates from this prediction is Baranski and Petrusic's doubt-scaling model (1998); a lesser known account that posits that confidence is inversely proportionate to the amount of non-diagnostic information present at the time a decision is made.

The first two experiments outlined in Chapter 3 demonstrated that confidence ratings were significantly lower when additional, non-diagnostic information was provided during a recognition task compared to trials in which there is no non-diagnostic information present. In these recognition tasks, participants studied a series of stimuli consisting of both whole and partial images (with the "partial" stimuli consisting of only the top half of the image). Participants were also tested using both whole and partial images, after first being instructed that as long as the image depicted the same target (person, landscape or house), a partial image and a whole image would constitute a correct match. Thus, non-diagnostic information was present in trials in which participants studied a partial image, but were tested with the corresponding whole image. Confidence decreased in the presence of non-diagnostic information (compared to trials in which participants studied a partial image and were tested with the same partial image), and this was consistent for face stimuli (Experiment 1), and houses and landscapes (Experiment 2). The use of houses and landscapes as stimuli allowed us to rule out some competing accounts of the observed patterns of results and increase our confidence in the conclusion that reductions in confidence were associated with increased non-diagnostic information. These findings were further supported by Experiment 3, which showed a decrease in confidence associated with an increase in the proportion of non-diagnostic information present in a perceptual discrimination task, independent of any effects on the amount of diagnostic available at test. The results of the three experiments together

provide strong support for Baranski and Petrusic's (1998) doubt-scaling model of confidence, which until now has received very little consideration in the literature.

Evidence of a Doubt-Scaling Model of Confidence

Having found evidence in support of Baranski and Petrusic's (1998) doubt-scaling model of confidence, we set out to explicitly test predictions drawn from the theory. From what we could gather from the existing literature, the doubt-scaling model has two key hypotheses. First, that confidence is inversely proportion to the amount of non-diagnostic information present when a decision is made. This means that as the amount of non-diagnostic information increases, confidence should decrease accordingly, a claim supported by the results detailed in Chapter 3 (see Experiment 3). Second, as the amount of non-diagnostic information increases, the likelihood of a guess response increases accordingly, resulting in a decrease in accuracy associated with an increase in non-diagnostic information. As Baranski and Petrusic did not provide explicit hypotheses pertaining to the effects of non-diagnostic information on reaction time (RT), we made the following assumptions: As an increase in non-diagnostic information is linked to an increased reliance on guessing, we would expect "guess" responses to be faster as the amount of non-diagnostic information increases. However, when participants are *not* guessing, we would expect their RT to slow as the amount of non-diagnostic information increases. This prediction is based upon Audley's runs model (1960) which suggests that evidence accumulation is a competitive process, meaning that an increase in the accumulation of non-diagnostic evidence would result in a decrease in the accumulation of diagnostic evidence (and therefore slower responses). To test this, we conducted a series of perceptual experiments, in which participants were asked to make judgements pertaining to the primary colour of a dynamic grid consisting of blue, orange and (sometimes) white pixels. Participants would be asked to determine whether the grid was majority blue or majority orange, with the white pixels serving as non-diagnostic

information. Using a perceptual task allowed us to manipulate the proportion of non-diagnostic information present in each trial, the amount of diagnostic information present in each trial (Experiment 1 only), and to put constraints on response times. While Experiment 1 asked participants to simultaneously indicate their decision judgement (orange or blue) and confidence rating (low, moderate, or high), Experiment 2 only asked for a decision judgement. This allowed us to investigate whether the effects of non-diagnostic information on accuracy and RT held when participants were no longer required to consider confidence (Baranski & Petrusic, 2001, 2003). To determine the effects of non-diagnostic information on RT, we used Conditional Accuracy Functions (Thomas, 1974) to plot the accuracy of responses (classified into ranges) as a function of their RT. We observed the following in terms of how non-diagnostic information influenced confidence, accuracy, and RT. First, as the proportion of non-diagnostic information increased, confidence in the decision decreased in a monotonic fashion. Second, as the proportion of non-diagnostic information increased, participants' ability to discriminate between correct and incorrect responses (as indexed by *d'*) decreased accordingly. Third, as the proportion of non-diagnostic information increased, RTs slowed overall. Further, we found that less accurate responses sped up relative to more accurate responses. This was evidenced by the CAFs generally having the steepest slope when there was no non-diagnostic information and flattening as the proportion of non-diagnostic information increased. These fast errors suggest an increased reliance on guessing in trials with higher proportions of non-diagnostic information, whereas the slow accurate responses suggest that the presence of non-diagnostic information does slow the accumulation of diagnostic evidence.

These findings both support and expand upon the predictions made by Baranski and Petrusic's doubt-scaling model (1998). Further, they suggest that existing models of confidence and decision making (e.g., SDT, accumulator models etc.) may benefit from

incorporating a mechanism to account for the effects of non-diagnostic information on the decision-making process.

Measuring Confidence

If we are to truly understand the processes that underpin recognition-ratings, it is important to ensure that we are measuring confidence in a way that allows participants to faithfully represent the underlying construct. While it is common for researchers in the field to use ratings from 0-100% to represent confidence (e.g., Brewer & Wells, 2006; Koriat, 2011; Sauer, Weber & Brewer, 2012 etc.), it has been argued that such probabilistic representations are at odds with how people conceptualise their own uncertainty (Windschitl & Wells, 1996). Further, it has been suggested that reducing the number of response options available could reduce measurement noise (Benjamin, Tullis & Lee, 2013).

We attempted to clarify whether different measurement scales would influence recognition-ratings in Experiments 1 and 2 of Chapter 3. Both studies consisted of a recognition task in which participants were asked to make recognition judgments pertaining to images of faces (Experiment 1), houses or landscapes (Experiment 2), and included a between-subject manipulation of scale-type, with half of the participants responding on a fine-grained numerical scale (0-100% confidence, increasing in 10% increments), and the other half using a coarse-grained verbal scale ("low confidence", "moderate confidence", "high confidence"). We found no tangible effects of scale type on responding in either experiment, suggesting that fine-grained numerical scales and coarse-grained verbal scales are equally reliable measures. This is not completely inconsistent with the existing literature, as some researchers have suggested that participants are able to use both verbal and numerical measures of confidence equally (e.g., Weber, Brewer & Margitich, 2008; Tekin & Roediger, 2017). Based on these findings, we turned our research attention elsewhere. However, this is not to that scale type is unimportant. Obviously, our manipulation

confounded scale granularity (i.e., more vs fewer response options) with anchor types (verbal vs numeric labels). A potentially fruitful avenue for future research would be to disentangle these components, using more coarse-grained numerical scales (e.g., 0%, 25%, 50%, 75% and 100%, etc.) and more fine-grained verbal scales. Further, our research focused on the effects of scale type on the measurement of confidence. However, in applied contexts where confidence is measured, it is typically because expressions of uncertainty may help others to evaluate the reliability of a decision (e.g., triers of fact assessing the likely accuracy of an identification). Thus, future research may wish to investigate the efficacy of ratings-based measures of recognition collected on different scales for informing third-party evaluations of judgement reliability.

Using Ratings in Perceptual Tasks

As discussed previously, a major focus of our work has been on the use of confidence ratings provided without an accompanying decision judgement. While we were initially interested in the theoretical mechanisms that shape such ratings when used in the context of recognition memory, we were also interested in the extent to which they could be used for perceptual tasks. This was first addressed in Chapter 3 (Experiment 3), which showed that participants were able to use a rating-only approach to confidence to convey their confidence in a perceptual discrimination task. Chapter 5 aimed to expand upon this further, first testing the utility of a rating-only approach for a basic perceptual task (identifying the main colour of a dynamic grid, Experiment 1), as well as for a more complex visual search task (identifying whether there was a knife present in an x-ray luggage screening image, Experiment 2). The results showed varying levels of support for the use of a rating-only confidence approach for both types of tasks. Although participants confidence ratings were well calibrated with the likelihood of a dynamic grid consisting predominantly of the target colour (Experiment 1), efficacy for the use of ratings was less clear-cut for the luggage screening task (Experiment

2). While participants demonstrated an ability to use the rating-only scale to discriminate between target present and target absent trials (as evidenced by *d'*), there was evidence of overconfidence. Further, when the data from the rating-only condition was converted to classifications, it proved less accurate than that of the control condition (i.e., a binary decision accompanied by a confidence rating).

Overall, these results suggest that while the benefits of a rating-only approach have been demonstrated in the recognition memory literature, they are less clear-cut for perceptual discrimination tasks. While such an approach proved sound for a basic perceptual task, it proved less useful when employed in a more complex visual-search experiment. More research is needed to comment on whether this discrepancy has occurred due to a problem with our experimental paradigm, or whether it is due to the fundamental differences between the two tasks.

Implications

The findings detailed in this thesis provide strong support for Baranski and Petrusic's (1998) doubt-scaling model of confidence. Specifically, the presence of non-diagnostic information at test reduces confidence ratings in a memory task (Chapter 3), as well as confidence in a perceptual discrimination task (Chapters 3 and 4). Further, accuracy rates decreased as the proportion of non-diagnostic information increased, and RT showed an overall slowing as non-diagnostic information increased. These findings have implications in terms of our understanding of existing theories of confidence in recognition, which, aside from the doubt-scaling model, do not have specific mechanisms to account for the presence of non-diagnostic information. One such example is Hawkins and Heathcote's (2021) timed racing diffusion model. Baranski and Petrusic's model posits that a "guess" response is triggered when the amount of non-diagnostic information accumulated reaches a predetermined threshold before one of the choice accumulators. Hawkins and Heathcote's

model, on the other hand, suggests that the guessing accumulator is driven passively by the passing of time. Thus, if there is not enough diagnostic information to trigger a decision via one of the choice accumulators, it will eventually lead to a guess response. This raises the question of whether the presence of non-diagnostic information may modulate the rate of the guessing accumulator, driving faster guesses when non-diagnsotic information is high, and slower guesses with the presence of non-diagnsotic information is low.

Our research may also have implications for the diagnostic-feature-detection model of eyewitness identification (Wixted & Mickes, 2014; Wixted et al., 2018). The theory was developed to account for higher accuracy rates for police lineups in which the suspects are presented simultaneously compared to when they are presented sequentially, suggesting that providing members of a police lineup simultaneously allows participants to discount "non-diagnostic" features (i.e., features shared between multiple lineup members). This, in turn, allows participants to focus on the features that are unique to each of the faces. This is in direct contrast to our findings, which suggests that participants were unable to simply discount non-diagnostic information, whether it be in a recognition memory task (Chapter 3) or a perceptual discrimination task (Chapters 3 and 4). The deviation between our findings and suggestions put forward by the diagnostic-feature-detection account may be in the differing ways in which "non-diagnsotic information" were operationalized. While our work focused on non-diagnsotic information in terms of additional, irrelevant information that is added at the time of decision, diagnostic-feature accounts define non-diagnostic information as information that was present at the encoding phase, but is no longer considered diagnostic at the decision phase. According to the doubt-scaling model, non-diagnostic information is defined as any information present at the time of decision that is "irrelevant" to the decision. Thus, it may be that common shared features do not fall into this category, as they still provide evidence in favour/against a given response option (though the weight assigned to

this evidence may be lowered based on the number of faces in the lineup that share this trait). Thus, future research may aim to investigate the predictions of the doubt-scaling model utilising different conceptualisations of non-diagnostic information to clarify exactly what meets the definition of "irrelevant" information.

Understanding the effects of non-diagnostic information on confidence and decision making is also important in applied contexts. For example, if an eyewitness witnesses a crime in which the perpetrator was wearing a face-mask, they would likely be tasked with identifying the unmasked suspect from a lineup. In this scenario, the additional facial information would be non-diagnostic, as it was covered at the time of encoding. The results detailed in this thesis suggest that the witness would not only be less confident in their identification (Chapter 3), but likely less accurate as well (Chapter 4). These findings may well explain the results of Manley et al. (2019), who found that participants were less confident identifying a studied-face that had been wearing a face-mask at encoding, but not at test (c.f. a masked face at test).

While the findings detailed in Chapter 5 do not make a compelling case for the use of a rating-only approach to visual search, the two experiments do highlight the importance of further research on the topic to clarify whether the utility of such an approach can extend beyond basic perceptual tasks.

References

Angell, J. R. (1907). The province of functional psychology. *Psychological Review, 14*, 61–91. doi:10.1037/h0070817

Audley, R. J. (1960). A stochastic model for individual choice behaviour. *Psychological Review, 67*(1), 1-15. doi:10.1037/h0046438

Baranski, J. V., & Petrusic, W. M. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *The American Journal of Psychology, 110*(3), 543–572. doi: 10.2307/1423410

Baranski, J. V., & Petrusik, W. M. (1998). Probing the Locus of Confidence Judgments: Experiments on the Time to Determine Confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 929-945. doi: 10.1037//0096-1523.24.3.929

Baranski, J. V., & Petrusik, W. M. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology/Revue Canadienne de psychologie experimentale, 55(*3), 195-206. doi:10.1037/h0087366

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1601-1608. doi: 10.1037/a0031849

Bernbach, H. A. (1971). Strength theory and confidence ratings in recall. *Psychological Review, 78*(4), 338–340. doi:10.1037/h0031034

Bradfield, A.L., Wells, G.L., & Olson, E.A. (2002). The damaging effect of

confirming feedback on the relation between eyewitness certainty and identification accuracy. *Journal of Applied Psychology, 87*, 112-120. doi:10.1037/0021-9010.87.1.112

Brewer, N., Burke, A. (2002). Effects of Testimonial Inconsistencies and Eyewitness Confidence on Mock-Juror Judgments. *Law Human Behaviour 26*, 353–364 (2002). doi:10.1023/A:1015380522722

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*, 1208–1214. doi:10 \.1177/0956797612441217

Brewer, N., Weber, N., & Guerin, N. (2020). Police lineups of the future? *American Psychologist, 75*(1), 76–91. doi:10.1037/amp0000465

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, functional size and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11-30. doi: 10.1037/1076-898X.12.1.11

Brewer, N., & Williams, K. D. (2005). *Psychology and Law, An Empirical Perspective*. New York: Guilford Press.

Bruer, K. C., Fitzgerald, R. J., Price, H. L., & Sauer, J. D. (2017). How sure are you that this is the man you saw? Child witnesses can use confidence judgments to identify a target. *Law and Human Behavior, 41*(6), 541–555. doi:10.1037/lhb0000260

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153–178. doi:10.1016/j.cogpsych.2007.12.002

Busey, T. A., Tunnicliff, J., Loftus G. R., & and Loftus, E., F. (2000). Accounts of

the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26-48. doi:10.3758/BF03210724

Carlson, C. A., Young, D. F., Weatherford, D. R., Carlson, M. A., Bednarz, J. E. & Jones, A. R. (2016). The Influence of Perpetrator Exposure Time and Weapon Presence/Timing on Eyewitness Confidence and Accuracy. *Applied Cognitive Psychology. 30*(6), 898-910. doi: 10.1002/acp.3275

Core Team, R. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Cleary, A. M., & Greene, R. L. (2000). Recognition without identification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*(4), 1063–1069. doi: 10.1037/0278-7393.26.4.1063

Cutler, B.L., Penrod, S.D. & Dexter, H.R. (199). Juror sensitivity to eyewitness identification evidence. *Law Hum Behav* **14,** 185–191 (1990). doi:10.1007/BF01062972

Deffenbacher, K.A., & Loftus, E.F. (1982). Do jurors share a common understanding concerning eyewitness behavior? *Law and Human Behavior 6*(1), 15–30 (1982). doi:10.1007/BF01049310

Diederich, P., B., French, J., W., & Sydell, T. C. (1961). Factors in judgement of writing ability. *ETS Research Bulletin Series, 1961*(1), i-93. doi: 10.1002/j.2333-8504.1961.tb00286.x

Egan, J. P., Schulman, A. L., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31*(6), 768–773. doi:10.1121/1.1907783

Elliott, D., Strickland, L., Loft, S. & Heathcote, A. (accepted 10/November/2021). Integrated responding improves prospective memory accuracy. *Psychonomic Bulletin & Review.*

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45*(1), 1-18. doi: https://doi.org/10.1016/0749-5978(90)90002-Q

Festinger, L. (1943). Studies in decision: I. Decision-time, relative frequency of judgement and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32*(4), 291-306. doi:10.1037/h0056685

Filliter, J. H., Glover, J. M., McMullen, P. A., Salmon, J. P., & Johnson, S. A. (2016). The DalHouses: 100 new photographs of houses with ratings of typicality, familiarity, and degree of similarity to faces. *Behavior Research Methods, 48*, 178-183. doi: 10.3758/s13428-015-0561-8

Fox J, Weisberg S (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Franks, J. J., Bilbrey, C. W., Lien, K. G., & McNamara, T. P. (2000). Transfer-appropriate processing (TAP). *Memory & Cognition, 28*(7), 1140-1151.

Gigerenzer, G., Hoffrage, U., & Kleinboelting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528. doi: 10.1037/0033-295X.98.4.506

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

Hanczakowski, M., Pasek, T., Zawadzka, K., & Mazzoni, G. (2013). Cue familiarity and 'don't know' responding in episodic memory tasks. *Journal of Memory and Language, 69*(3), 368-383. doi:10.1016/j.jml.2013.04.005

Handmer, J., & Proudley, B. (2007). Communicating uncertainty via probabilities: the case of weather forecasts. *Environmental Hazards, 7*(2), 79-87. doi:10.1016/j.envhaz.2007.05.002.

Hawkins, G.E., & Heathcote, A. (2021). Racing against the clock: Evidence-based vs. time-based decisions. *Psychological Review, 128*, 222-263.

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

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging, 5*(2), 215–227. doi:10.1037/0882-7974.5.2.215

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review, 18*(3), 186–201. doi:10.1037/h0074579

Horry, R., & Brewer, N. (2016). How target–lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145*(12), 1615–1634. doi:10.1037/xge0000227

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian Origins of Uncertainty in Judgement: A Sampling Model of Confidence in Sensory Discrimination. *Psychological Review, 104*(2), 344-366. doi: 10.1037/0033-295X.104.2.344

Juslin, P., Winman, A., & Olson, H. (2000). Naïve Empiricism and Dogmatism in Confidence Research: Critical Examination of the Hard-Easy Effect. *Psychological Review, 107*(2), 384-396. doi: 10.1037//0033-295X.107.2.384.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*(3), 217–273. doi: 10.1016/0001-6918(91)90036-Y

Klein, S. P., & Hart, F. M., (1968). Chance and systematic factors affecting essay grades. *Journal of Educational Measurement, 5*(3), 197-206.

Kohl, T. A., Sauer, J. D., Palmer, M., Brooks, J., & Heathcote, A., (2022) *The Effects of Non-Diagnostic Information on Confidence and Decision Making*. Manuscript under

review.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological review. 100*(4), 609-639. doi: 0.2307/1422236

Koriat, A. (1977). Monitoring One's Own Knowledge During Study: A Cue-Utilization Approach to Judgments of Learning. *Journal of Experimental Psychology: General, 126*(4), 349-370. doi: 10.1037/0096-3445.126.4.349

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General, 140*(1), 117–139. doi:0.1037/a0022171

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490–517. doi:10.1037/0033-295X.103.3.490

Kuznetsova A., Brockhoff P.B., Christensen R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1-26, doi:10.18637/jss.v082.i13

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological Science, 25*(9), 1663-1673. doi:10.1177/0956797614541991

Macmillan, N. A., & Creelman, C. D. (1991). Detection theory: A user's guide. New York: Cambridge University Press.

Martinez, A. M., & Benavente, R. (1998). *The AR face database* (CVC Technical Report No. 24). Barcelona, Spain: Universitat Autonoma de Barcelona, Computer Vision Center.

Manley, K.D., Chan, J. C. K., & Wells, G. L. (2019). Do masked-face lineups facilitate eyewitness identification of a masked individual? *Journal of Experimental Psychology: Applied, 25*(3), 396-409. doi:10.1037/xap0000195.

Mansour, J. K. (2020). The confidence-accuracy relationship using scale versus other methods of assessing confidence. *Journal of Applied Research in Memory and Cognition, 9*(2), 215-231. doi:10.1016/j.jarmac.2020.01.003

MATLAB and Statistics Toolbox Release R2016b, The MathWorks, Inc., Natick, Massachusetts, United States.

Maurer, D., Le Grand R., & Mondloch, C. J. (2002). The many faces of configural processing. *TRENDS in Cognitive Sciences, 6*(6), 255-260. doi: 10.1016/S1364-6613(02)01903-4

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15*(5), 302–306. doi: 10.1111/j.0956-7976.2004.00673.x

Mickes, L. (2015). Receiver Operating Characteristic Analysis and Confidence-Accuracy Characteristic Analysis in Investigations of System Variables and Estimator Variables that Affect Eyewitness Memory. *Journal of Applied Research in Memory and Cognition.* doi: 10.1016/j.jarmac.2015.01.003

Miletić, S., Boag, R.J., Trutti, A. C., Stevenson, N., Forstmann, B.U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks, *eLife*. doi:10.7554/eLife.63055

Morris, D. C., Bransford J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519-533. doi: 10.1016/S0022-5371(77)80016-9

National Academy of Sciences (NAS) (2014). Identifying the Culprit. Assessing eyewitness identification. Washington DC: The National Academic Press.

Neil v. Biggers, 409 U. S. 188 (1972).

Nold, E. W., & Friedman, S. W. (1977). An analysis of readers' responses to essays.

Research in the Teaching of English, 11(2), 164-174.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*(1), 55-71. doi:10.1037/a0031602

Peltier, C., & Becker, M. W. (2020). Individual differences predict low prevalence visual search performance and sources of errors: An eye-tracking study. *Journal of Experimental Psychology: Applied, 26*(4), 646–658. doi: 10.1037/xap0000273

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. Journal of Experimental Psychology: *Human Perception and Performance, 18*(4), 962-986. doi:10.1037/0096-1523.18.4.962

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgements. *Psychonomic Bulletin & Review, 10,* 177-183. doi:10.3758/BF03196482

Potter, R., & Brewer, N. (1999). Perceptions of witness behaviour-accuracy relationships held by police, lawyers and mock-jurors. *Psychiatry, Psychology and Law, 1*(1), 97-103. doi: 10.1080/13218719909524952

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., McKoon, G., & Tindal, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 763–785. doi: 10.1037/0278-7393.20.4.763

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face

processing. *Psychological Bulletin, 140*(5), 1281–1302. doi:10.1037/a0037004

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 185-208). Chichester: Wiley Blackwell.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple Confidence Estimates as Indices of Eyewitness Memory. *Journal of Experimental Psychology: General, 137*(3), 528-547. doi: 10.1037/a0012712

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347. doi:10.1007/s10979-009-9192-x

Sauer, J. D., Weber, N., & Brewer, N. (2012). Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness. *Psychonomic Bulletin & Review, 19*(3), 490-498. doi: 10.3758/s13423-012-0239-5

Sauer, J. D., Palmer, M. A., & Brewer, N. (2017). Mock-juror evaluations of traditional and ratings-based eyewitness identification evidence. *Law and Human Behavior, 41*(4), 375–384. doi:10.1037/lhb0000235

Sauerland, M., Raymaekers, L. H. C., Otgaar, H., Memon, A, Waltjen, T. T., Nivo, M … Smeets, T. (2016). Stress, stress-induced cortisol responses, and eyewitness identification performance. *Behavioral Sciences and the Law, 34*(4), 475-594. doi: 10.1002/bsl.2249

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. Journal of Mathematical Psychology, 32(2), 135-168. doi:10.1016/0022-2496(88)90043-0

Tanaka, J. W., & Gordon, I. (2011). Features, configuration, and holistic face processing. In G. Rhodes & J. Haxby (Eds.), *Oxford Handbook of Face Perception.* (pp. 177-195). Oxford University Press.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*:49, 1-13, doi: 10.1186/s41235-017-0086-z

Teodorescu, A. R., Moran, R., & Usher, M. (2016). Absolutely relative or relatively absolute: violations of value invariance in human decision making. *Psychonomic Bulletin & Review, 23*(1), 22–38. doi:10.3758/s13423-015-0858-8

Technical Working Group: Eyewitness Evidence (1999*). Eyewitness evidence: A guide for law enforcement.* US Department of Justice, Office of Justice Programs, National Institute of Justice. NCJ 178240.

Thomson, K. J., & Goodhew, S. C. (2021). The relationship between the subjective experience of real-world cognitive failures and objective target-detection performance in visual search. *Cognition, 217,* 104914. doi: 10.1016/j.cognition.2021.104914

Tillman, G., Zandt, T. V., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review, 27*(5), 911–936. doi:10.3758/s13423-020-01719-6

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

Usher, M., & McClelland, L. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 109*(3), 550-592. doi:10.1037/0033-295X.108.3.550

Van Ravenzwaaij, D., Brown, S. D., Marley, A. J., & Heathcote, A. (2020). Accumulating

advantages: A new conceptualization of rapid multiple choice. *Psychological Review, 127*, 186–215. doi:10.1037/rev0000166

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 582-600. doi: 10.1037/0278-7393.26.3.582

Van Zandt, T. V., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review, 7*, 208–256. doi: 10.3758/BF03212980

Vickers, D. (1970). Evidence for an Accumulator Model of Psychophysical Discrimination. *Ergonomics 13*(1), 37-58. doi:10.1080/00140137008931117

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgements. I. properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences, 2*(3), 169-194. doi:10.1023/A:1022371901259

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. Acta Psychologica, 50(2), 179-197. doi:10.1016/0001-6918(82)90006-3

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numeric terms. *Bulletin of the Psychonomic Society, 31*(2), 135-138. doi: https://doi.org/10.3758/BF03334162

Weber, N., & Varga, M. (2012). Can a modified lineup procedure improve the usefulness of confidence? *Journal of Applied Research in Memory and Cognition, 1*(3), 152-157. doi: 10.1016/j.jarmac.2012.06.007

Weber, N., & Brewer, N. (2004). Confidence–Accuracy Calibration in Absolute and Relative Face Recognition Judgments. *Journal of Experimental Psychology: Applied. 10*(3), 156-172. doi:10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied, 14*(1), 50–60. doi:10.1037/1076-898X.14.1.50

Weber, N., Brewer, N., & Margitich, S. (2008). The confidence-accuracy relation in eyewitness identification: Effects of verbal versus numeric confidence scales. In K. H. Kiefer (Ed.), *Applied psychology research trends* (pp. 103-118). Hauppauge, NY: Nova Science Publishers.

Wickelgren, A. W., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3*(2), 316-347. doi: 10.1016/0022-2496(66)90018-6

Windshitl, P. D., & Wells, G. L. (1996). Measuring Psychological Uncertainty: Verbal Versus Numeric Methods. *Journal of Experimental Psychology: Applied, 2*(4), 343-364. doi: 10.1037/1076-898X.2.4.343

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152-176. doi: 10.1037/0033-295X.114.1.152

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*, 262–276. doi:10.1037/a0035940

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81-114. doi:10.1016/j.cogpsych.2018.06.001

Wixted, J.T., Wells, G.,L. (2017). The relationship between eyewitness confidence and identification accuracy: a new synthesis. Psychological Science in the *Public Interest, 18*(1), 10-65. doi:10.1177/1529100616686966

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*(7041), 439–440. doi: 10.1038/435439a

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General, 136*(4), 623–638. doi: 10.1037/0096-3445.136.4.623

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*(3), 611–617. doi: 10.1037/0033-2909.110.3.611