Exploring the Basis of Confidence in Recognition – a Psychophysical Approach vs. an Inferential Approach to Ecphoric Confidence Ratings.

Amelia Kohl.

Word Count: 9819

A report submitted as a partial requirement for the degree of Bachelor of Arts with Honours in Psychology at the University of Tasmania, 2017.

Statement of Sources

I declare that this report is my own original work and that contributions of others have been duly acknowledged: Amelia Kohl, 19/10/2017.

Acknowledgements

I would like to thank Dr. Jim Sauer for his supervision, and for sharing his knowledge and enthusiasm for science with me throughout the year.

I would also like to thank Dr. Nathan Weber (of Flinders University) for developing the software that we used to run our study.

Table of Contents

List of Tables and Figures

Abstract

We investigated whether ecphoric confidence ratings were best accounted for by psychophysical or inferential models of metacognition. 60 participants (43 female; aged 16 to 75 years), undertook a facial recognition task. Participants saw a mix of full and partial faces at both test and study (partial faces displayed the top half of the face), and provided ecphoric confidence ratings (indexing recognition without a yes/no decision) for each face at test on a coarse-grained verbal or fine-grained probabilistic scale. Inferential models of metacognition propose that additional information at test, regardless of its diagnosticity, increases confidence. Therefore, we would expect confidence to be higher in trials where participants viewed a partial face at study (TS) followed by the corresponding full face at test (FT), than when they viewed a partial face followed by the corresponding partial face (TT). Psychophysical models, in contrast, propose that confidence indexes stimulus discriminability, and should be unaffected by additional non-diagnostic information, TS/FT = TS/TT. The doubt-scaling model argues that non-diagnostic information should decrease confidence, and therefore TS/TT > TT/FT. Linear mixed effects models supported the doubt-scaling account. However, these results must be interpreted with caution given potential limitations. Scale type did not affect results.

In a variety of recognition domains, effects on an individual's decision criteria can influence decision-making independent of the quality of the individual's memory, stimulus discriminability, or the strength of the individual's recognition experience. In applied recognition domains, this can contribute to costly errors. For example, in the eyewitness identification domain, factors resulting in a more lenient criterion (e.g., instructions that state or imply that the culprit is present in the lineup) can result in a false identification when the culprit is absent and contribute to wrongful convictions (Carlson, Gronlund & Clark, 2008; Malpass, 1981). Alternatively, factors that induce a stricter response criterion can lead the witness to reject a lineup when the culprit is present and hinder the prosecution of the guilty party. Thus, researchers have recently advocated for an alternative method of collecting eyewitness identification in an effort to provide a more informative assessment of the witness's memory, and the degree of match between the police's suspect, and the witness's memory of the culprit.

Eschewing a categorical identification response (where the witness either picks a lineup member or rejects the lineup) in favour of a procedure where witnesses indicate, for each lineup member, their confidence that this person was the culprit may provide a richer source of information when evaluating the quality of a witness's memory, and better discriminate a target among foils in a lineup (Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012). This sort of confidence rating is made in the absence of a categorical recognition judgment is known as an ecphoric confidence, whilst confidence ratings made alongside a binary yes/no identification when presented with a stimuli is known as retrospective confidence (Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012).

If ecphoric confidence ratings are to be used in the field of eyewitness identification, whether it be in a research setting or within the legal system, it is important to understand the theoretical mechanisms underlying these judgments. There are currently two main schools of theory that are used to explain the basis of metacognitive judgments (of which confidence in recognition is an example): psychophysical models (e.g., those derived from Signal Detection Theory; Green & Swets, 1966; or accumulator models coupled with a balance of evidence hypothesis; Van Zandt, 2000) and inferential models (e.g., the cue utilization model; Koriat, 1997). Despite evidence demonstrating that inferential models are at play when forming the basis of other metacognitive judgments (Koriat, 1993; Koriat, 1997), most of the previous research into retrospective confidence in recognition has assumed a psychophysical basis (Baranksi & Petrusic,1998; Vickers, 1979). Previous literature in regards to ecphoric confidence has typically noted its conceptual similarity to retrospective confidence and, thus, has relied on psychophysical models as a basis for understanding ecphoric ratings (Sauer, Weber & Brewer, 2012). An understanding of the theoretical basis of ecphoric confidence is paramount if we are to determine whether confidence truly is a reliable indices of recognition, and furthermore it is necessary if we are to establishing the boundary conditions for these indices.

In the current study, we used a facial recognition task to test the suitability of psychophysical or inferential accounts of confidence in recognition. Specifically, we manipulated the amount of non-diagnostic information provided at study and test phases, and measured the effect of this additional, non-diagnostic information on participants' ecphoric confidence ratings. Our key interest was investigating how confidence ratings differed in trials in which participants were shown the same

stimulus at study and test with no extra non-diagnostic information provided, compared to when they were shown the same face at study and test, but with additional non-diagnostic information provided at the test phase. A secondary interest was in how any effects might vary as a function of the scale used to assess ecphoric confidence.

Metacognition and confidence

Metacognition refers to a person's understanding and appraisal of the processes that underpin their own cognition, and is characterized by introspection and monitoring of one's own memories and learning (Eisenacher & Zinc, 2017; Flavel 1979). Although metacognition is involved in a variety of essential cognitive processes, including memory, it is important to recognize that metacognition and memory are separate concepts, and to therefore distinguish between the two (Metcalfe & Dunlosky, 2008). Thus, if memory is indexed by the ability to recall or recognize studied items, metacognition could be indexed by an individual's ability to accurately predict future recall, or effectively assess the likely accuracy of a recognition decision.

Research into metacognition has largely focused upon two somewhat distinct domains: the first being the predictive value of metacognitive judgments, and the second being the use of metacognitive judgments to index the accuracy of a decision post-hoc. Judgments of Learning (JOLs) and Feelings of Knowing (FOKs) are both examples of metacognitive judgments used to predict performance, whereas retrospective confidence ratings are an example of a metacognitive judgment used to index the accuracy of a decision.

People's confidence in their memory, and whether or not confidence ratings can be used to index memory accuracy, is a field of research that has been of particular

interest in the literature, with previous research typically focusing upon the relationship between confidence and accuracy for recognition judgments (Brewer & Wells, 2006; Gigerenzer, Hoffrage & Kleinboelting,1991; Juslin, Winman & Olsson, 2000; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer, Brewer, & Weber, 2008; Sauer, Brewer, Zweck, & Weber, 2010). In these cases, confidence follows a categorical response and is intended to index the likely accuracy of that response. For the purpose of the current study, however, we investigated ecphoric confidence ratings. Unlike retrospective or typical confidence ratings, which involve the participant providing an indication of whether the stimulus was old or new alongside their confidence rating, an ecphoric confidence rating is not accompanied by any categorical recognition judgment (Sauer, Weber & Brewer, 2012). More recently, researchers have been interested in the confidence accuracy relationship in applied settings, predominantly in the field of eye-witness identification (Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012).

Previous studies have determined that ecphoric confidence ratings provide diagnostic information even without an accompanying binary decision, and that ecphoric confidence may actually be of greater diagnostic value than a binary decision (Brewer, Weber, Wootton & Lindsay, 2012; Sauer, Brewer & Weber, 2008). Expanding upon this, it has been found that factors that have been found to impair discriminability, such as retention interval and distinctiveness, have less of an effect on ecphoric confidence ratings than they do binary yes/no responses (Sauer, Weber & Brewer, 2012).

Weber and Varga (2012) used a modified facial recognition lineup procedure in which participants were asked to identify the member of the lineup that best matched their memory, give a rating as to how confident they were that that lineup member

was the target that they were supposed to be identifying (i.e., an ecphoric confidence rating), and finally give a binary yes/no decision as to whether the person they picked was in fact the target. They found that ecphoric confidence ratings were more informative than the binary responses in terms of determining whether the test stimulus had been viewed a study. This is important for two reasons. First, it reiterates that ecphoric confidence ratings provide valuable diagnostic information. Second, the increased diagnosticity of confidence ratings compared to the binary identification suggests that these two decision-making processes may not be based upon the same underlying mechanisms.

Finding reliable indices of recognition and establishing the boundary conditions for these indices is important in a variety of domains, with an example being the criminal justice system. 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. Whilst the endorsement of the U.S. Supreme Court may highlight an applied use of confidence ratings alongside (or instead of) recognition ratings, it is also important to recognize that there is a theoretical basis for expecting a relationship between confidence and accuracy. Both psychophysical and inferential accounts of confidence propose that memory strength/discriminability plays a role in formulating confidence levels (albeit to differing extents), and these theoretical accounts are supported by aforementioned studies that have found that confidence ratings can provide diagnostic information in the absence of a binary yes or no decision (Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012; Weber & Varga, 2012).

The argument for using ecphoric confidence ratings in an applied setting rather than a binary yes/no identification is two-fold. Firstly, it may reduce the influence of

non-memorial factors on individual's decision criteria, and attenuate the contribution of these influences to identification, compared to binary identification decisions (Sauer & Brewer, 2015). Secondly, it provides more information for those making a decision in a legal setting (jury members, judges, police etc.) to base their decision upon (Sauer & Brewer, 2015), because as mentioned before, ecphoric confidence ratings provide a richer source of information than a binary decision (Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012).

Whilst previous research indicates the potential value of using confidence ratings as an alternative to the more traditional methods of testing eyewitness recognition memory (Brewer, Weber, Wootton & Lindsay, 2012; Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012; Weber & Varga, 2012), it is important to understand the underlying mechanisms that form said confidence ratings. Specifically, to determine the underlying theoretical basis of confidence in recognition. This deeper understanding will allow for a greater comprehension of the conditions under which confidence ratings will or will not provide accurate, reliable information.

Psychophysical models

Psychophysical models are a class of theories that all propose that confidence indexes stimulus intensity or discriminability (Vickers, 1979; Wixted, 2007). In regards to recognition tasks, this would be how strongly the test stimulus matched the participant's memory of the stimulus during the encoding phase. Psychophysical models of confidence are based upon the direct-access hypothesis: that a person's recognition is reliant on the strength of the memory and the degree of match (or *ecphory*) between a stimulus presented at test and the participant's memory of a studied stimulus (King, Zechmeister & Shaughnessy, 1980; Koriat, 1997). Most of

the previous research into the foundation of confidence in recognition has assumed a psychophysical basis (Baranksi & Petrusic,1998; Vickers, 1979).

A simplistic lab-based example of how psychophysical models propose that confidence is formulated is as follows. First, a participant is presented with a stimulus to study. Later, they are presented with another stimulus and asked to determine whether this test stimulus was previously studied. When making this judgment, the participant compares the current stimulus in front of them with their memory of the stimulus that they studied earlier. This comparison would then generate some degree of match/sense of familiarity, which then forms the basis of both the participant's recognition judgment, as well as their confidence in said recognition judgment. This basic paradigm has often been used to explain the empirical relationship between confidence and accuracy for recognition judgments (Van Zandt, 2000; Vickers, 1979).

Although it is not out intent to distinguish between different theories that fall under the umbrella of psychophysical models, outlining the underlying mechanisms proposed by two of the most prominent psychophysical models would help elicit a deeper understanding of how the degree of ecphoric similarity/familiarity shapes confidence judgments. These two models are signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 1991; Wixted, 2007) and accumulator models combined with a balance of evidence mechanism (Vickers, 1979). SDT proposes that in a recognition memory test, once the strength of a memory signal for a given stimulus has surpassed the decision criteria, the stimulus will be classed as "familiar" or "old". If the strength of the memory signals fails to reach this criteria, the stimulus will be classed as "unfamiliar" or "new" (Wixted, 2007). Confidence is then based upon the difference between strength of the memory signal and the decision criteria

(Baranksi & Petrusic, 1998). Thus, as a signal gets stronger the extent by which it exceeds the relevant criterion increases and so does confidence.

Vickers (1979) proposed that confidence levels reflect the balance of evidence accumulated for each of two (or more) possible outcomes. For example, in a basic facial recognition task, there are generally only two possible answers: response A: the face was seen before (i.e., it was "old") or response B: it was not (i.e., it was "new"). As the degree of match between the test stimulus and the participant's memory for the studied stimulus increases, evidence favoring response A (old) increases. This increases the discrepancy between response A (old) and response B (new). Confidence then increases along with this discrepancy, with confidence reflecting the balance of evidence in favor of response A. In a situation where the balance of evidence favored response B over A, confidence would be decrease as per the amount of information in favor of response B. In a slightly more complex scenario, such as if a stimuli was partly obscured at initial viewing and therefore is not a perfect match at current viewing, there will be less evidence for the affirmative response (old) and/or more evidence for the negative response (new). Thus, the discrepancy in the balance of evidence for the competing response option, and therefore confidence, will decrease.

Importantly, both SDT and accumulator models put forward the idea that confidence in recognition is based upon the strength/discriminability of a stimulus against a criterion (Green & Swets, 1966; Vickers, 1979). If it is true that confidence ratings are based upon psychophysical processes, the current study should find that in trials where there is a correct match, participants' confidence ratings should remain the same regardless of the amount of non-diagnostic information present at test. This is because the presence of non-diagnostic information does not interfere

with the degree of match between the remembered stimuli and the current stimuli, nor does it interfere with strength of the memory itself. One exception to this predicted pattern of results that could still provide support for a psychophysical model, however, could be justified using Baranksi and Petrusic's (1998) doubt scaling model.

The doubt-scaling model (Baranksi and Petrusic, 1998) proposes that confidence levels are inversely based upon the amount of non-diagnostic information accrued during the decision-making process. That is, the more non-diagnostic information present, the lower confidence ratings will be for the selected response. The doubt-scaling model differs from the SDT-based criterion hypothesis and the balance of evidence models by focusing upon the amount of non-diagnostic information rather than the degree of match; though it is still similar in that it proposes that confidence is based upon strength of memory and/or stimulus discriminability.

In regards to the current study, the doubt-scaling model would predict quite a different pattern of results to the other psychophysical models. As the doubt-scaling model proposes that confidence is formulated based upon the amount of non-diagnostic information present during the decision making process, it would therefore predict that the presence of additional non-diagnostic information at test would decrease confidence ratings compared to when this non-diagnostic information was not present. Other psychophysical models, however, suggest that confidence increases as the degree of match increases, with non-diagnostic information playing no role in determining confidence ratings[1]. Although the doubt-scaling model provides an interesting and empirically testable prediction that is quite

[1] Assuming that providing additional non-diagnostic information at test does not undermine the salience/utility of the diagnostic information.

unique from other psychophysical models, it is worth noting that this theory has not been elaborated upon in the literature since it was introduced by Baranski and Petrusik in 1998.

Inferential models

Unlike psychophysical models of confidence, inferential models propose that metacognitive judgments are shaped by inferential processes rather than being primarily determined by properties of the stimulus (Koriat, 1993; Koriat, 1997). These inferential cues are not related just to the stimulus, but can also include participant's heuristics relating to perceptions of the study and test conditions (Koriat, 1997). Empirical support for an inferential approach has emerged in other domains of metacognition, such as such as JOLs (e.g., Koriat, 1993) and FOK (e.g., Koriat, 1997).

Early accounts of metacognition, including JOLs, assumed a direct-access mechanism of confidence (Hart, 1965). However, in a study that aimed to determine whether JOLs had a direct-access (i.e. psychophysical) or cue-utilization (i.e., inferential) basis, Koriat (1997) found evidence for an inferential basis. Whilst the direct-access approach suggested that JOLs were based solely on memory strength, the cue-utilisation approach suggested that JOLs were based upon inferential processes. Koriat identified three types of cues that form the basis of the cue-utilization approach: intrinsic, extrinsic, and mnemonic. Intrinsic cues relate directly to the item studied and impact how easy/difficult the participant will find 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. These include cues such as the ease of retrieval (e.g., retrieval fluency, or the amount of information retrieved in response to a test cue – regardless of whether the information retrieved is correct or not) and ease of processing the item.

Although Koriat's (1997) paper integrated four different experiments, we consider only the first of these here as it is a good example of a typical JOL task. In the study phase, participants were presented with 50 word pairs, ranging from easier associations (e.g., cow-milk) to harder associations (e.g., citizen-fox). After participants finished studying each pair they were asked to indicate, from 0-100%, how likely they thought they were to recall the second word of the pair when presented with the first in the test phase (i.e., provide a JOL). In the test phase, participants were presented with the first word of the pair and given 10 seconds to respond with the paired word. Each participant undertook two study phases and two test phases, with participants in the experimental condition seeing the same set of words for both study phases (as to examine the effects of re-studying items), and participants in the control group receiving two different lists. This first experiment was particularly interested in the effect of studying the same set of words twice, as well as general transfer of learning, both of which are examples of extrinsic factors.

What was found was that viewing the word pair more than once (i.e., those in the experimental group) had a larger effect on recall than it did JOLs. Inversely, the judged difficulty of an item effected JOLs more strongly than it did recall. These findings provided evidence for a cue-utilization approach to JOLs, as a direct-access model would propose that both the amount of times studied and judged difficulty

should have had the same effect on JOLs as they did accuracy. Experiments 2-4 each produced similar evidence in support for a cue-utilization approach.

In summary, Koriat (1997) demonstrated that intrinsic and extrinsic factors, as well as mnemonic cues, all contributed to the formulation of JOLs. This provided evidence in support of the cue-utilization approach – that JOLs are not based purely on memory strength. An inferential approach for JOLs has also been supported by Hertzog, Dixon and Hultsch (1990), who showed that JOLs can be influenced by a person's beliefs about their own abilities in regards to memory (i.e., an inference distinct from the properties of the target stimuli).

Another paper by Koriat (1993), this time focusing on the theoretical basis of FOKs, also provides evidence for an inferential model of metacognition. In the first experiment of three, participants viewed 40 tetragrams (with tetragrams being strings of consonants; e.g., RDFK). Each of the tetragrams were 4 letters long and randomly generated[2]. Each participant completed 4 stroop items[3] before being presented with a tetragram for 1000ms, followed by another 18s stroop task (the stroop tasks were used to minimize any interference from the previous tetragram). Participants were then asked to recall the target, and were told that they would gain 1 point per correct letter, but would gain no points at all if they reported a single incorrect letter (this was purely to encourage a good response rate). They then made a FOK judgment regarding the tetragram, after which they completed a recognition test for the item.

---

[2] Although it was guaranteed that each consonant would appear in a minimum of 7 and maximum of 9 of the tetragrams within one string.

[3] In a stroop task, participants are presented with the name of a colour written in that colour ink (e.g. the colour red written in red), and are required to report out-loud the colour in question (Stroop, 1935).

The recognition test involved participants viewing 8 tetragrams, one of which was the correct match for what they had seen before, the other 7 of which were designed to be of varying similarity to the target, and selecting the one that they believed was a correct match to what they had studied. This was repeated 40 times.

In scenarios in which participants failed to recall the entire stimulus, but were able to recall a few letters from said stimulus, future recognition was predicted by the accuracy of the partial information (i.e., individual letters) recalled. FOKs, however, were predicted by the amount of partial information (i.e., number of individual letters) recalled, irrespective of accuracy. Thus, FOKs are not based on direct-access, but rather on inferential cues associated with attempted retrieval (Koriat, 1993).

JOLs, FOKs and ecphoric confidence are all examples of metacognitive processes. Thus, these findings suggest the merit of exploring an inferential account of confidence in recognition. In addition to the conceptual similarity of these metacognitive processes providing justification for exploring an inferential basis of confidence, an examination of Thurstonian and Brunswikian accounts of uncertainty also provides support for such research.

Juslin and Olsson (1997) put forward two accounts of uncertainty: Thurstonian and Brunswikian. The Thurstonian approach characterises an internal form of uncertainty, and is centred around the idea that uncertainty is caused by noise in a persons information processing system rather than any issues relating to the stimuli itself. The Brunswikian approach, on the other hand, characterises an external form of uncertainty, and assumes that said uncertainty 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.

Given the conceptual similarity between the ideas, it seems as though psychophysical models of confidence would fit well within a Thurstonian account of uncertainty (as they are both based upon internalised processes, e.g., stimuli intensity etc.), where as inferential models of confidence would fit within the Brunswikian account (as they are both cue based). In their research, Juslin and Olsson (1997) suggested that the Thurstonian approach was 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.

It is difficult to class a recognition task such as ours, in which participants are asked to make a confidence judgement by comparing a stimuli in front of them to a memory, as a task that would clearly favour either of the Brunswikian or Thurstonian approaches. In the literature, researchers have tended to apply the Thurstonian approach to tasks such as ours, because models of confidence in recognition tend to draw on theoretical frameworks originally designed for perceptual discrimination tasks (e.g., SDT and Accumulator models). However, given that memory represents an imperfect source of information (cf. having two stimuli available for direct comparison), it could be that Brunswikian accounts (which emphasise the role of inference in assessments of uncertainty) have something to offer. This lack of clear distinction brings into question whether the current assumption in the literature that confidence in recognition is based upon psychophysical models has also been made without fully considering the complexity of the task.

Despite the current assumption in the literature that confidence in recognition is based upon psychophysical processes, one study has found evidence for an

inferential basis. Busey, Tunnicliff, Loftus and Loftus (2000) conducted a study in which they altered the luminescence 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. In this scenario, psychophysical models of confidence would suggest that confidence should be highest when luminescence levels are kept the same at study and test, as that is when the degree of match is highest. Results that differ from this might indicate that inferential processes were at play.

The results showed that when luminescence was increased from study to test, confidence levels increased. This showed that when more information was available at test, confidence increased, even when that information was non-diagnostic and did not increase the degree of match between the item at study and test. Interestingly, accuracy was stronger when luminescence at test was the same as that at encoding. These results lend themselves to an inferential model of confidence in recognition, as the discrepancy between confidence and accuracy suggests that the presence of additional, non-diagnostic information at test has an effect on confidence without having a corresponding effect on stimulus discriminability.

Due to the conceptual similarities between retrospective and ecphoric confidence, there is reason to believe that results mirroring those of Busey et al. (2000) may be found for using ecphoric confidence ratings. Without evidence, however, the generalizability of this phenomenon cannot be guaranteed, and therefore more research must be done before this can be accepted.

Furthermore, literature surrounding the relationship between feedback and retrospective confidence levels also provides support for an inferential basis for confidence in recognition. A variety of research has found that feedback given to participants after identifying a subject in a facial identification task effects

confidence levels (Bradfield, Wells & Olson, 2002; Semmler & Brewer, 2006; Semler, Brewer & Wells, 2004). Luus and Wells (1994) also found that when participants were asked to identify the perpetrator of a staged crime, confidence levels increased or decreased if told that another participants had given the same or a different answer respectively. Thus, retrospective confidence is affected by external cues.

Measuring confidence

Metacognitive judgments in general, and confidence in particular, index uncertainty. However, it is not clear how best to measure psychological uncertainty. 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 and ecphoric) as a percentage ranging from 0-100% (Brewer & Wells, 2006; Sauer, Brewer & Weber, 2008; Sauer, Weber & Brewer, 2012). However, Windshcitl & Wells (1996) argued that (a) numerical scales are not appropriate for recording confidence, as they do not reflect the way in which people typically conceptualize uncertainty, and (b) that verbal scales would be better suited. In fact, their study 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.

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. In their experiment, participants completed a study phase in which they viewed a sequence of words. They then completed a test phase in

which they were presented with a word and asked to indicate whether they had studied that word before on either a fine-grained or coarse-grained scale. What they found was that participants who used a scale with less options provided higher estimates of recognition than those who used a more fine-grained scale. Benjamin et al. (2013) concluded that this was most likely due to the fact that providing more response items increased cognitive load, which resulted due to the strain of maintaining criteria that is used to parcel subjective evidence into ratings.

Assessing uncertainty in an effective manner is important, particularly when people are asked to make their own judgments based upon somebody else's metacognitive assessment. An example of this includes jury members who may be instructed to take a witnesses confidence into account when interpreting their testimony. Determining the effects of scale type upon the validity of metacognitive assessments is an important step in regards to figuring out the best way to then communicate such uncertainty to others. Leading on from this, one of the interests of the current study was to see whether the key manipulation (i.e., manipulating the amount of non-diagnostic information at test) had differential effects on confidence ratings for participants who used a fine-grained probabilistic scale compared to those who used a coarse-grained verbal scale.

The current study

Given the established link between confidence and recognition, but the competing accounts for this relationship, we tested whether ecphoric confidence ratings were better accounted for by psychophysical or inferential accounts. The aforementioned study by Busey, Tunnicliff, Loftus and Loftus' (2000) forms the starting point for our experimental manipulation. Whilst Busey et al. altered the amount of non-diagnostic information at test by changing the levels of luminance at study and test,

we manipulated the amount of non-diagnostic information by manipulating whether participants viewed a partial face (i.e., top-half of the stimulus only) or a full face at study and test. Thus, the amount of diagnostic information remained the same.

Participants were shown a sequence of "full" and "partial" faces at study, and another set at test. When presented with a face, participants were asked to record how confident they were that they have seen the stimuli before, with half the participants using a scale of 1-100%, and the other half using a 3-point confidence scale with verbal markers (e.g. low confidence, moderate confidence and high confidence). It is important to note that for the purpose of the current study, a correct match did not have to be an exact match. For instance, participants may have viewed a partial face at study followed by the full version of that same at test, and this will have constituted a correct match.

One type of analysis that aims to best represent uncertainty, specifically confidence, is confidence accuracy characteristic analysis (CAC; Mickes, 2015). CAC generally uses a 100-point confidence scale to plot confidence against accuracy on a curve (known as the CAC curve). When representing the data on the CAC curve, however, it is transformed from numerical to categorical data. Whilst the original CAC curve was to have 5 categories (0-20, 30-40, 50-60, 70-80 and 90-100), low response rates in the first three categories resulted in them being collapsed into one larger category for the purpose of the analysis. Interestingly, many researchers who have chosen to use CAC since Mickes' (2015) publication have also adopted the same 3 categories to group their data, categorizing 0-60 as 1 category, 70-80 as the next, and 90-100 as the last (Carlson et al, 2016; Sauerland et al., 2016), rather than defining categories based upon trends encountered in their own data sets.

Inspired in part by the findings of Windshcitl and Wells (1996) and Benjamin at al. (2013), we have manipulated scale type as an independent variable in the current study, with some participants using a numerical scale of 0-100%, and others a 3-point verbal scale (low confidence, moderate confidence, high confidence). This will allow us to examine whether a simplistic, coarse-grained verbal scale may better represent the way in which people conceptualize uncertainty, therefore resulting in a more accurate portrayal of their confidence ratings. Our choice to use a 3-point confidence scale was inspired by Mickes' (2015) confidence accuracy characteristic analysis, which reduces confidence to three levels: "High", "Moderate", and "Low". However, unlike typical CAC research, this scale did not transform numerical data into categorical data. Instead, we simply examined how the collection of confidence data using a simplified, verbal scale (compared to the standard, probabilistic scale) affected the diagnostic value of confidence (i.e., calibration and resolution).

We do recognize that there is a confound between the numerical/verbal and 11(0-100%)/3 levels, however we are not specifically interested in comparing verbal scales to numeric scales or a 11-point scale (0-100%) to a 3-point scale. Instead, we simply wanted to contrast a simple scale alongside the traditional 0-100% to see if they reacted differently to the manipulation. This is of particular interest because, as mentioned before, there is currently a question surrounding how to best communicate uncertainty. If the results show that people are able to use a simplified scale to represent uncertainty in a sensible manner, then it could present itself as an easier method for communicating such uncertainty to others.

Hypothesis 1: Based on previous research, we expected that previously studied faces would receive higher confidence ratings than non-studied faces (Sauer, Brewer & Weber, 2008; Sauer, Weber and Brewer 2012).

The main interest of this study, however, was to test the applicability of psychophysical and inferential models of confidence. Specifically, we were interested in whether (as per inferential accounts) the provision of additional non-diagnostic information at test inflated ecphoric confidence ratings. To answer this question, we were primarily interested in comparing the ratings provided when participants viewed a partial face at study followed by the same partial face at test, compared to when they viewed a partial face at study followed by the corresponding full face at test.

Hypothesis 2a: Taking into account the previous literature demonstrating the inferential basis of both JOLs and FOKs (Koriat, 1993; Koriat, 1997), and the findings by Busey, Tunnicliff, Loftus, and Loftus (2000) that specifically support an inferential basis of confidence in recognition, we expected that when participants were provided with additional, non-diagnostic information at test, they would provide higher confidence ratings than when they were provided with the same level of information at test compared to study. Specifically, we expected that they would record higher confidence ratings in trials in which they were shown a full face at test after viewing the corresponding partial face at study, compared to instances in which they were shown shown a partial face at test followed by the same partial face at study.

Hypothesis 2b: According to psychophysical approaches, we would not expect confidence ratings to change with the inclusion of additional non-diagnostic information at test. Therefore, we would not expect any significant differences in ecphoric confidence for trials where participants viewed 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: Unlike other psychophysical models, Baranksi and Petrusic's (1998) doubt scaling model suggests that an increase in non-diagnostic information would results in a decrease in ecphoric confidence. According to this perspective, we would expect confidence levels to be higher in instances in which participants viewed partial face at study followed by a partial face at test, compared to when they viewed a partial face at study followed by a full 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 our manipulation had differential effects on fine-grained probabilistic scales and coarse-grained verbal scales.

Method

Design

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

Participants

60 participants (43 female), aged 16 to 75 years (M=30.63, SD=14.42), 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

333 colour photographs of male and female (predominantly Caucasian) faces were 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 exact same facial image but cropped so only the top half of the face (i.e., showing the tip of the nose and above). All stimuli and instructions were presented via computer, using purpose-developed experimental software.

Figure 1. An example of a full face (left) and the corresponding half face (right).

Procedure

Participants were tested in groups of up to five people, but completed the task individually. The process ranged from 20 minutes to 45 minutes. 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". In the next slide, they were shown the same but 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". After viewing the instructions on the computer, the experiment began. The experiment was broken up into 6 blocks of trials, with each block consisting of a study phase and a test phase. In the study phase, each participant viewed a sequence of 24 faces. Each of these facial stimuli were 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). As each image was presented, participants were asked to make a judgment as to how confident they were that they had seen the face before. 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, by participants using a mouse to click the on-screen button that corresponded to their level of confidence. The fine-grained scale was made up 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 stimulus as random factors in the models, which allowed random intercepts for these effects. We used the lme4 package (Bates, Maechler, Bolker, & Walker, 2013) 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 confidence changed relative to this reference point, when faces *were* studied (as either a full or partial face [*FS* and *PS*, respectively]) and when the test face was full (*FT*, cf. partial). When the 95% confidence intervals for a coefficient do not overlap zero, we can interpret the predictor as having a significant effect on the outcome at conventional levels of certainty (i.e., $p < .05$).

It is not possible to provide errors bars for figures produced based on MEM analysis. 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.

The primary aim for this study was to determine whether patterns indicative of an inferential or psychophysical basis of confidence in recognition was present in our data set. Evidence for 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. Evidence for the standard psychophysical approach would 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. Evidence for the doubt-scaling model (Baranski & Petrusik, 1998) would be provided if, for the data points labeled "Top only" on the X-axis of *Figure 1* and *Figure 2*, confidence is lower for the "Test: full face" bar than the "Test: top only" bar (as this would provide support for the doubt-scaling model).

First, the data show that confidence ratings were higher at test for stimuli that had been viewed at study than for stimuli that had not. This is visible in *Figure 2* and *Figure 3*, and supported by the findings that, for both the fine- and coarse-grained scales, the TS and FS coefficients are positive and the relevant 95% CIs do not overlap zero. These findings are consistent with those previously established in the literature (Sauer, Brewer & Weber, 2008; Sauer, Weber and Brewer 2012), and provide support for Hypothesis 1.

Interestingly, there did not seem to be any effect of scale-type. Although the different scales of measurement make a direct inferential comparison problematic, the patterns of results are very similar across conditions based on both a visual inspection of Figures 2 and 3, and the coefficients in Tables 1 and 2. This is at odds with our tentative prediction that fine-grained scales would be more vulnerable to inflation effects (Hypothesis 3).

The main interest of this study was to determine whether ecphoric confidence was 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 for each scale are summarised in *Figure 2* and *Figure 3*, and are labeled "Top only" on the X-axis. Both of the figures show that in trials where participants viewed a partial face at study followed by the same partial face, 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. These effects are also evident in the $TS \times FT$ coefficients in Tables 1 and 2. These results support Hypothesis 2c, providing some support for the (psychophysical) doubt scaling model account for confidence in recognition. However, as will be discussed later, this initial support comes with an important caveat.

Table 1

Fixed effect coefficients for linear mixed-effects model predicting confidence on a coarse-grained (verbal) scale.

Fixed effect	*Coefficient*	*SE*	*t*	*95% CI*
Intercept	0.64	0.04	16.75	[0.57, 0.71]
Top at study (TS)	0.47	0.03	16.36	[0.41, 0.52]
Full at study (FS)	0.42	0.03	14.76	[0.36, 0.48]
Full at test (FT)	-0.22	0.02	-9.47	[-0.27, -0.18]
TS × FT	-0.17	0.04	-4.23	[-0.25, -0.09]
FS × FT	0.35	0.04	8.60	[0.27, 0.43]

Table 2

Fixed effect coefficients for linear mixed-effects model predicting confidence on a fine-grained (numeric) scale.

Fixed effects	*Coefficient*	*SE*	*t*	*95% CI*
Intercept	41.50	1.7083	24.294	[38.10, 44.87]
Top at study (TS)	21.68	1.1075	19.574	[19.55, 23.88]
Full at study (FS)	19.09	1.1058	17.263	[16.91, 21.22]
Full at test (FT)	-12.08	0.9047	-13.352	[-13.81, -10.30]
TS × FT	-5.97	1.5674	-3.809	[-9.10, -2.97]
FS × FT	18.22	1.5659	11.636	[15.17, 21.30]

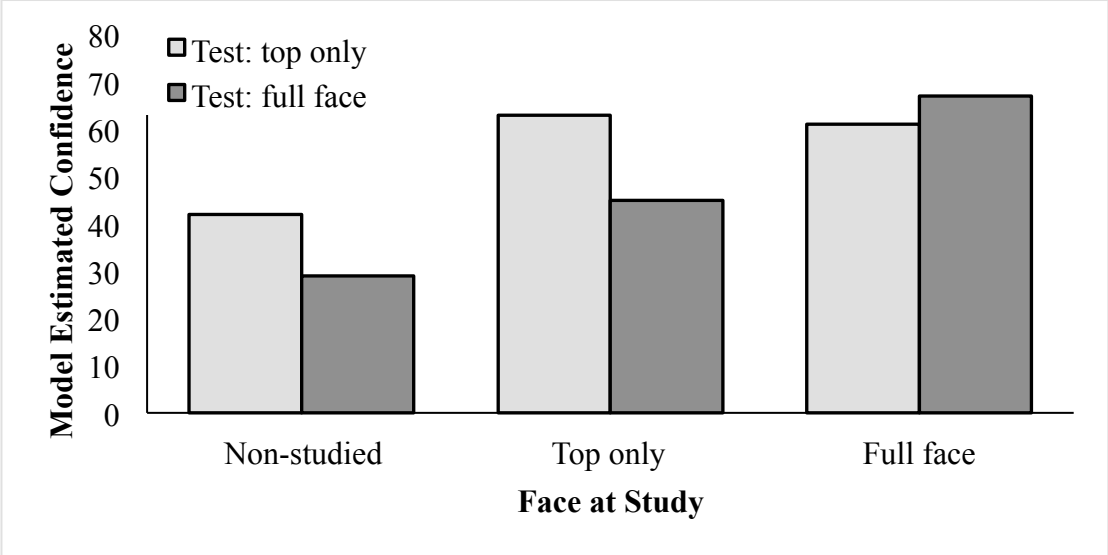


Figure 2. The model-estimated confidence ratings for participants using the fine-grained (numeric) scale, based upon the amount of information provided at study and test.

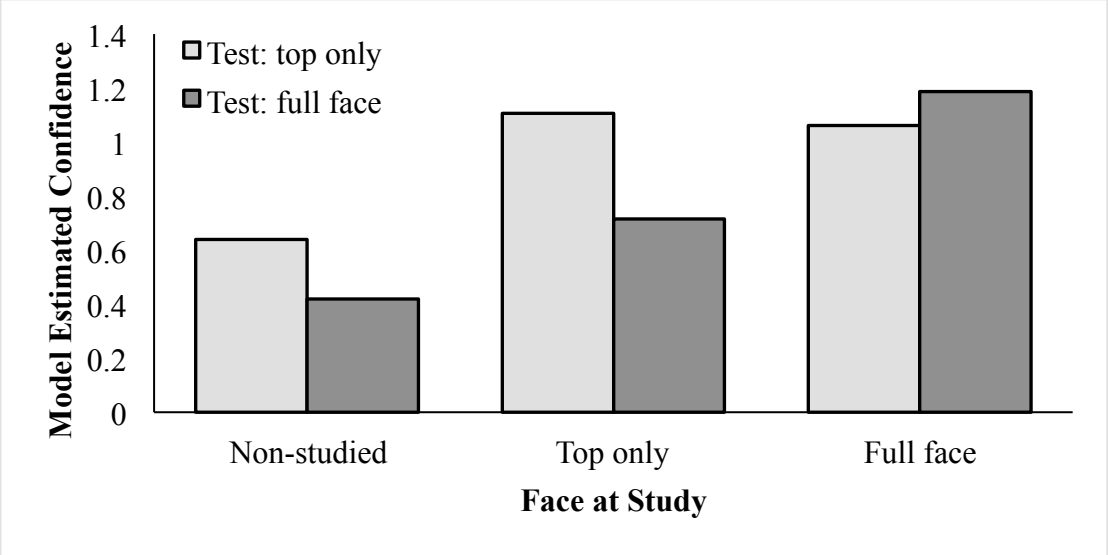


Figure 3. The model-estimated confidence ratings for participants using the coarse-grained (verbal) scale, based upon the amount of information provided at study and test.

Additional exploratory results

Confidence ratings for new faces were significantly lower for full face stimuli than partial face stimuli (see *Figure 2*, *Figure 3*), meaning that participants were better able to distinguish that they had not studied the face before when the full stimuli was presented.

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 (significance is achieved as the confidence intervals for the coefficient do not overlap 0 – see Table 3 and Table 4). These results are important in regards to issues covered in the discussion.

Table 3

Fixed effect coefficients for linear mixed-effects model comparing confidence between FF and FS on a fine-grained (probabilistic) scale.

Fixed effects	*Coefficient*	*SE*	*t*	*95% CI*
Intercept	60.83	2.31	26.33	[56.30, 65.26]
Full at study (FS)	5.84	1.32	4.42	[3.23, 8.46]

Table 4

Fixed effect coefficients for linear mixed-effects model comparing confidence between FF and FS on a coarse-grained (verbal) scale.

Fixed effects	*Coefficient*	*SE*	*t*	*95% CI*
Intercept	1.06	0.05	20.84	[0.96, 1.16]
Full at study (FS)	0.12	0.04	3.5	[0.06, 0.19]

Discussion

This study investigated whether ecphoric confidence ratings were best accounted for by psychophysical or inferential models of metacognition. Participants viewed a series of facial stimuli, with the amount of non-diagnostic information provided at the test phase being the main manipulation of interest. Inferential models of metacognition would propose that the as long as the degree of match was the same, any additional information available at test, regardless of the diagnosticity of that information, would result in higher confidence ratings than if that additional non-diagnostic information was not present (Busey, Tunnicliff, Loftus & Loftus, 2000; Koriat, 1993; Koriat, 1997). In contrast, psychophysical models generally argue that confidence indexes stimulus discriminability (Vickers, 1979; Wixted, 2007). Thus, additional non-diagnostic information should not increase confidence. Furthermore, the doubt-scaling model argues that non-diagnostic information should decrease confidence (Baranksi & Petrusic, 1998).

To manipulate the amount of non-diagnostic information at test, we used a combination of half and full faces (see *Figure 1*). The results showed that when participants viewed a partial face at study followed by the corresponding full face at test (TS/FT), their confidence generally decreased compared to when they viewed a partial face at study followed by the same partial face at test (TS/TT).

These results provide support for a particular model within the psychophysical school of theories of confidence in recognition. Specifically, Baranksi and Petrusic's (1998) doubt scaling model. Whilst the majority of psychophysical models, generally based on SDT or balance of evidence/accumulator models, put forward that confidence is proportionate to the degree of match, only the doubt-scaling model provides an explanation as to why confidence in recognition would decrease in the

presence of extra, non-diagnostic information at test. Unfortunately, there is almost no information in the literature regarding the doubt-scaling model, aside from two paragraphs in a journal article by Baranksi and Petrusic (1998) that referenced an earlier paper that does not seem to exist. Nonetheless, we do understand is that it proposes that confidence is negatively associated with the amount of ambiguous information sampled at test. Thus, an increase in non-diagnostic information (in this case the additional facial information at test) results in a decrease in confidence.

Interestingly, our findings are at odds with those reported by Busey et al. (2000). Busey et al. 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. There are a variety of explanations that could be put forward to explain this divergence in our results, all of which lie in the nature of the manipulations used. One important way in which our study differed from Busey et al.'s is that by changing the luminescence of the entire stimuli, they changed the nature of the entire stimuli. That is, adjusting the luminance of the stimulus changes the appearance of the stimulus as a whole. In contrast, by adding or removing the bottom half of the face, but 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*). Without a direct comparison of these manipulations (which our study did not include), however, this explanation is purely speculative. Another potential explanation for the difference in between our results and Busey et al.'s findings could lie in the way in which they altered the levels of luminescence. 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 skewed the results. , as these stimuli would have contained very little diagnostic information, which may have lead to an increased reliance on inferential cues.

A final potential explanation could relate to the holistic nature of facial processing. There is some evidence in the literature that faces may be processed differently to other stimuli. Specifically, some researchers have suggested that faces are processed holistically rather than using individual features to construct a whole (Maurer, Le Grand & Mondloch, 2002). This effect has not been found with any other class of stimuli, except for in instances when a person is considered an “expert” in that field (e.g., race-car drivers and car brands) (Le Grand, Mondloch, Maurer & Brent, 2004). Whether this reliance on holistic processing is predominantly related to faces, or more broadly characteristic of any stimulus-type with which the perceiver has sufficient expertise, there is compelling evidence that individuals processes faces holistically. The holistic processing of faces relies not only on an individual’s ability to recognize facial features, but their ability to recognize the interconnection between them (Maurer, Le Grand & Mondloch, 2002). This is demonstrated by the facial inversion effect, in which people’s facial recognition ability is significantly impaired when the face is inverted (Farah, Tanaka & Drain, 1995; Freire, Lee & Symons, 2000), and that this effect is disproportionate compared to other types of stimuli (Yin, 1969). It has also been demonstrated by studies that have found that participants show reduced ability to recognize a specific facial feature in isolation

compared to when it is presented as part of a face (Tanaka & Farah, 1993; Tanaka & Sengco, 1997).

This reliance on holistic processing may contribute to the differences in results between our study and that of Busey, Tunnicliff, Loftus and Loftus (2000). 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 the current study, however, we were actually modifying the physical makeup of the face by removing certain facial features. To explain this more thoroughly, participants encoded half a face a test (whether this would have involved holistic or non-holistic processing has not been established). They were then presented with a full-face stimulus which they processed holistically. They may then have tried to compare this new, holistic stimuli with the old, half-face stimuli and determined that the overall degree of match was lower than in trials where they viewed the partial face followed by the same partial face. That is, participants may have not viewed a full face at test (following a partial face at study) as a "matching stimulus with some additional information" but as a stimulus that, on the whole, did not match as well with the partial image encoded at study. This decrease in match could therefore account for the lower confidence ratings.

An example of the holistic processing of faces, which actually mirrors the manipulation in our study in some ways, is the composite-face effect. The composite-face effect refers to a phenomenon in which people are better able to identify the top half of a face as matching another, previously seen face if it is not aligned perfectly with the bottom half of the stimuli (Carey & Diamond, 1994; Le Grand, Mondlock, Maurer & Brent, 2004). A simple way to demonstrate this effect is to look at an experiment by Le Grand, Mondlock, Maurer and Brent (2004). In this

study, participants were shown a series of paired facial images, during which they had to indicate whether the top half of each face was the same of different. In half of the trials, both halves of the faces were aligned as to look like a "normal" face. In the other half of the trials, the top and bottom halves of each face were separated so that they did not resemble a "normal" face (with the ear on the right hand side being lined up with the center of the top half of the face). The researchers reported accuracy rates of 63% for the "normally" aligned faces, compared to 86% for the misaligned faces. The fact that participants were significantly worse in trials where the faces were aligned compared to when they were misaligned suggests that holistic processing plays a role (i.e., a tendency to perceive the "aligned" stimuli as a whole, rather than a composite constructed from two independent faces) – demonstrating the composite-face effect.

What is particularly interesting when talking about the composite-face effect in regards to our study is the similar findings the two manipulations produced; despite the fact that one was testing a persons' ability to discriminate between stimuli, and the other was testing the ability to recognize whether a stimuli had been seen earlier. Le Grand, Mondlock, Maurer and Brent (2004) demonstrated that participants were able to identify that two misaligned faces were the same, and this is mirrored by our results, which showed that participants were able to recognize the same partial face at test (*Figure 1, Figure 2)*. Although research looking at the composite-face effect included the bottom half of the face, the misalignment as to disrupt holistic processing meant that the stimuli itself was virtually the same as our partial faces. The findings that participants were less able to identify two matching top halves of faces when they were part of a full face with different bottom halves also has similarities to the reduced ability that participants in the current study had when

shown a partial face at study followed by the corresponding full face at test (*Figure 1, Figure 2)*. While the practical implications of these similarities are unclear, it is interesting from a theoretical perspective to note that the way in which people process stimuli as "matching" seems to be quite similar whether both stimuli are physically in front of them, or if they are relying on a mental representation.

That confidence was significantly lower in trials where participants viewed FS/TT compared to when they saw FS/FT provides further support for the idea that holistic facial processing may be playing a part (see Table 3, Table 4). If the holistic nature of facial processing caused a mismatch between stimuli in TS/FT trials, the same effect should be evidence for FS/TT trials. The fact that the decrease in confidence was larger when contrasting TS/FT and TS/TF compared to when contrasting FS/FT and FS/TT (see *Figure* 2 and *Figure* 3) suggests that both holistic facial processing and the doubt-scaling mechanisms could have shaped participants confidence ratings. To elaborate, the doubt-scaling model would predict that confidence should be higher in FS/TT trials than TS/FT trials, as there is no non-diagnostic information present in the FS/TT trials. Therefore, the fact that FS/TT and TS/FT produced lower confidence ratings than FS/FT and TS/TT respectively could be a result of holistic facial processing, but the fact that TS/FT showed a larger reduction in confidence than FS/TT (compared to TS/TT and FS/FT respectively) might suggest a doubt-scaling effect.

Relying only on our current results, we cannot determine whether the effects we observed are witnessing are attributable to a doubt-scaling mechanisms or a holistic facial processing mechanism. Consequently, a follow up study on aims to eliminate the potential contribution of holistic processing effects. In short, the follow up study will utilize the same method and design as the current study. The crucial difference,

however, lies in the nature of the stimuli. Where the current study used facial images, the follow up study will use images of houses and landscapes. This will allow us to isolate two potential mechanisms that could have contributed to the current pattern of results. First, it will isolate the effect of holistic processing.. If we find the same pattern for houses (a stimulus type not associated with greater reliance on holistic processing), this would provide evidence for the doubt-scaling model. If not, then we would have to concede that the current study may not have allowed for us to effectively compare the different theories of confidence in recognition that we set out to examine.

Secondly, it will allow us to determine whether *knowing* the study item is incomplete (i.e., a partial version of a full face) affects performance. Participants in the present study were obviously aware that, when viewing a partial face at study, they were viewing an incomplete stimulus. By using images of landscapes, which can be cropped without the result being obviously incomplete, we can create a situation where participants may not be able to tell that they are only seeing one half of an image. By comparing the pattern of results obtained for houses (for which the cropping will result in an obviously incomplete stimulus) and landscape stimuli, we can determine whether the knowledge that a study stimulus was incomplete affects ecphoric confidence at test.

Returning back to the current study, it was interesting to find that there appeared to be no interaction of our manipulations with scale type. Given the argument by Windshcitl and Wells (1996) that people are not well equipped to use probabilistic scales to represent uncertainty, there are three potential explanations for this. Firstly, it could be that the coarse-grained verbal scale that we formulated does not adequately address uncertainty either, however given that both scales showed

evidence of discrimination, (i.e., the ability to differentiate between studied and non-studied test times) this seems unlikely. Furthermore, the pattern of results suggests that, for both scales, confidence tracked the amount of useful information (i.e., degree of match) present at test (e.g., confidence when participants saw a full face followed by another full face was greater than full face followed by a partial face). Secondly, it could be that Windshcitl and Wells were not correct in their assumptions, and that people may well be able to use a probabilistic scale to represent confidence. The final explanation could be that scale-type effects depend on the type of uncertainty being measured: Thurstonian or Brunswikian.

In their experiments, Windshcitl and Wells (1996) required participants to predict an uncertain outcome (e.g., in Experiment 1, participants were asked to predict the likelihood of someone winning a lucky-draw prize), which falls into the category of Brunswikian uncertainty. It was mentioned earlier in this paper that it is not clear-cut whether a task like ours represents a Brunswikian or Thurstonian approach to uncertainty, however the fact our results most closely represent the doubt-scaling model suggest that perhaps a Thurstonian approach is most appropriate (due to the reliance on internal monitoring). Therefore, it could be that people are better able to use probabilistic scales to represent uncertainty that arises due to noise in ones own information processing (Thurstonian) than uncertainty which arises due to the uncertain nature of future predictions of performance (Brunswikian) given the conceptual difference between the two.

In sum, our study set out to test predictions derived from psychophysical or inferential models of confidence in recognition. Although our results map neatly on to a psychophysical approach (specifically, for the doubt-scaling model), further research is needed to determine whether our results have been affected upon by

potentially confounding mechanisms reflecting the holistic nature of facial processing.

References

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. M., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4 classes. Retrieved from http://lme4.r-forge.r-project.org

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

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., 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., & 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

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. Retrieved from https://link.springer.com/journal/13423

Carey, S., & Diamond, R. (1994). Are faces perceived as configurations more by adults than by children? *Visual Cognition, 1*, 253–274. doi: 10.1080/13506289408402302

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 14*(2), 118-128. doi: 10.1037/1076-898X.14.2.118

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.

Eisenacher, S., & Zink, M. (2017) The Importance of Metamemory Functioning to the Pathogenesis of Psychosis. Front. Psychol. 8:304. doi: 10.3389/fpsyg.2017.00304

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What Causes the Face Inversion Effect? *Journal of Experimental Psychology: Human Perception and Performance 21*(3), 628-634. doi: 10.1037/0096-1523.21.3.628

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive developmental inquiry. *Am. Psychol. 34*, 906–911. doi: 10.1037/0003-066X.34.10.906

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a defecit in

the encoding of configural information: Direct evidence. *Perception, 29*, 159-170. doi:10.1068/p3012

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.

Hart, J, T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*, 208-216. doi: 10.1037/h0022263

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.

King, J. F., Zechmeister, E. B. & Shaughnessy, J. J. (1980) Judgements of knowing. The influence of retrieval practice. *American Journal of Psychology. 93*(2), 329-343. Retrieved from http://www.press.uillinois.edu/journals/ajp.html

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

Le Grand, R., Mondloch, C. J., Maurer, D., & and Brent, H. P. (2004). Impairement

in Hollistic Face Processing Following Early Visual Deprevation. *Psychological Science, 15*(11), 762-768. doi: 10.1111/j.0956-7976.2004.00753.x

Luus, C. A. E., Wells, G. L (1994) The Malleability of Eyewitness Confidence: Co-Witness and Perseverance Effects. *Journal of Applied Psychology, 79*(5), 714-723. doi: 10.1037/0021-9010.79.5.714

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

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior, 5*(4), 299-309. doi: 10.1007/BF01044945

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

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

Metcalfe, J., & Dunlosky, J. (2008). Metamemory. In H. L. Roediger, III (Ed.), *Learning and Memory: A Comprehensive Reference* (Vol. 2, pp. 349-362): Elsevier.

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

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

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

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. Retrieved from https://link.springer.com/journal/10979

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

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

Semmler, C., & Brewer, N. (2006). Postidentification feedback effects on face recognition confidence: Evidence for metacognitive influences. Applied Cognitive Psychology, 20, 895-916. doi:10.1002/acp.1238

Semmler, C., Brewer, N., & Wells, G.L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification confidence. Journal of Applied Psychology, 89, 334-346. doi:10.1037/0021-9010.89.2.334

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662. doi: 10.1037/h0054651

Tanaka, J.W., & Farah, M.J. (1993) Parts and wholes in face recognition. *The Quaterly Journal of Experimental Psychology, 46*a, 225–245. doi: 10.1080/14640749308401045

34 Tanaka, J.W. and Sengco, J. (1997) Features and their configuration in face recognition. *Memory and Cognition, 25*, 583–592. doi: 10.3758/BF03211301

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. (1979). Decision processes in visual perception. New York: Academic Press.

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

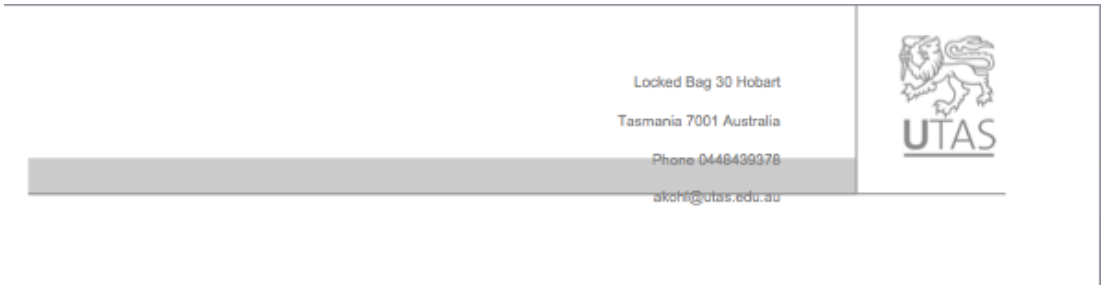
Windshitl, P. D., & Wells, G. L. (1996). Measuring Psychological Uncertainty: Verbal Versus Numeric Methods. *Journal of Experimental Psychology: Applied, 2*(4), 343-364. doi: 10.1037/1076-898X.2.4.343

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

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141-145. doi:10.1037/h0027474

Appendix A

Consent Form.



Facial Recognition Study

Participant Consent Form

1. I agree to take part in the research study named above.
2. I have read and understood the Information Sheet for this study.
3. The nature and possible effects of the study have been explained to me.
4. I understand that the study involves viewing a series of stimuli and answering questions about them.
5. I understand that participation involves no foreseeable risks.
6. I understand that all research data will be securely stored on the University of Tasmania premises for five years from the publication of the study results, and will then be destroyed unless I give permission for my data to be archived.

   I agree to have my study data archived. (Note that your data will be stored anonymously.)

   Yes ☐ No ☐
7. Any questions that I have asked have been answered to my satisfaction.
8. I understand that the researchers will maintain confidentiality and that any information I supply to the researcher will be used only for the purposes of the research.
9. I understand that the results of the study will be published so that I cannot be identified as a participant.
10. I understand that my participation is voluntary and that I may withdraw at any time without any effect.

    I understand that I will not be able to withdraw my data after completing the experiment as my data will be anonymous.

Participant's name: _____

Participant's signature: _____

Date: _____

Locked Bag 30 Hobart
Tasmania 7001 Australia
Phone 0448439378
akohl@utas.edu.au

Statement by Investigator

☐ I have explained the project and the implications of participation in it to this volunteer and I believe that the consent is informed and that he/she understands the implications of participation.

If the Investigator has not had an opportunity to talk to participants prior to them participating, the following must be ticked.

☐ The participant has received the Information Sheet where my details have been provided so participants have had the opportunity to contact me prior to consenting to participate in this project.

Investigator's name: _____

Investigator's signature: _____

Date: _____

Appendix B

Debriefing Form.

Participant no:_____

Initial Debrief

Study:	**Facial Recognition Task**
Researcher:	Amelia Kohl, Psychology Honours Student, University of Tasmania, akohl@utas.edu.au

What were the aims of this study?
This study investigated factors affecting confidence in memory for studied items. In order to preserve the scientific rigour of the research (by ensuring that future participants remain naive to the purpose of the experiment and the full experimental hypotheses) we will not be providing a full debrief at this time. However, we will provide a full debrief as soon as data collection is complete.

You can also obtain a summary of the results of the study by writing an email to Amelia Kohl using the contact information above. We expect that such a summary will be available by October 2017. It will be emailed to you automatically if you enter your email address into our results request list.

If you have any questions about the study please ask the experimenters, they will be happy to answer them now (although they are unable to reveal the exact purpose/hypotheses of the experiment).

If, for any reason, you wish to withdraw your data once you have left you can do this by writing an email to this effect to Amelia Kohl using the contact information provided above and quoting your participant number at the top of this sheet.