How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks

Ruth Horry

Swansea University and Flinders University

Neil Brewer

Flinders University

In press at *Journal of Experimental Psychology: General*

Ruth Horry, Department of Psychology, Swansea University , and School of Psychology, Flinders University; Neil Brewer, School of Psychology, Flinders University;

This research was supported by ARC-DP150101905 awarded to N. Brewer et al. and a Faculty of Social and Behavioural Sciences Research Grant from Flinders University, awarded to N. Brewer and R. Horry. We thank Nicola Guerin for her assistance with data collection.

Data from some of the experiments reported in this manuscript were orally presented at the 2013 annual conference of the European Association of Psychology and Law (Coventry, UK), and at the 2015 meeting of the Society for Applied Research in Memory and Cognition (Victoria, BC, Canada).

Correspondence concerning this article should be addressed to Ruth Horry, Department of Psychology, College of Human and Health Sciences, Vivian Tower, Swansea University, Swansea, SA2 8PP. Email: r.horry@swansea.ac.uk.

Abstract

Confidence judgments in two-alternative decisions have been the subject of a great deal of research in cognitive psychology. Sequential sampling models have been particularly successful at explaining confidence judgments in such decisions, and the relationships between confidence, accuracy, and response latencies. Across five experiments, we derived predictions from sequential sampling models and applied them to more complex decisions: multiple-alternative decisions, and compound decisions, such as eyewitness identification tasks, in which a target may be present or absent within the array of items that can be selected. We hypothesized that, when a decision-maker chooses an item, confidence in that decision reflects the relative evidence for the chosen item over all unchosen items. We tested this hypothesis by manipulating the similarity between the target (or target-replacement, for trials in which the target was not present in the array) and the weakest lure(s). As target-lure similarity decreased, confidence in correct target identifications increased, while response latencies decreased. When the decision-maker chose none of the items, the similarity between the target-replacement and the lures was unrelated to confidence. We conclude that similar mechanisms underpin confidence judgments in multiple-alternative and positive compound decisions as in simpler, two-alternative decisions. A goal of future research should be to formally extend sequential sampling models to more complex decisions, such that it will be possible to establish whether diffusion or accumulator models provide a better fit to the data.
Keywords: *Confidence, response latencies, n-alternative decisions, compound decisions, eyewitness identification.*

How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks

There is a long history in psychology and psychophysics of using confidence ratings to understand the processes through which an individual arrives at a decision (e.g., Angell, 1907; Henmon, 1911; Williamson, 1915). In recent decades, a concerted effort has been made to understand how these confidence judgments are produced, with sequential sampling models proving particularly successful (e.g., Brown & Heathcote, 2008; Ratcliff & Starns, 2009; Usher & McClelland, 2001). However, these models have largely been used to predict confidence in simple, binary-choice decisions, such as determining which of two stimuli is brighter (Teodorescu, Moran, & Usher, 2016) or which of two lines is longest (Yu, Pleskac, & Zeigenfuse, 2015). The aim of this paper is to test predictions derived from these models in the context of more complex decision tasks that are more closely aligned to the types of decisions we make in daily life. Specifically, we investigate *n*-alternative forced-choice decisions, such as might be faced by a candidate sitting a multiple-choice exam, or by a physician deciding which of several treatment regimens to prescribe to a patient. We also investigate compound decisions, in which the decision-maker must decide whether a target item is present in or absent from an array of stimuli, such as might be faced by an eyewitness attempting to identify a culprit from a police photo-array or lineup. To foreshadow, our findings suggest that similar cognitive mechanisms may produce confidence judgments in these complex decisions as in the simple decisions that have formed the basis of much cognitive research.

Theoretical perspectives on confidence judgments

In basic cognitive psychology research, confidence judgments are frequently elicited following memorial (Heathcote, Bora, & Freeman, 2010; Mickes, Hwe, Wais, & Wixted, 2011; Stretch & Wixted, 1998) or perceptual (Baranski & Petrusic, 1994; Festinger, 1943; Juslin & Olsson, 1997) decisions. In a typical paradigm, the decision-maker is faced with two

response options: Is line A or line B longest? Was this item on the study list or is it a new item? Following the decision, the participant indicates her confidence that the decision was correct.

Over many decades, several regularities have emerged regarding the relationships between decision accuracy, confidence, and response latencies. First, when difficulty is manipulated, conditions with higher accuracy are associated with higher mean confidence and shorter decision latencies than conditions with lower accuracy (e.g., Baranski & Petrusic, 1998; Björkman, Juslin, & Winman, 1993; Festinger, 1943). Second, when difficulty is held constant, there is a positive relationship between confidence and accuracy (e.g., Baranski & Petrusic, 1994; Hiller & Weber, 2013; Juslin & Olsson, 1997). In non-speeded tasks, there is also a negative relationship between confidence and decision latencies (e.g., Kiani, Corthell, & Shadlen, 2014; Pleskac & Busemeyer, 2010; Vickers & Packer, 1982). To explain these regularities, a class of models called sequential sampling models have been proposed. There are many sequential sampling models, which differ in their details. It is beyond the scope of this paper to provide a comprehensive review of these models and how they differ, and it is not our intention to critically test these models against each other. Rather, we focus on the common underlying assumptions of the models.

All sequential sampling models assume that, when an observer is faced with a decision, evidence accumulates over time until a decision threshold is reached. Let us imagine a two-alternative perceptual discrimination task in which the decision-maker must decide which of two lines, A or B, is longest. In diffusion, or random walk models, (e.g., Ashby, 1983; Ratcliff, 1978), the accumulated evidence is stored on a single variable. At any moment in time, the state of accumulated evidence is represented as a single value that indicates the extent to which the evidence favours one response option over the other. This variable drifts towards one of two decision thresholds – one for choice A and one for choice

B. When the evidence reaches one of these thresholds, the appropriate decision is made. In accumulator, or race models (e.g., Brown & Heathcote, 2008; Usher & McClelland, 2001; Vickers, 1970), each response option is associated with its own counter upon which evidence accumulates. Thus, at any single moment in time, the state of evidence accumulation is presented by *n* values, where *n* is the number of response options. In the line discrimination example, line A and line B would each have a corresponding evidence counter. Just as with diffusion models, evidence accumulates until a decision threshold is reached, at which point the corresponding decision is made.

Sequential sampling models have been very successful at explaining relationships between decision accuracy and response latencies. If the evidence more strongly favors one option over the other, the evidence will accumulate rapidly, leading to a fast decision. Furthermore, when evidence strongly favors one decision over another, that decision is likely to be correct. Thus, these models can explain why conditions in which stimuli are more discriminable produce faster, more accurate decisions than conditions in which stimuli are less discriminable. The models also explain the negative accuracy-latency relationship observed within conditions; trial-to-trial variability in evidence accumulation is associated with variability in both response latencies and accuracy.

The extension of sequential sampling models to confidence judgments is more complex. In accumulator models, confidence is assumed to index the difference in the end state of accumulated evidence between the chosen and unchosen response options; this has been termed the *balance-of-evidence* hypothesis (P. L. Smith & Vickers, 1988; Van Zandt, 2000). Thus, a trial that produced much more evidence in favor of option A over option B will be associated with high confidence; conversely, a trial in which the race between the two counters was much closer will be associated with low confidence. In this way, accumulator models can account for: i) the difference in mean confidence between high- and low-

discriminability conditions (higher discriminability is associated with a larger difference in the end-state of the evidence between the response options); ii) the positive relationship between confidence and accuracy (trials in which the evidence strongly favours one option over the other are more likely to be correct, and will produce higher confidence, than those in which the evidence is more closely balanced); and iii) the negative relationship between confidence and response latencies (when evidence accumulates rapidly on the winning counter, the decision will be made more quickly; furthermore, the balance-of-evidence will more strongly favor the winning response, producing high confidence judgments).

Diffusion models cannot be extended to confidence judgments without some extra assumptions (Usher & McClelland, 2001). Recall that in diffusion models the evidence is represented on a single variable, which indexes the relative evidence for one option over the other, and that the decision is made when this variable reaches the decision threshold. Consequently, there is no variability in the end state of the evidence at the time of the decision, and therefore, no basis upon which to produce graded confidence judgments. To overcome this obstacle, several two-stage models have been proposed, in which evidence continues to accumulate after the decision is made (e.g., Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010). Confidence is scaled not from the state of evidence at the time of the decision (*t1*), but from the state of the evidence at the time of the confidence judgment (*t2*). It is the variability in the rate of evidence accumulation between *t1* and *t2* that allows for graded confidence judgments, with rapid accumulation of evidence associated with high confidence. These two-stage diffusion models are also able to account for: i) the difference in mean confidence between high- and low-discriminability conditions (the rate of the post-decisional evidence accumulation will be higher in high-discriminability conditions, producing higher mean confidence judgments); ii) the positive relationship between confidence and accuracy (accurate trials are associated with higher post-decisional drift rates

than inaccurate trials, consequently producing higher confidence judgments); and iii) the negative relationship between confidence and response latencies (pre- and post-decisional drift rates are associated, such that a decision made quickly will continue to accumulate evidence rapidly in the post-decision period, leading to a high confidence judgment).

In sum, sequential sampling models have been developed to account for commonly observed relationships between decision accuracy, confidence, and response latencies. Though there are some differences between models, they all assume that evidence accumulates over the course of a trial until a decision can be made. Importantly, across all sequential sampling models, confidence judgments express the *relative* evidence for the chosen response over the unchosen response. In the case of accumulator models, confidence judgments index the balance-of-evidence stored in the accumulators; in diffusion models, confidence is scaled from a single variable that represents evidence favoring one response option over the other.

Confidence judgments in *n*-alternative decisions

The models outlined above were largely developed in the context of 2-alternative forced-choice (2AFC) decisions. However, the types of decisions we face in day-to-day life are often not as simple as deciding between two options. In this section, we discuss the extension of sequential sampling models to forced-choice decisions with more than two alternatives.

A major strength of accumulator models is that they can be easily applied to multiple-alternative decisions (Brown & Heathcote, 2008; Usher & McClelland, 2001). Each response option is associated with its own counter upon which evidence accumulates. There is no limit to the number of counters and, hence, no limit to the number of response options that can be considered. In some accumulator models (e.g., Usher & McClelland, 2001), lateral inhibition occurs, such that a build-up of evidence on one counter comes to dominate the others; in

other models (e.g., Brown & Heathcote, 2008), evidence accrual is independent across counters.

However, when sequential sampling models have been applied to multiple-alternative decisions, the focus has usually been on explaining the effect of choice set size on accuracy and response times in absolute identification tasks (Hick, 1952; Lacouture & Marley, 2004; Usher, Olami, & McClelland, 2002). Here, we try to answer a rather different question: given a multiple-alternative decision with a specific set size, how are confidence judgments formed?

As argued previously, it is widely agreed that confidence judgments are largely relative (though they may also carry information about absolute stimulus values; Teodorescu et al., 2016; Zawadzka, Higham, & Hanczakowski, 2016). In a 2AFC task, confidence indexes the degree to which the evidence favors the chosen option over the unchosen option. Thus, we would expect that, within an *n*-alternative forced choice task, confidence will index the degree to which the evidence favors the chosen option over the unchosen options. The question we ask is this: do confidence judgments in *n*AFC decisions contrast the strength of evidence associated with the chosen alternative with the strength of evidence for *all* unchosen alternatives? We investigate this possibility within a task analogous to an eyewitness identification test, which would allow us to consider both the theoretical and applied implications of such a finding.

There is reason to think that this may be the case. Windschitl and Chambers (2004) showed that likelihood assessments of a particular outcome increase when some implausible alternatives are added. The authors argued that the inclusion of "duds" increases the perceived evidence strength of the focal alternative. Charman, Wells, and Joy (2011) extended these findings to confidence judgments in eyewitness identification decisions by adding implausible filler items to a lineup. The inclusion of these duds boosted confidence in

the chosen (plausible) item. Such an effect could be the result of top-down meta-cognitive inferences (e.g., "I was able to exclude a couple of items, so my memory must be pretty good!"), or it could be the bottom-up result of a sequential sampling system in which confidence indexes relative evidence strengths between a chosen item and all non-chosen items (e.g., P. L. Smith & Vickers, 1988; Van Zandt, 2000).

If sequential sampling models can be successfully extended to *n*AFC tasks, then we might expect the general findings from 2AFC tasks to transfer. Namely, we would predict that: i) Mean confidence will be higher and mean response latencies will be shorter under conditions of higher discriminability; ii) when discriminability is held constant, confidence and accuracy will be positively correlated; and iii) when discriminability is held constant, response latencies and accuracy will be negatively correlated. Furthermore, if confidence judgments carry information from *all* items within the choice array, then we would expect confidence to increase as the evidence associated with the weakest item decreases. Importantly, if such an effect were observed even when the weakest item was not an obvious 'dud', this would be suggestive of a bottom-up process, rather than top-down metacognitive processes.

Confidence judgments in compound decisions

In compound decision tasks, decision-makers are presented with an *n*-alternative choice array that may or may not include a target, and they must decide: i) whether a target is present or absent; and ii) the identity or location of the target. Examples of real-world compound decision tasks include airport security personnel scanning luggage for dangerous or prohibited items, radiologists searching images for particular types of abnormalities, and eyewitnesses attempting to identify a perpetrator from a police lineup (Duncan, 2006). These tasks are considerably more complicated than 2AFC tasks. In addition to the *n*-alternative

nature of compound decision tasks, the decision-maker has the opportunity to decide that the target item is not present. We focus on the latter of these complications in this section.

Let us consider a recognition task. On each test trial, the participant must decide whether a specific item was previously studied or not. Accumulator models of recognition assume that positive and negative samples of evidence accumulate on separate counters (Van Zandt, 2000). With each sample, the degree of match between the test item and the stored item (or items) is assessed (which may include familiarity and recollective detail; Wixted & Mickes, 2010). If a sample produces low evidence of match, (or evidence of mismatch), the sample will accumulate on the "new" counter; if the sample produces high evidence of match, the sample will accumulate on the "old" counter. Just as with a 2AFC task, a decision is made when one of the counters reaches the appropriate threshold. In these models, confidence is scaled from the balance of evidence in accumulated "old" and "new" evidence.

An alternative accumulator model of recognition is Ratcliff and Starns' (2013) RTCON2 model. In this model, match-to-memory is available to the decision-maker not as a single value but as a distribution. The strength distribution is partitioned into separate regions, with each region corresponding to a particular combination of decision type (e.g., "old" or "new") and confidence (e.g., "high", "medium", or "low"). Each possible decision-confidence pairing is associated with an accumulator, racing towards its own threshold. The rate at which evidence accumulates on each counter is determined by the area of the strength distribution that falls within the corresponding region. One disadvantage of this model is that the number of accumulators grows rapidly with each additional response option (e.g., choosing between 2, 3, or 4 items) and with increasingly fine-grained confidence scales. For this reason, we will not focus on this particular model here, instead extracting predictions from simpler models that scale more easily to the complex 4-alternative, compound decisions that we investigate in Experiments 4 and 5.

In diffusion models of recognition memory, the decision variable on which evidence accumulates indexes the degree of match between a target item and an item, or items, stored in memory (e.g., Ratcliff, 1978; Ratcliff, Thapar, & McKoon, 2004). Without extension, these models would suffer from the same problems as in a 2AFC task; a lack of variability in the end state of the accumulated evidence would provide no basis for confidence. Thus, additional assumptions, such as post-decisional processing, would be required, in which confidence is scaled from the evidence state at the time of the confidence judgment (e.g., Moran et al., 2015; Pleskac & Busemeyer, 2010).

Let us now consider a compound recognition task, such as that faced by an eyewitness viewing a police lineup. The decision-maker is presented with *n* items, and must decide whether any of these items match a previously seen target. To our knowledge, sequential sampling models have not been formally extended to compound decision tasks, and so we necessarily must become speculative here. Because accumulator models are more easily extended to multiple-choice decisions than diffusion models, we will take an accumulator perspective here. Let us assume that each of the *n* items is associated with a counter, as in a *n*AFC task (e.g., Brown & Heathcote, 2008; Usher & McClelland, 2001); we will refer to these as *item counters*. Furthermore, negative evidence accumulates on a separate, "not present" counter (similar to how negative evidence accumulates in recognition decisions; Van Zandt, 2000); we will refer to this as the *negative counter*. We will also assume that the decision-maker sets two thresholds; one for making a positive decision (i.e., choosing an item), and one for making a negative decision (i.e., deciding that no target is present). The decision-maker responds when one of two events occur: when the evidence on one of the item counters reaches the positive decision threshold; or when the evidence on the negative counter reaches the negative decision threshold.

When a positive decision is reached, we assume that confidence is produced by the same processes as in a *n*AFC task. Specifically, confidence is computed from the balance-of-evidence between the chosen item and the unchosen items. A growing body of research in the eyewitness identification literature has documented a robust confidence-accuracy relationship for witnesses who make a positive choice (e.g., Brewer & Wells, 2006; Horry, Palmer, & Brewer, 2012; Palmer, Brewer, Weber, & Nagesh, 2013; Sauerland & Sporer, 2009). For positive decisions, there is also a negative response latency-accuracy relationship (e.g., Brewer, Caon, Todd, & Weber, 2006; Sauerland & Sporer, 2009; S. M. Smith, Lindsay, Pryke, & Dysart, 2001). Furthermore, conditions associated with higher accuracy also produce higher mean confidence (e.g., full vs. divided attention; Palmer et al., 2013). Thus, the same regularities that have been observed in 2AFC tasks have also been observed for positive decisions in more complex, compound decision tasks.

When a negative decision is reached, we assume that this is because the evidence on the negative counter accumulated to threshold. The rate of accumulation should be faster, on average, for target-absent trials than for target-present trials, producing a negative response latency-accuracy relationship. But how are confidence judgments scaled in negative decisions? Given a negative decision, a decision-maker must evaluate their confidence that *none* of the items matched the target (Weber & Brewer, 2004; Lindsay et al., 2013). This would seem to place somewhat different demands on the decision-maker than a confidence judgment for a positive decision, in which an individual item is chosen. One possible mechanism would be to scale confidence from the distance between the evidence accumulated for the best (though unchosen) item and the positive decision threshold. Consequently, confidence judgments in negative decisions would become less of a relative index of evidence strengths across items, and more of an absolute judgment of match associated with the best item. If this were the case, we would expect to see a positive

confidence-accuracy relationship for negative decisions, as the distance between the best-match and the positive threshold should vary systematically with target-presence. However, we would not expect the perceptual similarity between the target-replacement and the lure to affect confidence, as this relative difference should have little or no impact on the absolute strength of the best match.

In summary, we expected to replicate the standard relationships between confidence, response latencies, and accuracy in compound decision tasks. We also predicted that confidence in positive decisions would increase as target-lure similarity decreased. We did not expect to observe such an effect for negative decisions, as we reasoned that these judgments would index the absolute strength of the best match, rather than the relative strengths of the items.

Overview of the experimental paradigm

In each of the experiments described in this paper, we used the same basic paradigm. Participants completed a series of trials in which they briefly studied a target face, and then attempted to identify that target from an array consisting of two items (Experiments 1 and 3) or four items (Experiments 2, 4, and 5), before rating their confidence in their decision. Our approach was to build up the complexity of the decision gradually over the course of several experiments. Thus, we began with a simple 2AFC task (Experiment 1) before moving on to a 4AFC task (Experiment 2). In Experiments 3 to 5, we added a compound decision component (via the inclusion of target-absent trials and a "not present" response option).

In each experiment, we manipulated the similarity between the target or target-replacement and the lure(s) to vary the difference in evidence strengths between the strongest and weakest items in the array.[1] Systematically varying the similarity between two faces produces some practical challenges, as faces are complex, multidimensional stimuli. To achieve tight experimental control, we created computer generated faces using FaceGen

Modeller (Singular Inversions Inc.). These faces are three-dimensional, are suitably realistic and they have been used in many investigations of face perception and recognition (e.g., Chen, Yang, Wang, & Fang, 2010; Papesh & Goldinger, 2010; Potter & Corneille, 2008; Xu, Yue, Lescroart, Biederman, & Kim, 2009). The software includes a "Genetic" function, which generates sibling faces for any referent face. Each of the siblings is allowed to vary by a specified amount from the referent along multiple dimensions (full details are given in the Experiment 1 Method). We used this method to create a series of siblings for each target face (and for each target-replacement) that varied along a continuum of perceived similarity (see Figure 1 for examples).

Experiment 1

In Experiment 1, we used a simple, 2AFC task. The purpose of Experiment 1 was to replicate the standard findings predicted by sequential sampling models, as a test of our stimuli and experimental paradigm. Specifically, we predicted that, when the target is correctly identified: 1) Mean confidence would increase as the lure became less similar to the target; and 2) Mean response latencies would decrease as the lure became less similar to the target. Furthermore, when target-lure similarity was held constant, we predicted that: 3) Confidence and accuracy would be positively related; 4) Response latencies and accuracy would be negatively related; and 5) Response latencies and confidence would be negatively related.

Method

Participants and Design

Participants were recruited from an email list of people who had expressed interest in taking part in research. Most of the people on the list were undergraduates, though the list also included postgraduates and university staff. Twenty-four participants with a mean age of 25.4 years (*SD* = 5.4 years) took part in Experiment 1, of whom 70.8% were female.

G*Power was used to determine a sample size. We assumed that there would be a reasonably strong correlation amongst the repeated measures in each of the analyses (r = .60). We thus determined that a sample of 19 would be sufficient to detect a medium effect with 80% power.

Experiment 1 was a single factor (target-lure similarity) repeated measures design, with four levels: High, Medium-High, Medium-Low, and Low. Note that we use these terms for ease of interpretation, though, as can be seen from Figure 1, even the "low similarity" lures were reasonably similar to the target (i.e., they could not reasonably be considered "duds", cf. Charman et al., 2011). The dependent measures were decision accuracy, confidence, and response latencies.

Materials and Apparatus

The experiments were created using E-Prime 2.0 (Psychology Software Tools, Inc.). The stimuli were digital faces created using the software FaceGen Modeller 3.5 (Singular Inversions, Inc.). All faces were designed to appear male, Caucasian, and between 20 and 40 years of age. To make the faces appear more lifelike, skin texturing was added using the Texture Overlay function of the software. None of the faces included hair.

To create the stimuli, 96 target faces were created using the random face generator in FaceGen Modeller. If a randomly generated face appeared somewhat atypical, then the structure, skin coloring, and/or the asymmetry of the face was adjusted toward the typical face as appropriate. From each of these 96 identities, siblings were created using the Genetic function of the software. This function creates an array of eight faces that are similar to the target face, but which have freedom to vary along multiple dimensions. The user is able to specify by how much the sibling faces are allowed to differ from the target face; a randomness value of 0 would create a face identical to the target; a randomness value of 1 would create a face that is free to vary completely randomly on all dimensions, and so would

bear little resemblance to the target. We chose a randomness value of .30 for the creation of our stimuli. From each array of siblings, one face was randomly selected. Once again, if a face had become somewhat atypical in appearance, we adjusted the skin tone, structure, or asymmetry, as appropriate. That face then became the referent face for the next iteration of the process. The process was repeated seven times, creating a chain of seven generations of sibling faces that were increasingly dissimilar to the target face. Only generations 1 (High similarity), 3 (Medium-High similarity), 5 (Medium-Low similarity) and 7 (Low similarity) were used in the experiments. See Figure 1, top row, for an example of a target and its four siblings. We conducted two pilot tests to determine whether our manipulation was successful in affecting perceptual similarity, and to ensure that the low-similarity siblings were no more distinctive than the targets. Full details of the pilot experiments can be found in the Supplemental Materials. Following these pilot tests, we ended up with 64 sets of stimuli, which were used in all of the experiments.

Procedure

Each trial began with a target face presented in the centre of the screen for 500 ms, at a size of 400 × 400 pixels. A visual mask consisting of patches of varying skin tones was then presented for 2000 ms to minimize reliance on low level features of the stimulus, such as skin tone. A two-alternative choice array was then presented, with each face shown at a size of 300 × 300 pixels. The two faces were presented side by side, and the position of the target was counterbalanced across trials. At the top of the screen was the question "Which face did you see?" Beneath each image was a number corresponding to the key that the participant should use to choose that image. The images remained on screen until a response was made.

Following a recognition decision, the participants were asked to rate their confidence in their decision from 0% to 100%. Prior to beginning the experiment, they were told to use 0% if they were guessing and to use 100% if they were completely certain, and to use the

whole range of numbers between. The images remained on screen while the confidence judgment was made and the chosen face was highlighted with a red border. Participants typed their response in a text box and pressed ENTER to submit their confidence rating. A blank screen was shown for 1000 ms before the next trial began.

The participants completed 64 trials in two blocks. The order of the trials within a block was randomized for each participant, and the order of the blocks was counterbalanced across participants.

Results

Participants were excluded on the basis of two criteria. First, participants were excluded if they failed to provide a valid confidence rating on more than 25% of the trials. One participant met this exclusion criterion. Second, participants were excluded if they used the confidence scale very narrowly or inappropriately. One participant was excluded as he provided confidence ratings of 100% on all but two trials, and one further participant was excluded for providing confidence judgments in the range 3 to 9, suggesting a misunderstanding of the confidence scale. Thus, the final sample included 21 participants. For all post-hoc comparisons, a Bonferroni-corrected α of .008 was applied.

Before we addressed our main hypotheses, we checked whether our manipulation of target-lure similarity significantly affected discriminability. Discriminability (d') was estimated using the formula for n-alternative forced-choice tasks provided by Alexander (1990). Descriptive statistics are shown in Table 1, and a full breakdown of the proportions of each decision type can be found in the Supplemental Materials (Table S1). A one-way ANOVA showed that mean discriminability was, indeed, affected by similarity, $F(3, 60) = 13.85$, $p < .001$, $\eta_p^2 = .41$, 90% CI [.20, .53].[2] Post-hoc comparisons revealed that discriminability increased from High to Medium-High similarity trials, $t(21) = 3.56$, $p = .002$,

d = 0.75, 95% CI [0.26, 1.23], and from Medium-High to Low similarity trials, $t(18) = 3.26$, $p = .004$, d = 0.68, 95% CI [0.21, 1.16].

Our first main hypothesis was that, when a participant correctly identified a target, mean confidence would be higher under conditions of lower target-lure similarity. Thus, these analyses included only correct target identifications, excluding those in which the confidence rating was missing or invalid. A consequence of partitioning data by response type is that each participant provides a different number of trials to the analyses; in such situations, it is appropriate to use mixed effects models that allow the inclusion of participant as a random effect (Wright, Horry, & Skagerberg, 2009; Wright & London, 2009). To test our hypothesis, we created two regression models, with confidence as the outcome variable, and compared their fit. The only predictor in the baseline model was the random intercept for participant; in the comparison model, the fixed effect of similarity (coded as a categorical factor) was entered. The difference in model fit was statistically significant, χ^2 (3, N = 1059) = 94.73, $p < .001$, indicating that confidence varied systematically with similarity. A plot of means is shown in Figure 2A, and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S2). Mean confidence significantly increased from High to Medium-High similarity trials (95% CI_{diff} [6.25, 14.28]), and from Medium-High to Medium-Low similarity trials (95% CI_{diff} [2.73, 10.37]).

Our second hypothesis was that response latencies for correct target identifications would be shorter under conditions of lower target-lure similarity. Trials were removed if the latency was more than two standard deviations from the participant's condition mean, or if the latency was less than 500 ms. Using these criteria, 49 trials (4.55% of total correct identifications) were removed[3]. We used the same mixed-effects modelling approach as that described above for confidence ratings. Prior to analyses, response latencies were log-transformed to reduce skew. Figure 2B shows a plot of back-transformed means. Similarity

was a significant predictor of response latency, χ^2 (3, N = 1027) = 98.96, p < .001. Response latencies significantly decreased from High to Medium-High, and from Medium-High to Medium-Low trials.

Next, we explored the confidence-accuracy relationship. For each participant, at each level of similarity, we calculated Goodman Kruskal gamma (*G*). The mean *G* coefficients were then compared to zero in a series of one-sample *t* tests. At all levels of similarity, the mean *G* coefficient was positive and significantly greater than zero: High similarity, M = .34, SD = .36, t(20) = 4.42, p < .001, 95% CI_{diff} [.18, .51]; Medium-High similarity, M = .48, SD = .47, t(19) = 4.59, p < .001, 95% CI_{diff} [.26, .70]; Medium-Low similarity, M = .55, SD = .54, t(17) = 4.25, p = .001, 95% CI_{diff} [.27, .82]; Low similarity, M = .67, SD = .36, t(9) = 5.90, p < .001, 95% CI_{diff} [.41, .93].[4] Thus, as predicted, confidence and accuracy were positively related, when target-lure similarity was held constant.

We also predicted that response latency and accuracy would be negatively related. For each participant, at each level of similarity, we calculated a point-biserial correlation between log-transformed response latencies (to reduce skew) and accuracy. The mean correlation coefficients were then compared to zero in a series of one-sample *t* tests. At all levels of similarity, (log) response latency and accuracy were negatively related: Medium-High similarity, M = -.26, SD = .27, t(19) = 4.40, p < .001, 95% CI_{diff} [-.39, -.14]; Medium-Low similarity, M = -.28, SD = .39, t(18) = 4.25, p < .001, 95% CI_{diff} [-.42, -.14]; Low similarity, M = -.35, SD = .29, t(9) = 3.90, p = .004, 95% CI_{diff} [-.56, -.15], though the mean coefficient was not significantly different from zero in the High similarity condition, M = -.11, SD = .31, t(20) = 1.64, p = .12, 95% CI_{diff} [-.25, .03].

Finally, we predicted that we would see a negative relationship between confidence and (log) response latency. For each participant, at each level of confidence, we calculated Pearson's *r*. The mean coefficients were then compared to zero in a series of one-sample *t*

tests. At all levels of similarity, confidence and (log) response latency were negatively correlated: High similarity, M = -.40, SD = .26, $t(20) = 6.98$, $p < .001$, 95% CI_{diff} [-.52, -.28]; Medium-High similarity, M = -.44, SD = .26, $t(20) = 7.85$, $p < .001$, 95% CI_{diff} [-.55, -.32]; Medium-Low similarity, M = -.43, SD = .32, $t(19) = 6.08$, $p < .001$, 95% CI_{diff} [-.58, -.28]; Low similarity, M = -.47, SD = .29, $t(20) = 7.44$, $p < .001$, 95% CI_{diff} [-.61, -.34].

Discussion

In Experiment 1, we replicated the standard effects of target-lure similarity on mean confidence and mean response latency. We also replicated the standard relationships between accuracy, confidence, and response latency. This pattern of results is entirely consistent with a range of sequential sampling models of decision-making, in which evidence accumulates over time until a decision can be reached, and in which confidence is scaled from the extent to which the evidence favours the chosen option over the unchosen option. In the following experiments, we increase the complexity of the decision task, first by increasing the number of alternatives from two to four, and then by introducing a compound decision component.

Experiment 2

In Experiment 2, we added one layer of complexity to the task by increasing the number of alternatives from two to four. In all trials, the choice array included the target and two high-similarity lures. The similarity of the weakest lure was varied in the same way as in Experiment 1. Thus, we selectively manipulated the strength of the weakest item in the array. We hypothesised that if confidence judgments index the balance-of-evidence between all items in the choice set (or, indeed, if the diffusion variable reflects the evidence associated with all items), then mean confidence in correct target identifications would increase, and mean response latency would decrease, as the similarity of the weakest item to the target decreased. We also expected to see the same relationships between confidence, accuracy, and response latency as in Experiment 1.

Method

Participants and Design

Twenty-five participants took part in Experiment 2, of whom 60.0% were female. The mean age was 28.4 years (*SD* = 10.3). The participants were recruited from the same email list used in Experiment 1, though no participants had taken part in the previous experiment.

Experiment 2 was a single factor (similarity of worst-matching lure) repeated measures design, with four levels: High, Medium-High, Medium-Low, and Low. The dependent variables were decision accuracy, confidence, and response latencies.

Materials

Following the procedure described for Experiment 1, we created two additional sets of siblings for each of the 64 target faces so that we could create four-alternative choice arrays. The materials for Experiment 2 included 64 target faces and 768 lures, though only 192 of the lures were seen by any participant.

Procedure

The procedure was similar to Experiment 1, except that the number of faces in the choice array was increased from two to four. The faces were presented in two rows of two. Of the three distractor faces, two were always highly similar to the target, leaving the third distractor free to vary in similarity (High, Medium-High, Medium-Low, Low).

Results

We used the same exclusion criteria as for the previous experiments. One participant was excluded because he provided confidence ratings of 100% on all trials. The final analysis included 24 participants. For all post-hoc comparisons, a Bonferroni-corrected α of .008 was applied.

Once again, discriminability was estimated using Alexander's (1990) formula for *n*-alternative forced-choice tasks. Descriptive statistics are shown in Table 1, and a full

breakdown of responses can be found in the Supplemental Materials (Table S1). A one-way ANOVA revealed that discriminability was significantly affected by similarity, $F(3, 69) = 3.33$, $p = .03$, $\eta_p^2 = .13$, 90% CI [.01, .22]. Post-hoc tests indicated an increase in accuracy from High to Medium-Low similarity trials, $t(23) = 2.80$, $p = .01$, $d = 0.55$, 95% CI [0.12, 0.98], and from High to Low similarity trials, $t(23) = 2.92$, $p = .008$, $d = 0.58$, 95% CI [0.14, 1.01], though neither of these met the Bonferroni-corrected α level of .008.

We tested our hypothesis that confidence in correct target identifications would increase as the similarity of the weakest item to the target decreased. We used the same analysis as in Experiment 1, namely a mixed-effects regression model with a random intercept for participants. As predicted, adding similarity as a categorical fixed effect significantly improved model fit, $\chi^2(3, N = 716) = 16.11$, $p = .001$. The means are plotted in Figure 3A , and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S3). Confidence was significantly lower in High similarity trials than in the Medium-High (95% CI_{diff} [3.72, 12.81]), or Low similarity (95% CI_{diff} [3.30, 12.22]) trials. There were no significant differences between the Medium-High, Medium-Low, and Low similarity conditions.

Next, we tested our hypothesis that response latencies for correct identifications would decrease as the similarity of the weakest lure to the target decreased. Once again, we used a mixed-effects regression model, including participant as a random intercept. As in Experiment 1, trials were excluded if the response latency was less than 500 ms or more than 2 SDs from the participant's condition mean. Using these criteria, 33 trials (4.56% of all correct identifications) were removed. Response latencies were then log-transformed to reduce skew. Unexpectedly, adding similarity as a fixed effect did not improve the fit of the model, $\chi^2(3, N = 691) = 6.05$, $p = .11$. Means are plotted in Figure 3B. However, when we

repeated the analysis with the outliers included, similarity did significantly improve the model fit, $\chi^2(3, N = 724) = 12.74$, $p = .005$ (see Supplemental Materials).

As in Experiment 1, we explored the relationship between confidence and accuracy, by calculating *G* for each participant, at each level of similarity. The mean coefficients were significantly different from zero only in the Medium-High similarity condition, $M = .28$, $SD = .35$, $t(23) = 3.96$, $p = .001$, 95% CI_{diff} [.13, .43], and in the Low similarity condition, $M = .25$, $SD = .47$, $t(23) = 2.59$, $p = .02$, 95% CI_{diff} [.05, .45]. In the remaining conditions, the coefficients were not significantly different from zero: High similarity, $M = .06$, $SD = .47$, $t(23) = 0.68$, $p = .51$, 95% CI_{diff} [-.13, .26]; Medium-Low similarity, $M = .03$, $SD = .47$, $t(23) = 0.32$, $p = .75$, 95% CI_{diff} [-.17, .23].

Correlations between log-transformed response latency and accuracy were explored next. The mean coefficient was significantly only in the Low similarity condition, $M = -.12$, $SD = .24$, $t(23) = 2.54$, $p = .02$, 95% CI_{diff} [-.22, -.03]. In the remaining conditions, the coefficients were not significantly different from zero: High similarity, $M = -.11$, $SD = .31$, $t(23) = 1.75$, $p = .09$, 95% CI_{diff} [-.24, .02]; Medium-High similarity, $M = -.08$, $SD = .27$, $t(23) = 1.47$, $p = .15$, 95% CI_{diff} [-.19, .03]; Medium-Low similarity, $M = -.02$, $SD = .29$, $t(23) = 0.39$, $p = .70$, 95% CI_{diff} [-.15, .10].

Finally, we examined correlations between confidence and log-transformed response latency. The mean coefficients were significantly below zero in the High, $M = -.15$, $SD = .32$, $t(23) = 2.26$, $p = .03$, 95% CI_{diff} [-.28, -.01], and Medium-High conditions, $M = -.18$, $SD = .32$, $t(23) = 2.78$, $p = .01$, 95% CI_{diff} [-.32, -.05]. In the remaining conditions, the coefficients were not significantly different from zero: Medium-Low similarity, $M = -.09$, $SD = .33$, $t(23) = 1.25$, $p = .22$, 95% CI_{diff} [-.23, .06]; Low similarity, $M = -.13$, $SD = .34$, $t(23) = 1.91$, $p = .07$, 95% CI_{diff} [-.27, .01].

Discussion

In Experiment 2, participants completed a 4AFC task, in which the similarity between the target and the weakest item was systematically manipulated. As predicted from a sequential sampling perspective, mean confidence ratings in correct target identifications increased as the similarity between the target and the weakest item decreased. These findings suggest that, in a *n*AFC task, confidence indexes the relative strength of evidence for the chosen item over all unchosen items. The mechanism for such an effect could be through assessing the *balance-of-evidence* across multiple counters (Usher & McClelland, 2001; Van Zandt, 2000), or it is possible that a single diffusion variable combines information from multiple items, and confidence ratings are scaled from this variable at the time of the confidence judgment (Pleskac & Busemeyer, 2010).

The standard relationships between confidence, response latency, and decision accuracy were replicated. However, the relationships were less strong than in Experiment 1. This may be in part due to the increased variability in response latency, and the decrease in accuracy. Though accuracy was above floor in even the most difficult condition, it is clear that the task was a reasonably difficult one, with two highly similar lures in each array. Nonetheless, the relationships that were observed lend further support to a sequential sampling perspective on *n*AFC judgments.

Experiment 3

We turn our attention now to compound decisions, in which a decision-maker must decide whether a target item is present within a choice array, and if so, which item corresponds to that target. To reduce the complexity of the task, we used a 2-alternative compound task. In target-present trials, the target appeared alongside a single lure, whose similarity to the target was systematically varied. In target-absent trials, the target was replaced by a target-replacement, who appeared alongside a lure that varied in its similarity to

the target-replacement in the same way as the lures in the target-present trials. Thus, the perceptual similarity relationships between items were maintained in the target-present and –absent trials.

For correct target identifications, we predicted that we would see the same pattern of results as in forced-choice decisions; namely, mean confidence would increase, and mean response latency would decrease, as target-lure similarity decreased. We also predicted that, across all positive decisions, we would see the same relationships between confidence, response latency, and decision accuracy as in Experiments 1 and 2.

For correct rejections of target-absent lineups, we predicted that we would not observe any influence of target-lure similarity on confidence. We reasoned that confidence judgments might be scaled in a different way (i.e., the distance between the evidence associated with the best match, and the positive decision threshold), which would not lend itself to a relative confidence judgment. Based on findings from the face recognition and eyewitness identification literatures (e.g., Brewer & Wells, 2006; Horry et al., 2012; Weber & Brewer, 2004), we did not expect to observe a strong relationship between confidence and accuracy across all negative decisions. We had no clear predictions about the relationship between response latency and accuracy or confidence.

Method

Participants and Design

Twenty-five participants took part in Experiment 2, of whom 84.0% were female. The mean age was 22.5 years (*SD* = 4.07). The participants were recruited from the same source as in the previous experiments; no participant had taken part in a previous experiment.

Experiment 2 followed a 2 (target-presence) × 4 (similarity: High, Medium-High, Medium-Low, Low) repeated measures design.

Materials

For target-present trials, the materials were identical to those used in Experiment 1. For target-absent trials, 64 target-replacements were created, each of which were seven generations removed from its respective target. Thus, the target-replacement was equivalent in similarity to the target as a "Low similarity" lure. The target-replacement was a unique identity, who never appeared as a lure in any target-present trials. Furthermore, for each target-replacement, 4 siblings were created, one at each level of similarity. Thus, in target-absent trials, we manipulated the similarity between the target-replacement and the lure, rather than the similarity between the target and the lure.

Procedure

The procedure was similar to Experiment 1, with the following exceptions. First, half of the trials were target-absent. In place of the target, the array included a target-replacement alongside the lure. The participants could choose either of the items in the array or they could make a "not present" response. Following a "not present" response, all of the images remained on screen while the participant provided his or her confidence judgment. The participants were told that in some of the trials, the target would not be present, but they were not informed of the ratio of target-present to target-absent trials.

Results

One participant was excluded because she only ever provided confidence ratings of 0% or 100%. One further participant was excluded because she provided confidence ratings from 0 to 9, suggesting a misunderstanding of the confidence scale. Thus, the analyses included 23 participants.

Discriminability and choosing

With a compound decision, estimating discriminability is not straightforward. We took the approach here of breaking the decision down into its constituent components: detection and identification. Detection refers to the participant's ability to distinguish

between target-present and target-absent trials. Any positive decision on a target-present trial is coded as a hit (regardless of the accuracy of that decision), and any positive decision on a target-absent trial is coded as a false alarm. *d'*-detection is then calculated as *z*(HR) – *z*(FAR), where HR is the proportion of target-present trials that result in a positive decision, and FAR is the proportion of target-absent trials that result in a positive decision. Second, identification refers to the participant's ability to identify the target, given that a positive decision was made on a target-present trial. *d'*-identification is then estimated using the same formula as for an *n*-alternative forced-choice task (Alexander, 1990), but only target-present trials are included, with negative decisions excluded. As a proxy for response bias, we examined choosing rates, collapsing across target-presence. Descriptive statistics are shown in Table 2, and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S1). For all post-hoc comparisons, a Bonferroni-corrected α of .008 was adopted.

Similarity had a statistically significant effect on *d'*-identification, $F(3, 66) = 9.51$, $p < .001$, $\eta_p^2 = .30$, 90% CI [.13, .41]. Post-hoc comparisons revealed that *d'*-identification significantly increased from High to Medium-Low similarity trials, $t(22) = 6.09$, $p < .001$, $d = 1.21$, 95% CI [0.62, 1.81], and from Medium-High to Medium-Low similarity trials, $t(22) = 3.20$, $p = .004$, $d = 0.64$, 95% CI [0.15, 1.13]. Thus, unsurprisingly, the participant's ability to distinguish between the target and the lure in a target-present array was affected by the similarity between the target and the lure.

Neither *d'*-detection, $F(3, 66) = 0.42$, $p = .74$, $\eta_p^2 = .02$, 90% CI [.00, .06], nor choosing rates were significantly affected by similarity, $F(3, 66) = 1.99$, $p = .12$, $\eta_p^2 = .08$, 90% CI [.00, .17]. Thus, similarity did not seem to influence willingness to choose, nor the ability to discriminate between target-present and target-absent trials.

Positive decisions

The following analyses focus on confidence and response latency for correct target identifications. As in Experiments 1 and 2, we used a mixed effects regression approach, allowing random intercepts for participant. For confidence, adding similarity as a fixed, categorical factor significantly improved the fit of the model, $\chi^2(3, N = 561) = 12.29$, $p < .001$. The means are plotted in Figure 4A. Confidence was significantly lower in High similarity trials than in the Low similarity trials (95% CI_{diff} [3.48, 13.06]), and in the Medium-High similarity trials than in the Low similarity trials, (95% CI_{diff} [1.76, 10.66]).

As in Experiments 1 and 2, trials with response latencies of less than 500 ms or more than 2 SDs from the participant's standard deviation were excluded. Using these criteria, 24 trials (4.21% of all correct target-present decisions) were removed. Response latencies were log transformed to reduce skew. Adding similarity as a fixed, categorical predictor significantly improved model fit, $\chi^2(3, N = 546) = 42.21$, $p < .001$. Back-transformed means are shown in Figure 4B, and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S4). Response latencies significantly decreased from High to Medium-High similarity trials, and from Medium-High to Medium-Low similarity trials.

For our analyses of the relationships between confidence, accuracy, and response latency, we included all positive decisions, whether from target-present or target-absent trials. First, we calculated the confidence-accuracy relationship for positive decisions at each level of similarity. Goodman-Kruskal's gamma (G) was calculated for each participant, and the mean coefficients were compared to zero in a series of one-sample t tests. At each level of similarity, the mean coefficient was significantly greater than zero: High similarity, $M = .28$, $SD = .44$, $t(21) = 3.01$, $p = .007$, 95% CI_{diff} [.09, .48]; Medium-High similarity, $M = .36$, $SD = .36$, $t(22) = 4.81$, $p < .001$, 95% CI_{diff} [.20, .51]; Medium-Low similarity, $M = .52$, $SD = .38$,

$t(21) = 6.31$, $p < .001$, 95% CI_{diff} [.35, .69]; Low similarity, $M = .58$, $SD = .34$, $t(21) = 8.13$, $p < .001$, 95% CI_{diff} [.43, .73].

To examine the relationship between response times and accuracy, we calculated a point-biserial correlation between accuracy and log-transformed response latency for each participant, at each level of similarity. The mean coefficients were then compared to zero in a series of one-sample *t* tests. As predicted, all coefficients were significantly below zero: High similarity, $M = -.20$, $SD = .23$, $t(21) = 4.04$, $p < .001$, 95% CI_{diff} [-.30, -.10]; Medium-High similarity, $M = -.28$, $SD = .31$, $t(22) = 4.20$, $p < .001$, 95% CI_{diff} [-.41, -.14]; Medium-Low similarity, $M = -.35$, $SD = .35$, $t(22) = 4.87$, $p < .001$, 95% CI_{diff} [-.50, -.20]; Low similarity, $M = -.45$, $SD = .25$, $t(22) = 8.62$, $p < .001$, 95% CI_{diff} [-.56, -.34].

Finally, we examined the relationship between (log) response latency and confidence ratings. Once again, the mean correlation coefficients were compared to zero in a series of one-sample *t* tests. At all levels of similarity, confidence and (log) response latency were negatively correlated: High similarity, $M = -.46$, $SD = .26$, $t(22) = 8.54$, $p < .001$, 95% CI_{diff} [-.57, -.35]; Medium-High similarity, $M = -.47$, $SD = .24$, $t(22) = 9.45$, $p < .001$, 95% CI_{diff} [-.57, -.37]; Medium-Low similarity, $M = -.46$, $SD = .19$, $t(22) = 11.51$, $p < .001$, 95% CI_{diff} [-.55, -.38]; Low similarity, $M = -.38$, $SD = .37$, $t(22) = 5.04$, $p < .001$, 95% CI_{diff} [-.54, -.23].

Negative decisions

The following analyses examine confidence and response latency in correct negative decisions (i.e., rejections of lineups when the target was absent). We used the same mixed-effects regression analyses as for positive decisions, with random intercepts allowed for participants. We first examined the effect of similarity on confidence. We excluded trials in which the confidence rating was missing or invalid. Adding similarity to the regression model did not significantly improve fit, $\chi^2(3, N = 359) = 1.53$, $p = .68$. Means are shown in Figure

5A, and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S5).

Next, we examined the effect of similarity on log-transformed response latency. We removed trials in which the response latency was less than 500 ms or more than 2 SDs from the participant's condition mean. In total, 13 trials (3.59% of correct negative decisions) were removed. Adding similarity to the regression model significantly improved model fit, $\chi^2(3, N = 349) = 16.10$, $p = .001$. Back-transformed means are shown in Figure 5B. Log-transformed response latency significantly decreased from High to Medium-Low similarity trials, and from High to Low similarity trials. However, when the analyses were repeated with the outliers included, the effect was no longer significant, $\chi^2(3, N = 362) = 5.10$, $p = .17$ (see Supplemental Materials). Therefore, this effect may not be robust, and may have been produced by a small number of outlying trials.

To examine the relationships between confidence, accuracy and response latency, we included all negative decisions, whether from target-absent or target-present trials. We were unable to examine these correlations separately at each level of similarity, as there were too few trials per cell for many participants. Consequently, for all correlations, we collapsed across similarity. Confidence and accuracy were positively related; the mean Goodman-Kruskal G was significantly higher than zero, $M = .54$, $SD = .62$, $t(17) = 3.69$, $p = .002$, 95% CI_{diff} [.23, .85]. Log-response times and accuracy were negatively related; the mean point-biserial correlation was significantly below zero, $M = -.16$, $SD = .26$, $t(17) = 2.66$, $p = .02$, 95% CI_{diff} [-.29, .-.03]. Confidence and (log) response latency were negatively correlated; the mean correlation coefficient was significantly below zero, $M = -.44$, $SD = .32$, $t(20) = 6.33$, $p < .001$, 95% CI_{diff} [-.58, .-.29].

Discussion

In Experiment 3, participants completed a 2-alternative compound task. When the participant correctly identified the target, we observed the same pattern of results as in Experiments 1 and 2: confidence increased, and response latency decreased, as target-lure similarity decreased. Across all positive decisions, we observed the standard pattern of relationships between confidence, response latency, and decision accuracy. Thus, when a decision-maker chooses an item in a compound decision task, the patterns of confidence and response latency map neatly onto the predictions we had derived from sequential sampling models.

As predicted, when the decision-maker correctly rejected a target-absent array, the similarity between the target-replacement and the lure did not significantly affect confidence. It may be that the framing of the question (how confident are you in your decision) forces the decision-maker to evaluate the evidence differently, focusing more on the absolute match of the best (though unchosen) item. It is also possible that negative decisions serve as a "catch-all" response when the decision-maker is unsure – although that seems unlikely in this context, as mean confidence in negative decisions was similar to, if not higher than mean confidence in positive decisions. Somewhat surprisingly, we did see a positive confidence-accuracy relationship for negative decisions, though we had to collapse across similarity levels to ensure a sufficient number of trials per participant for stable estimates. Thus, for negative decisions, we can make no strong claims about these relationships while holding task difficulty constant.

Experiment 4

In Experiment 3, we demonstrated that, in a compound decision task, patterns of confidence and response latency are consistent with predictions derived from a sequential sampling framework, at least when the decision-maker chooses an item from the array. In

Experiment 4, we add one additional layer of complexity to the task, by increasing the number of alternatives from two to four. As in Experiment 2, participants were presented with four items, and the similarity between the target (or target-replacement, in target-absent trials) and the weakest lure was systematically manipulated. For correct target identifications, we expected to replicate the findings of Experiment 2; for correct rejections, we expected to observe no significant effect of target-lure (or target-replacement-lure) similarity on confidence.

Method

Participants and Design

Twenty-three participants took part in Experiment 4, of whom 56.5% were female. The mean age was 23.0 years (*SD* = 6.40). The participants were recruited from the same source as in the previous experiments; no participant had taken part in a previous experiment.

Experiment 4 followed a 2 (target-presence) × 4 (similarity: High, Medium-High, Medium-Low, Low) repeated measures design.

Materials

For target-present trials, the materials were identical to those used in Experiment 3. For target-absent trials, the target-replacements were the same as those used in Experiment 2. The lures used in the target-absent trials included those used in Experiment 2, and an additional eight distractors per target-replacement (two at each level of similarity), to create four-alternative arrays.

As in Experiment 3, two of the distractors in each array were highly similar to the target/target-replacement. Only one of the distractors varied in similarity to the target/target-replacement.

Procedure

The procedure was similar to Experiment 2, except that the choice arrays included four faces instead of two. The items were arranged in a two-by-two grid, and the position of the target/target-replacement was counterbalanced within participants.

Results

No participants met the exclusion criteria in Experiment 4. Thus, data from all participants were analyzed.

Discriminability and choosing

Discriminability was estimated in the same way as in Experiment 2. *d'*-identification, *d'*-detection, and choosing rates were analyzed in separate one-way ANOVAs. Descriptive statistics are shown in Table 2, and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S1).

Similarity did not significantly affect *d'*-identification, $F(3, 66) = 0.98$, $p = .41$, $\eta_p^2 = .04$, 90% CI [.00, .10], *d'*-detection, $F(3, 66) = 0.18$, $p = .91$, $\eta_p^2 = .01$, 90% CI [.00, .03], or choosing rates, $F(3, 66) = 0.12$, $p = .95$, $\eta_p^2 = .01$, 90% CI [.00, .02]. Thus, the similarity between the target, or target-replacement, and the worst-matching lure did not observably affect discriminability or choosing.

Positive identification decisions

As in Experiment 3, we examined the effect of similarity on confidence and response latency when the target was correctly identified. For the analysis of confidence, trials were excluded if the confidence rating was missing or invalid; for the analysis of response latency, trials were removed if the response latency was less than 500 ms or more than 2 SDs from the participant's condition mean. In total, 11 trials (3.54% of correct positive decisions) were removed. Response latencies were then log-transformed prior to analysis. As in the previous experiments, we used a mixed-effects regression approach, with random intercepts for

participant. Contrary to predictions, adding similarity did not significantly improve model fit for confidence, $\chi^2(3, N = 307) = 0.84$, $p = .84$, or response latency, $\chi^2(3, N = 300) = 5.98$, $p = .11$. Mean confidence and back-transformed response latency are shown in Figure 6. A full breakdown of confidence frequencies can be found in the Supplemental Materials (Table S6).

Next, we explored the relationships between confidence, accuracy, and response latency. We included all positive decisions, whether from target-present or target-absent trials. For the confidence-accuracy relationship, at all levels of similarity, mean *G* was positive; however, the coefficients were only significantly above zero in the Medium-High and Low similarity conditions: High similarity, $M = .22$, $SD = .52$, $t(21) = 1.94$, $p = .07$, 95% CI_{diff} [-.02, .45]; Medium-High similarity, $M = .35$, $SD = .57$, $t(21) = 2.94$, $p = .008$, 95% CI_{diff} [.10, .61]; Medium-Low similarity, $M = .17$, $SD = .56$, $t(22) = 1.47$, $p = .16$, 95% CI_{diff} [-.07, .42]; Low similarity, $M = .22$, $SD = .47$, $t(21) = 2.14$, $p = .04$, 95% CI_{diff} [.01, .43].

Unexpectedly, log-response latency was not significantly correlated with accuracy at any level of similarity: High similarity, $M = .06$, $SD = .36$, $t(22) = 0.85$, $p = .40$, 95% CI_{diff} [-.09, .22]; Medium-High similarity, $M = -.14$, $SD = .34$, $t(20) = 1.96$, $p = .06$, 95% CI_{diff} [-.30, .01]; Medium-Low similarity, $M = -.13$, $SD = .37$, $t(22) = 1.68$, $p = .11$, 95% CI_{diff} [-.29, .03]; Low similarity, $M = -.01$, $SD = .36$, $t(21) = 0.09$, $p = .93$, 95% CI_{diff} [-.17, .15].

Finally, we examined the relationship between confidence and (log) response latency. At all levels of similarity, the mean correlation coefficients were negative, though they differed significantly from zero only in the High, Medium-High, and Medium-Low similarity conditions: High similarity, $M = -.20$, $SD = .42$, $t(22) = 2.27$, $p = .03$, 95% CI_{diff} [-.38, -.02]; Medium-High similarity, $M = -.29$, $SD = .27$, $t(22) = 5.08$, $p < .001$, 95% CI_{diff} [-.40, -.17]; Medium-Low similarity, $M = -.19$, $SD = .31$, $t(22) = 2.95$, $p = .007$, 95% CI_{diff} [-.32, -.06]; Low similarity, $M = -.11$, $SD = .31$, $t(22) = 1.78$, $p = .09$, 95% CI_{diff} [-.25, .02].

Negative identification decisions

Next, we examined the effect of similarity on confidence and response latency in correct negative decisions. For the analysis of confidence, trials were excluded if the confidence rating was missing or invalid; for the analysis of response latency, trials were removed if the response latency was less than 500 ms or more than 2 SDs from the participant's condition mean. In total, 4 correct negative decisions (1.72% of all correct negative decisions) were removed. Response latencies were then log-transformed prior to analysis. As in the previous experiments, we used a mixed-effects regression approach, with random intercepts for participant. Adding similarity did not significantly improve model fit for confidence, $\chi^2(3, N = 227) = 0.84$, $p = .84$, or response latency, $\chi^2(3, N = 229) = 3.71$, $p = .29$. Mean confidence and back-transformed response latency are shown in Figure 7. A full breakdown of confidence frequencies can be found in the Supplemental Materials (Table S7).

Finally, we explored the relationships between confidence, accuracy, and response latency in negative decisions. As in Experiment 3, we collapsed across levels of similarity to ensure sufficient numbers of decisions for stable correlations. Confidence and accuracy were positively related; the mean Goodman-Kruskal *G* was significantly greater than zero, $M = .38$, $SD = .51$, $t(16) = 3.09$, $p = .007$, 95% CI_{diff} [.12, .64]. Log-transformed response latency and accuracy were negatively related; the mean correlation coefficient was significantly below zero, $M = -.26$, $SD = .29$, $t(17) = 3.86$, $p = .001$, 95% CI_{diff} [-.41, -.12]. Unexpectedly, confidence and (log) response latency were not significantly correlated, $M = -.11$, $SD = .42$, $t(19) = 1.11$, $p = .28$, 95% CI_{diff} [-.30, .09].

Discussion

In Experiment 4, participants completed a 4-alternative compound decision task. In contrast to Experiment 2, our manipulation of similarity between the target (or target-replacement) and the weakest lure did not significantly influence mean confidence ratings, or

mean response latency for correct identifications. Why did we fail to replicate this finding in Experiment 4? Perhaps the multiple-alternative compound task is just too complex for such effects to hold up. However, this seems unlikely; we found such an effect in a 4AFC task (Experiment 2), and in a compound decision task (Experiment 3), which suggests that neither element of the task individually quashes the effect. A more likely explanation is that the effect size was simply too small to be reliably detected; the weakest item in the array should exert some influence on confidence judgments and response latency, but this pull may be quite weak. In Experiment 5, we address this possibility by manipulating the similarity of the *two* weakest items in the array to the target (or target-replacement).

Experiment 5

In Experiment 5, participants again completed a four-alternative compound decision task. To increase the influence of the weakest items on confidence, we manipulated the similarity of *two* members of the array to the target (or target-replacement). Thus, each array featured a target or target-replacement, one high-similarity lure, and two lures that varied systematically in their similarity to the target (or target-replacement). In addition, to increase reliability of the confidence ratings associated with correct identifications, we increased the proportion of target-present trials from .50 to .66. Participants were not informed of the ratio of target-present to target-absent trials.

Method

Participants and Design

Thirty-two participants took part in Experiment 5, of whom 76.7% were female. The mean age was 29.8 years (*SD* = 12.8 years). The participants were recruited from the same source as in the previous experiments; no participant had taken part in a previous experiment.

Experiment 5 followed a 2 (target-presence) × 4 (similarity: High, Medium-High, Medium-Low, Low) repeated measures design.

Materials

The materials were identical to those used in Experiment 4.

Procedure

The procedure was similar to Experiment 4, with two exceptions: First, the similarity of the *two* worst-matching lures was manipulated. These lures always varied in step, such that they were equivalent in their similarity to the target or target-replacement. Second, the proportion of target-present trials was increased from .50 to .66. The participants were not informed of the proportion of target-present trials.

Results

Two participants were excluded for failing to provide valid confidence judgments on more than 25% of trials. For all analyses, a Bonferroni-corrected α of .008 was applied.

Discriminability and choosing rates

As in Experiments 2 and 4, we analyzed *d'*-identification, *d'*-detection, and choosing rates in separate one-way ANOVAs. Means and standard deviations are shown in Table 1, and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S1).

Similarity had a statistically significant effect on *d'*-identification, $F(3, 87) = 6.63$, $p < .001$, $\eta_p^2 = .19$, 90% CI [.06, .28]. Post-hoc comparisons showed that discriminability increased from High to Medium-Low similarity trials, $t(29) = 3.41$, $p = .002$, $d = 0.61$, 95% CI [0.22, 0.99]. No other comparisons were statistically significant to the level of the Bonferroni-corrected criterion. *d'*-detection was not statistically affected by similarity, $F(3, 87) = 0.85$, $p = .47$, $\eta_p^2 = .03$, 90% CI [.00, .08].

Choosing rates were significantly affected by similarity, $F(3, 87) = 4.53$, $p = .005$, $\eta_p^2 = .14$, 90% CI [.03, .23]. Participants chose more frequently in the High similarity trials than in any other type of trial: High vs. Medium-High, $t(29) = 2.94$, $p = .006$, $d = 0.52$, 95% CI

[0.14, 0.90]; High vs. Medium-Low, $t(29) = 2.97$, $p = .006$, $d = 0.53$, 95% CI [0.15, 0.91]; High vs. Low, $t(29) = 3.21$, $p = .003$, $d = 0.57$, 95% CI [0.18, 0.96]. No other comparisons met the Bonferroni-corrected criterion for statistical significance.

Positive identification decisions

We first examined the effect of similarity on confidence in correct target identifications, using the same mixed-effects regression approach as in the previous experiments. Random intercepts were allowed for participant number, and trials were excluded if the confidence rating was missing or invalid. Adding similarity to the model significantly improved fit, $\chi^2(3, N = 529) = 17.24$, $p < .001$. Means are plotted in Figure 8A, and a full breakdown of response frequencies can be found in the Supplemental Materials (Table S8). Confidence significantly increased from High to Medium-Low similarity trials, (95% CI_{diff} [1.89, 10.86]), High to Low similarity trials, (95% CI_{diff} [4.29, 13.12]), and from Medium-High to Low similarity trials, (95% CI_{diff} [1.54, 10.20]).

Next, we examined the effect of similarity on response latency when the target was correctly identified. Trials were excluded if the response latency was less than 500 ms, or more than 2 SDs from the participant's condition mean. In total, 26 trials (4.76% of all correct positive decisions) were removed. Response latencies were then back-transformed and analysed in a mixed-effects regression, with participant as a random intercept. Adding similarity to the regression model significantly improved model fit, $\chi^2(3, N = 520) = 9.20$, $p = .03$. Back-transformed means are plotted in Figure 8B. Response latency was significantly longer in High and Medium-High similarity trials than in Medium-Low and Low similarity trials.

To explore the confidence-accuracy relationship, we once again calculated Goodman-Kruskal's *G* for each participant, at each level of similarity. The mean coefficients were then compared to zero in a series of one-sample *t* tests. At each level of similarity, confidence and

accuracy were positively related, with mean *G* significantly greater than zero: High similarity, $M = .36$, $SD = .44$, $t(28) = 4.36$, $p < .001$, 95% CI_{diff} [.19, .53]; Medium-High similarity, $M = .23$, $SD = .47$, $t(28) = 2.69$, $p = .01$, 95% CI_{diff} [.06, .41]; Medium-Low similarity, $M = .23$, $SD = .46$, $t(29) = 2.71$, $p = .01$, 95% CI_{diff} [.06, .40]; Low similarity, $M = .37$, $SD = .44$, $t(28) = 4.50$, $p < .001$, 95% CI_{diff} [.20, .54].

Next, we examined the relationship between (log) response latency and accuracy. The mean point-biserial correlation coefficient, at each level of similarity, was compared to zero. At all levels of similarity, the mean correlation was negative and significantly below zero: High similarity, $M = -.23$, $SD = .35$, $t(28) = 3.61$, $p = .001$, 95% CI_{diff} [-.37, -.10]; Medium-High similarity, $M = -.15$, $SD = .33$, $t(28) = 2.43$, $p = .02$, 95% CI_{diff} [-.28, -.02]; Medium-Low similarity, $M = -.21$, $SD = .30$, $t(29) = 3.88$, $p = .001$, 95% CI_{diff} [-.32, -.10]; Low similarity, $M = -.23$, $SD = .33$, $t(28) = 3.86$, $p = .001$, 95% CI_{diff} [-.36, -.11].

Finally, we examined the relationship between confidence and (log) response latency. At each level of similarity, the mean correlation coefficient was negative and significantly below zero: High similarity, $M = -.41$, $SD = .31$, $t(29) = 7.11$, $p < .001$, 95% CI_{diff} [-.52, -.29]; Medium-High similarity, $M = -.38$, $SD = .26$, $t(29) = 8.17$, $p < .001$, 95% CI_{diff} [-.48, -.29]; Medium-Low similarity, $M = -.20$, $SD = .38$, $t(29) = 2.94$, $p = .006$, 95% CI_{diff} [-.34, -.06]; Low similarity, $M = -.32$, $SD = .38$, $t(28) = 4.58$, $p < .001$, 95% CI_{diff} [-.46, -.18].

Negative identification decisions

Finally, we turned our attention to the effect of similarity on confidence and response latency for correct rejections. The same analytic techniques were used as for positive decisions. Using the same cutoff criteria as in all previous experiments, one trial was excluded from the response latency analysis. Adding similarity to the regression model did not significantly improve the fit for confidence, $\chi^2(3, N = 335) = 4.31$, $p = .23$, or for response

latency, $\chi^2(3, N = 337) = 0.49$, $p = .92$. Means are plotted in Figure 9, and a full breakdown of confidence frequencies can be found in the Supplemental Materials (Table S9).

Next, we examined the relationships between confidence, accuracy, and (log) response latency in negative decisions. To ensure sufficient numbers of decisions per cell, we collapsed across similarity. Confidence and accuracy were positively related; the mean *G* coefficient was significantly above zero: $M = .36$, $SD = .40$, $t(26) = 4.63$, $p < .001$, 95% CI_{diff} [.20, .52]. Log response latency and accuracy were negatively related; the mean point-biserial correlation coefficient was significantly below zero: $M = -.18$, $SD = .22$, $t(27) = 4.51$, $p < .001$, 95% CI_{diff} [-.27, -.10]. Confidence and (log) response latency were negatively related also; the mean correlation coefficient was significantly below zero: $M = -.38$, $SD = .36$, $t(26) = 5.47$, $p < .001$, 95% CI_{diff} [-.52, -.24].

Discussion

In Experiment 5, we manipulated the similarity between the target (or target-replacement) and the two weakest items in the array. As predicted, for correct target identifications, mean confidence increased, and mean response latency decreased, as the weakest lures became increasingly dissimilar to the target. These results suggest that, even in complex decision tasks, confidence ratings incorporate information about the relative strength of evidence for the chosen item over the unchosen items. Furthermore, across all positive decisions, we replicated the standard relationships between confidence, accuracy, and response latency. Taken together, these findings suggest that predictions from sequential sampling models of decision-making can be scaled up to understand decision-making in complex, multiple-alternative, compound tasks.

As in Experiments 3 and 4, the similarity between the target-replacement and the lures did not affect confidence or response latency in correct rejections of target-absent arrays. We

consider the theoretical and applied implications of these consistent findings in the General Discussion.

For the first time, in Experiment 5, we observed a significant association between target-lure similarity and choosing rates. Specifically, participants were most likely to choose from a High-similarity array. These results raise the intriguing possibility that in certain situations, target-lure similarity can have paradoxical effects on willingness to choose and on the confidence expressed in those choices. Of course, given that we did not observe any such effect in Experiments 3 or 4, it would be necessary to replicate such findings before drawing any strong conclusions.

General Discussion

Across five experiments, we attempted to answer the question: what information forms the basis of confidence judgments in complex, *n*-alternative, compound decisions? Across five experiments, we manipulated the similarity between the target, or target-replacement, and the lures, systematically testing predictions derived from sequential sampling models to increasingly complex decisions. We began with a simple, 2AFC task, such as has been the focus of much research in cognitive psychology. We replicated the standard pattern of findings that have been successfully explained by sequential sampling models of decision making (e.g., Brown & Heathcote, 2008; Ratcliff & Starns, 2009; Usher & McClelland, 2001). Specifically, we found that mean confidence increased, and mean response latency decreased, as target-lure discriminability decreased. Furthermore, when holding target-lure discriminability constant, confidence was positively related to accuracy, response latency was negatively related to accuracy, and confidence was negatively related to response latency. In the following experiments, we increased the complexity of the decision in two ways: by increasing the number of items in the choice array from two to four (Experiments 2, 4, and 5), and by changing the decision task from forced-choice to

compound (Experiments 3-5). Overall, our results suggested that predictions derived from sequential sampling frameworks can be applied to complex decisions, and that confidence (at least in positive decisions) is scaled from the relative evidence for the chosen item over unchosen items in the array.

Within multiple-item arrays, we systematically varied the similarity of the weakest items in the array to the target (or target-replacement). We did so to test whether the relative nature of confidence judgments would incorporate information from *all* items in the array, or whether confidence judgments would be scaled based on the relative information from the chosen and next-best item. In Experiments 3 and 5, we found that confidence increased as the weakest item(s) became increasingly dissimilar to the target (or target-replacement), which provides some support for the hypothesis that confidence judgments incorporate information from all items within the array. These findings are consistent with the "dud effect" reported by Windschitl and Chambers (2004), and extended upon by Charman et al. (2011), in which the addition of a weak item to a choice set increases confidence in a chosen item. Our results suggest that a dud effect is a natural product of a sequential sampling system in which confidence reflects the difference in evidence strengths between items. Importantly, our results show that even small changes in item plausibility can affect confidence judgments – an obvious dud is not required.

The magnitude of the similarity effect was larger in 2-alternative tasks than in 4-alternative tasks; indeed, in Experiment 4, we were unable to detect any significant effect (and the effect size was close to zero). This may have been because the effects of manipulating a single lineup member within a multiple-item array are fairly subtle, with that single item exerting only a small pull overall balance-of-evidence. When we manipulated the similarity of the two weakest items in the array in Experiment 5, we were once again able to detect a significant effect, which provides some support for this suggestion.

With compound decisions, we investigated positive decisions (i.e., decisions in which an item was chosen) and negative decisions (i.e., decisions in which no item was chosen) separately. Our motivation for doing so was two-fold: theoretically, there are reasons to expect that confidence in negative decisions may be scaled from different information than confidence in positive decisions (e.g., Weber & Brewer, 2004); additionally, research from the eyewitness identification literature has consistently shown asymmetrical relationships between confidence and accuracy in positive and negative decisions (e.g., Brewer & Wells, 2006; Horry et al., 2012; Sauerland & Sporer, 2009). As we predicted, while target-lure similarity had consistent, predictable effects on confidence and response latency in positive decisions (in line with the patterns observed within forced-choice tasks), we observed no such effects on negative decisions.

Why is there a discrepancy in findings between positive and negative decisions? There are several possibilities. We may have lacked statistical power to detect subtle effects with negative decisions, due to the smaller number of negative decisions, and the relatively large error variance associated with them. However, looking at the patterns of means for the negative decisions in Figures 5, 7, and 9, it is clear that they do not follow any predictable pattern as would be expected if we had simply lacked power to detect a small but real effect. A second possibility is that negative decisions are heterogeneous, capturing some true rejections of the items in the array, as well as default responses when the participant has insufficient information to make a positive choice (perhaps because the target wasn't encoded effectively due to a lapse in attention) (Sporer, Penrod, Read, & Cutler, 1995). If that were the case, we would expect longer average response latencies for negative than positive decisions, due to those trials in which the decision "timed out" due to insufficient information. However, response latencies for correct negative decisions were similar to, and in some cases, shorter than, correct positive decisions.

A third possibility is that the consistent effects of similarity on confidence for positive decisions were driven by the ecophoric similarity of the items in the array (i.e., the similarity between each test item and the memorial representation of the target; Tulving, 1981), rather than the perceptual similarity relationships amongst the items. In our target-absent arrays, ecphoric similarity would have been rather low across all items (given that the target-replacement was already seven generations removed from the target). Consequently, in the target-absent trials, the crucial source of similarity may have been at floor. At the suggestion of a reviewer, we examined the effects of similarity on confidence in negative decisions from target-present arrays. In these trials, the perceptual and ecphoric similarity relationships between items are identical for positive and negative decisions. Because target-present trials were rejected infrequently, we combined all target-present rejections from Experiments 3 to 5 to achieve a sufficient number of trials (N = 301). In the regression model, we included an additional random intercept for experiment number. Adding similarity to a baseline model did not significantly improve model fit, χ^2 (N = 301) = 3.16, p = .37. This null effect would seem to speak against the possibility that low ecphoric similarity in the target-absent trials was responsible for the positive-negative discrepancy. Of course, we must be cautious to avoid drawing any strong conclusions from null results, particularly as the number of trials was low, and the variance was high. A stronger test of this potential explanation provides an avenue for future research.

Any potential explanation for our findings must be able to account for two facets of our findings: i) that the similarity between the target-replacement and the lures did not significantly affect confidence ratings; and ii) that confidence and accuracy were positively related for negative decisions. Our working hypothesis is that confidence judgments in negative decisions incorporate different information than confidence judgments in positive decisions. Specifically, negative decision confidence may index the difference between the

amount of evidence required for a positive decision and the amount of evidence associated with the best match – in other words, by how much did the best candidate fall short of the threshold? This hypothesis could explain why similarity did not affect confidence, as the similarity between items should have no (or minimal) influence on the distance between the best candidate and the threshold. It could also explain why we consistently found positive confidence-accuracy relationships for negative decisions; on average, the distance between the best candidate and the threshold (and therefore, confidence) should be higher on target-absent trials than on target-present trials. The cognitive and neural correlates of confidence judgments in negative decisions provide another interesting avenue for future research.

From a theoretical perspective, our results suggest that similar mechanisms can explain confidence judgments in multiple-alternative and compound decisions, and in simpler yes/no or 2-alternative decisions. Future research should aim to clarify exactly what those mechanisms are. Are confidence judgments derived from a diffusion process, with some post-decisional accrual of information (e.g., Moran et al., 2015; Pleskac & Busemeyer, 2010)? Does each possible decision-confidence pairing accumulate on its own counter (Ratcliff & Starns, 2013)? Are confidence judgments derived from the balance-of-evidence across independent accumulators (e.g., Brown & Heathcote, 2008; Van Zandt, 2000)? Or are confidence judgments produced by some other process? To begin answering these sorts of questions, current models of decision-making would need to be formally extended, so that they can incorporate multiple choices (which is already the case with Brown & Heathcote's linear ballistic accumulator model, but is not true for the vast majority of other models) and compound decisions.

Across all of our experiments, confidence judgments and response latency were negatively related, even when target-lure similarity was held constant. Of course, this relationship is predicted by sequential sampling models; however, it has also been suggested

that response latency makes an independent contribution to confidence judgments (Kiani et al., 2014). Specifically, decision makers evaluate not only the relative strengths of evidence associated with items, but they also consider how long the decision took, based on a learned association that more difficult decisions are made more slowly. Our data are unable to provide any critical tests of this hypothesis, but exploring further the relationships between response latency and confidence in complex decisions (for example, using signal-to-respond or deadline procedures) would be an interesting avenue of research.

We observed consistent relationships between confidence, accuracy and response latencies for negative decisions, as well as for positive decisions. These relationships make sense from a sequential sampling perspective, if one assumes that negative evidence accumulates over the course of a trial. Specifically, negative evidence should accumulate more rapidly when the target is absent, producing faster decisions than when the target is present. Furthermore, assuming that confidence is based on a valid cue of target-absence (as we have speculated, the distance between the best candidate and the positive decision threshold), confidence should be higher, on average, for correct decisions than for incorrect decisions. These relationships are often not observed in more naturalistic, single-trial eyewitness identification paradigms for reasons that are currently unknown (e.g., Brewer & Wells, 2006; Horry et al., 2012; Lindsay et al., 2013). One possibility is that, in single-trial paradigms, participants may reject a lineup when they don't know who to choose, thus contaminating true rejections with uninformative 'default' responses (Sporer et al., 1995; Weber & Perfect, 2012). There are also many reasons why the negative counter might 'win' the race, some of which may not be useful for discriminating between target-present and target-absent trials. For example, the witness may have failed to adequately encode the face of the perpetrator, the memory trace may have deteriorated due to forgetting or retroactive interference, or the target in the lineup may appear quite differently than how he appeared at

encoding (i.e., the target may have low ecphoric similarity to himself). Any of these issues could reduce the strength of the confidence-accuracy relationship in a single-trial eyewitness identification study.

From an applied perspective, perhaps the most well-studied compound decision task is the eyewitness identification task, in which an eyewitness attempts to identify a target (who may or may not be present) from a lineup. Confidence in eyewitness identifications has garnered much interest in recent years, with a consensus emerging that, when a witness makes a positive decision, confidence is a meaningful predictor of accuracy (Brewer & Wells, 2006; Palmer et al., 2013). Of course, the strength of the confidence-accuracy relationship is likely to be moderated by many variables (e.g., Horry et al., 2012; Palmer et al., 2013; Sauer, Brewer, Zweck, & Weber, 2010), and it is important to understand when confidence and accuracy might become dissociated.

If our findings generalize to the eyewitness identification context, there are some important applied implications. Expressions of confidence from an eyewitness shape how that eyewitness, and the evidence that he or she provides, is perceived by investigators, judges, and juries (e.g., Brewer & Burke, 2002). Here, we have shown that the presence of a lure who is relatively dissimilar to the target (though still a plausible match) can boost confidence, with the potential downstream effect of increasing the perceived reliability of the decision. Taken to extremes, the presence of a very dissimilar lineup member (a "dud") can produce even more marked overconfidence (Charman et al., 2011). On the other side of the coin, the presence of a highly similar lure can reduce confidence, thus potentially decreasing the perceived reliability of the decision. Luus and Wells (1991) argued that there is likely to be some inverted-U shaped function between target-lure similarity and eyewitness identification performance. We would argue that there is also likely to be a similar function corresponding to optimum confidence-accuracy calibration. If similarity is too low, witnesses may be

overconfident in their decisions; if similarity is too high, witnesses may be underconfident in their decisions. Given recent very strong claims that highly confident witnesses are very likely to be correct (e.g., Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted, Mickes, Dunn, Clark, & Wells, 2016), it is important to further investigate the extent to which over- and under-confidence are produced by lineup composition.

It is important to note that, in real-world tasks, confidence judgments are likely to incorporate information from external cues, as well as those that are generated by the decision process. Charman, Carlucci, Vallano, and Hyman Gregory (2010) put forward the cues-based inference model of confidence judgments, a three-stage process for generating confidence judgments in lineup decisions. First, the decision-maker assesses the internal cues that formed the basis of their decision. If those cues are weak, the decision-maker proceeds to search for external cues (such as feedback from the lineup administrator; Charman & Wells, 2012; Wells & Bradfield, 1998). Finally, if those cues are sufficiently credible, they are integrated with the internal cues to produce a confidence judgment. We have focused on the first of these three stages, the assessment of internal cues. However, we do not mean to downplay the role of external cues. The internal signals that indicate evidence strengths are likely very fragile, and may decay quite quickly (Brewer, Keast, & Rishworth, 2002). Thus, rather than contradicting the cues-based inference model, the current results supplement it by elaborating on the nature of the internal cues assessed in the first stage of the model. Refining our understanding of how these internal cues related to confidence should be a high priority for future research.

Our paradigm differs in many ways from the standard eyewitness identification paradigm, in which participants typically encode a single target engaged in a mock crime, and then, after a variable delay, view a single target-present or target-absent lineup. For example, our participants completed many trials, they knew that they would need to attempt to identify

each target as he was being encoded, and they only had to store the memorial representation for a short time (2000 ms). Furthermore, we used the same image of the target at study and test, potentially allowing for low-level image matching rather than true face matching (Bruce, 1982). In contrast, real eyewitnesses must cope with the demands placed on their visual processing systems by changes in lighting, viewpoint, distance, and, in the cases of a photographic lineup, the shift from a dynamic encoding stimulus to a static test image (Megreya & Burton, 2008). However, despite these differences, it is important to stress that the decision frame is the same: the participant has a target stored in memory, and must search for that target in an array of faces that may or may not include that target. Though lacking in ecological validity, the 'mini-lineup' paradigm adopted here is better suited to the sorts of repeated testing that is associated with exploring and refining aspects of theory (see, for example, Weber & Brewer, 2004, 2006). In contrast, in a single-trial identification paradigm, individual differences can easily overwhelm subtle effects, which can hinder theoretical advances. Of course, before generalizing these findings to eyewitness identification decisions, it will be necessary to test whether they 'scale up' to more ecologically valid paradigms.

Conclusions

Across five experiments, we provided evidence that similar cognitive mechanisms are involved in producing confidence judgments in complex decisions (i.e., multiple-alternative and compound decisions) as in simple, two-alternative decisions. When we systematically varied the similarity between the target and the weakest item in the choice array, confidence in positive decisions increased, and response latency decreased. These results suggest that confidence judgments for positive decisions incorporate information about the relative evidence for a chosen item over the unchosen items. Of course, confidence judgments also appear to contain information about the absolute strength of the chosen item (e.g.,

Teodorescu et al., 2016); indeed, recent research using words as stimuli suggests that absolute information may play a large role in confidence judgments (Zawadzka et al., 2016). It is possible that the contributions of relative and absolute evidence to confidence judgments vary as a function of task demands and of the complexity of the stimuli. Formal modelling of compound decisions seems warranted to answer questions such as whether the relationships between confidence, accuracy, and response latencies in compound decisions are better explained by diffusion (e.g., Moran et al., 2015; Pleskac & Busemeyer, 2010) or accumulator (e.g., Brown & Heathcote, 2008; Van Zandt, 2000) processes, and to clarify the informational basis of confidence judgments in negative decisions.

References

Alexander, J. R. M. (1990). An approximation to *d'* for n-alternative forced choice. Unpublished manuscript. Accessible at http://eprints.utas.edu.au/475/ as of 23rd March 2016

Angell, F. (1907). On judgments of "like" in discrimination experiments. *The American Journal of Psychology, 18,* 253-260. doi: 10.2307/1412416

Ashby, F. G. (1983). A biased random walk model for two choice reaction times. *Journal of Mathematical Psychology, 27,* 277-297. doi: 10.1016/0022-2496(83)90011-1

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55,* 412-428. doi: 10.3758/BF03205299

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics, 54,* 75-81. doi: 10.3758/BF03206939

Baranski, J. V. & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception & Performance, 26,* 929-945. doi: 10.31037/0096-1523.24.3.929

Brewer, N. & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock juror judgments. *Law & Human Behavior, 26,* 353-364. doi: 10.31023/A:1015380522722

Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. *Law & Human Behavior, 30,* 31-50. doi: 10.1007/s10979-006-9002-7

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8,* 44-56. doi:10.1037//1076-898X.8.1.44

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, 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,* 153-178. doi: 10.1016/j.cogpsych.2007.12.002

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73,* 105-116. doi: 10.1111/j.2044-8295.1982.tb01795.x

Charman, S. D., Carlucci, M., Vallano, J., & Hyman Gregory, A. (2010). The selective cue integration framework: A theory of postidentification witness confidence assessment. *Journal of Experimental Psychology: Applied, 16,* 204-218. doi: 10.1037/a0019495

Charman, S. D., & Wells, G. L. (2012). The moderating effect of ecphoric experience on post-identification feedback: A critical test of the cues-based inference conceptualization. *Applied Cognitive Psychology, 26,* 243-250. doi: 10.1002/acp.1815

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law & Human Behavior, 35,* 479-500. doi: 10.1007/s10979-010-9261-1

Chen, J., Yang, H., Wang, A., & Fang, F. (2010). Percpetual consequences of face viewpoint adaptation: Face viewpoint aftereffect, changes of differential sensitivity to face view, and their relationship. *Perception, 10,* 1-11. doi: 10.1167/10.3.12

Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence-accuracy inversions in scene recognition: A remember-know analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 1306-1315. doi: 10.1037/0278-7393.24.5.1306

Duncan, M. (2006). *A signal detection model of compound decision tasks.* (Tech. Rep. No. TR2006-256). Toronto, ON: Defence Research and Development Canada.

Festinger, L. (1943). Studies in decision: 1. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32,* 291-306. doi: 10.1037/h0056685

Heathcote, A., Bora, B., & Freeman, E. (2010). Recollection and confidence in two-alternative forced choice episodic recognition. *Journal of Memory & Language, 62,* 183-203. doi: 10.1016/j.jml.2009.11.003

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. doi: 10.3758/PBR.16.5.824

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

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4,* 11-26. doi: 10.1080/17470215208416600

Hiller, R. M., & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2,* 185-191. doi: 10.1016/j.jarmac.2013.07.001

Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied, 18,* 346-360. doi: 10.1037/a0029779

Juslin, O., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344-366. doi: 10.1037/0033-295X.104.2.344

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84,* 1329-1342. doi: 10.1016/j.neuron.2014.12.015

Lacouture, Y., & Marley, A. A. J. (2004). Choice and response time processes in the identification and classification of unidimensional stimuli. *Perception & Psychophysics, 66,* 1206-1226. doi: 10.3758/BF03196847

Lindsay, R. C. L., Kalmet, N., Leung, J., Bertrand, M. I., Sauer, J. D., & Sauerland, M. (2013). Confidence and accuracy of lineup selections and rejections: Postdicting rejection accuracy with confidence. *Journal of Applied Research in Memory and Cognition, 2,* 179-184. doi: 10.1016/j.jarmac.2013.06.002

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14,* 364-372. doi: 10.1037/a0013464

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information intergration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology, 78,* 99-147. doi: 10.1016/j.cogpsych.2015.01.002

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239-257. doi: 10.1037/a0023007

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,* 55-71. doi: 10.1037/a0031602

Papesh, M. H., & Goldinger, S. D. (2010). A multidimensional scaling analysis of own- and cross-race face spaces. *Cognition, 116,* 283-288. doi: 10.1016/j.cognition.2010.05.001

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

Potter, T., & Corneille, O. (2008). Locating attractiveness in the face space: Faces are more attractive when closer to *their* group prototype. *Psychonomic Bulletin & Review, 15,* 615-622. doi: 10.3758/PBR.15.3.615

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59-108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R. & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116,* 59-83. doi: 10.1037/a0014086

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review, 120,* 697-719. doi: 10.1037/a0033152

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50,* 408-424. doi: 10.1016/j.jml.2003.11.002

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

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy. *Journal of Experimental Psychology: Applied, 15,* 46-62. doi: 10.1037/a0014560

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

Smith, S. M., Lindsay, R. C. L., Pryke, S., & Dysart, J. E. (2001). Postdictors of eyewitness errors: Can false identifications be diagnosed in the cross-race situation? *Psychology, Public Policy, & Law,* 7, 153-169. doi: 10.1037//1076-8971.7.1.153

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin,* 118, 315-327. doi: 10.1037/0033-2909.118.3.315

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164-182. doi: 10.1037/1082-989X.9.2.164

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 1397-1410. doi: 10.1037//0278-7393.24.6.1397

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,* 22-38. doi: 10.3758/s13423-015-0858-8

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning & Verbal Behavior, 20,* 479-496. doi: 10.1016/S0022-5371(81)90129-8

Usher, M. & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550-592. doi: 10.1037//0033-295X.108.3.550

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's Law in a stochastic race model with speed-accuracy tradeoff. *Journal of Mathematical Psychology, 46,* 704-715. doi: 10.1006/jmps.2002.1420

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & 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,* 37-58. doi: 10.1080/0140137008931117

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,* 179-197. doi: 10.1016/0001-6918(82)90006-3

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10,* 156-172. doi: 10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2006). Positive versus negative face recognition decisions: Confidence, accuracy, and response latency. *Applied Cognitive Psychology, 20,* 17-31. doi: 10.1002/acp.1166

Weber, N., & Perfect, T. J. (2012). Improving eyewitness identification accuracy by screening out those who say they *don't know. Law and Human Behavior, 36,* 28-36. doi: 10.1037/h0093976

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83,* 360-376. doi: 10.1037/0021-9010.83.3.360

Williamson, G. F. (1915). Individual differences in belief, measured and expressed by degrees of confidence. *The Journal of Philosophy, Psychology and Scientific Methods, 12,* 127-137. doi: 10.2307/2012776

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 198-215. doi: 10.1037/0278-7393.30.1.198

Wixted, J. T. & Mickes, L. (2010). A continuous dual-process model of Remember/Know judgments. *Psychological Review, 117,* 1025-1054. doi: 10.1037/a0020874

Wixted, J. T., Mickes, L., Clark, S., Gronlund, S., & Roediger, H. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70,* 515-526. doi: 10.1037/a0039510

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113,* 304-309. doi: 10.1073/pnas.1516814112

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods,* 41, 257-267. doi: 10.3758/BRM.41.2.257.

Wright, D. B. & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical & Statistical Psychology,* 62, 439-456.

Xu, X., Yue, X., Lescroart, M. D., Biederman, I., & Kim, J. G. (2009). Adaptation in the fusiform face area (FFA): Image or person? *Vision Research, 49,* 2800-2807. doi: 10.1016/j.visres.2009.08.021

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General, 144,* 489-510. doi:10.1037/xge0000062

Zawadzka, K., Higham, P. A., & Hanczakowski, M. (2016). Confidence in forced-choice recognition: What underlies the ratings? *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Footnote

[1]Tulving (1981) drew a distinction between two types of similarity: perceptual similarity (the perceived similarity between two physically present items), and ecphoric similarity (the similarity between a physically present item and a memorial representation). Though some researchers have attempted to dissociate perceptual and ecphoric similarity (e.g., Dobbins, Kroll, & Liu, 1998; Heathcote, Freeman, Etherington, Tonkin, & Bora, 2009), we make no such attempt here. Rather, perceptual and ecphoric similarity are confounded (as a lure becomes less perceptually similar to the target, it is also less ecphorically similar to the representation of that target). As a consequence, it should be noted that, though we preserved the perceptual similarity relationships across target-present and target-absent trials, we did not preserve the ecphoric similarity relationships; ecphoric similarity would have been substantially higher on target-present trials than on target-absent trials. The implications of this are discussed in the General Discussion.

[2]As η^2 can only be positive, a 90% confidence interval, rather than a 95% confidence interval, is appropriate (Steiger, 2004).

[3]Analyses were repeated with the outliers included. These analyses are reported in the Supplementary Materials. In all but two cases, the results were not changed by the inclusion of the outliers. The discrepancies are noted where they occur.

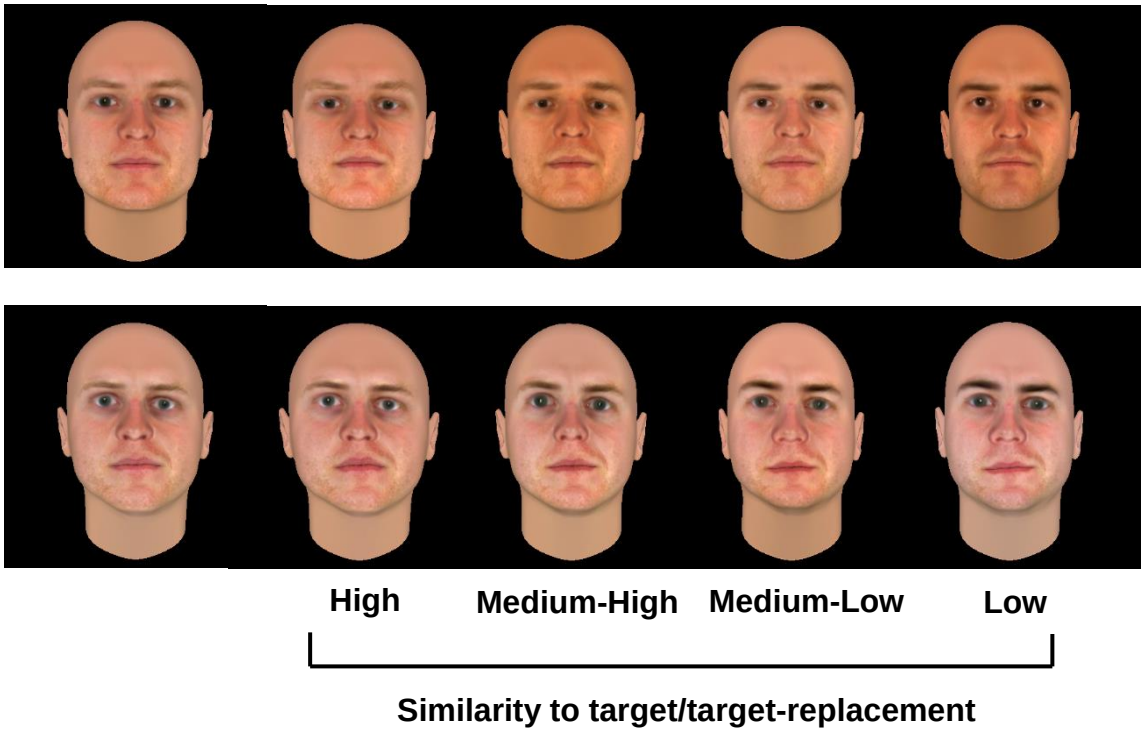
[4]In all of the correlational analyses, the degrees of freedom vary between conditions, as a coefficient could not always be calculated for each participant – for example, if confidence or accuracy was constant.

Table 1.

Mean discriminability and choosing rates, with standard deviations, as a function of similarity

	Similarity level							
	High		Medium-High		Medium-Low		Low	
Experiment	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
				d'-identification				
1	0.81	0.50	1.42	0.88	1.72	1.06	2.50	1.41
2	0.53	0.38	0.75	0.49	0.84	0.46	0.87	0.49
3	1.78	1.29	2.71	1.08	3.89	1.26	2.87	1.41
4	0.48	0.50	0.50	1.16	0.84	0.65	0.65	0.93
5	0.55	0.79	0.86	0.91	1.58	1.21	1.26	0.87
				d'-detection				
3	1.81	1.71	2.06	1.46	2.16	1.41	2.14	1.31
4	1.29	1.31	1.60	1.36	1.41	1.55	1.46	1.36
5	1.61	1.29	1.35	1.17	1.14	1.17	1.29	1.39
				Choosing rates				
3	.73	.17	.75	.14	.72	.14	.68	.15
4	.81	.15	.81	.14	.81	.16	.80	.14
5	.79	.18	.70	.16	.67	.22	.72	.19

Figure 1.



Example target face with siblings (top) and target-replacement with siblings (bottom).

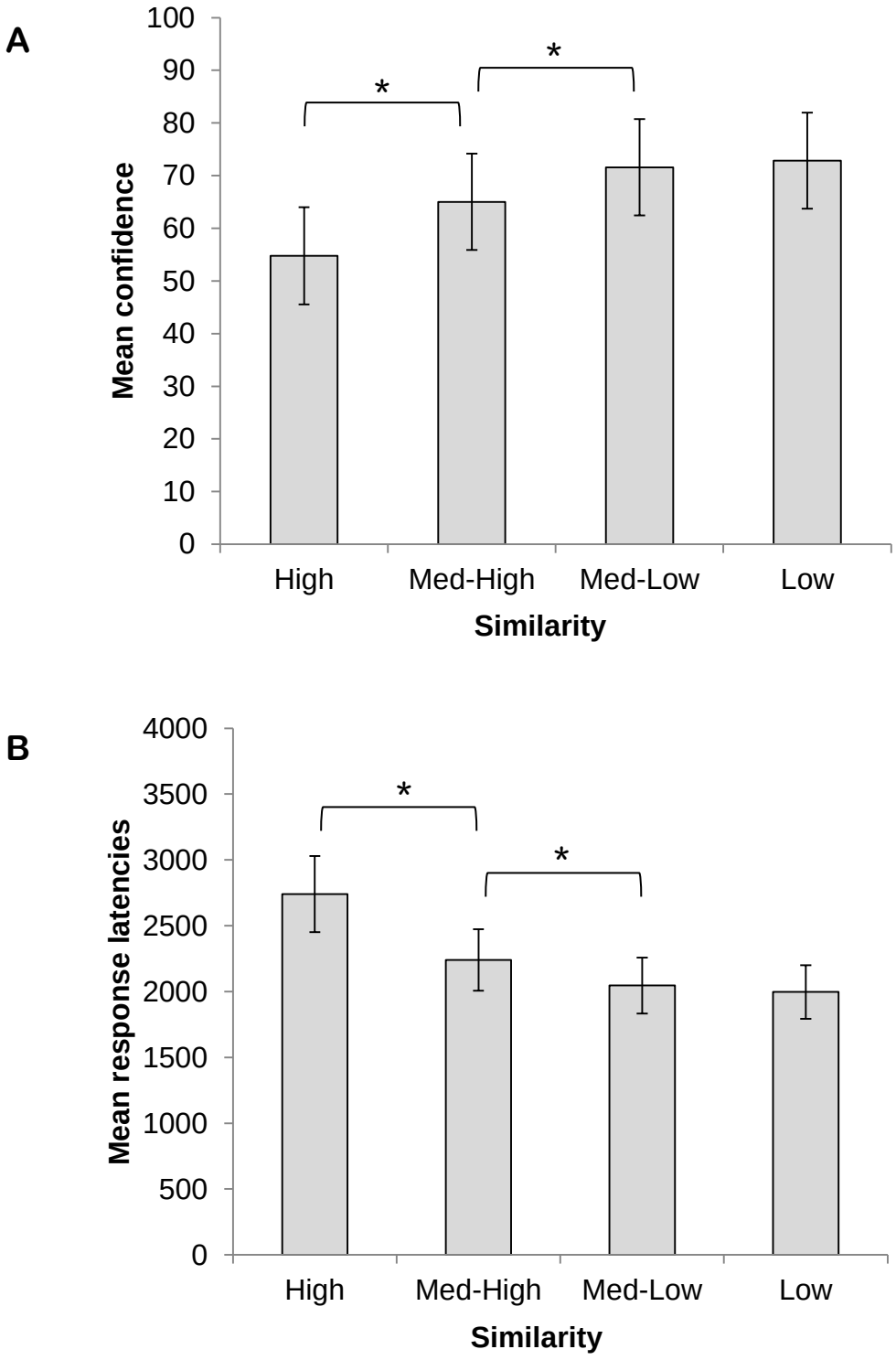


Figure 2. Mean confidence ratings (top panel) and response latency (bottom panel) for correct target identifications in Experiment 1. Error bars represent 95% confidence intervals. An asterisk indicates that the 95% CI of the difference excluded zero.

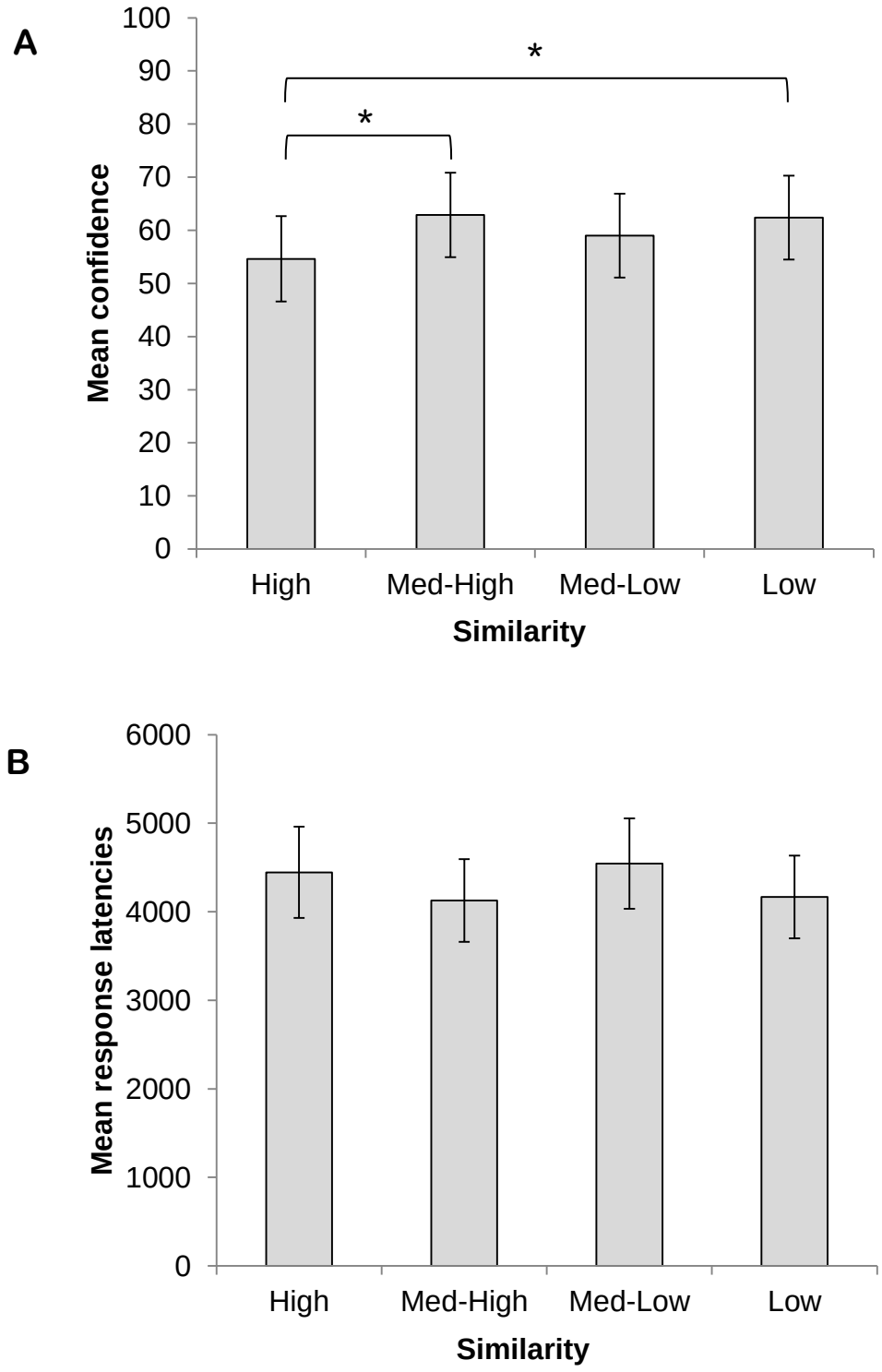


Figure 3. Mean confidence ratings (top panel) and response latency (bottom panel) for correct target identifications in Experiment 2. Error bars represent 95% confidence intervals. An asterisk indicates that the 95% CI of the difference excluded zero.

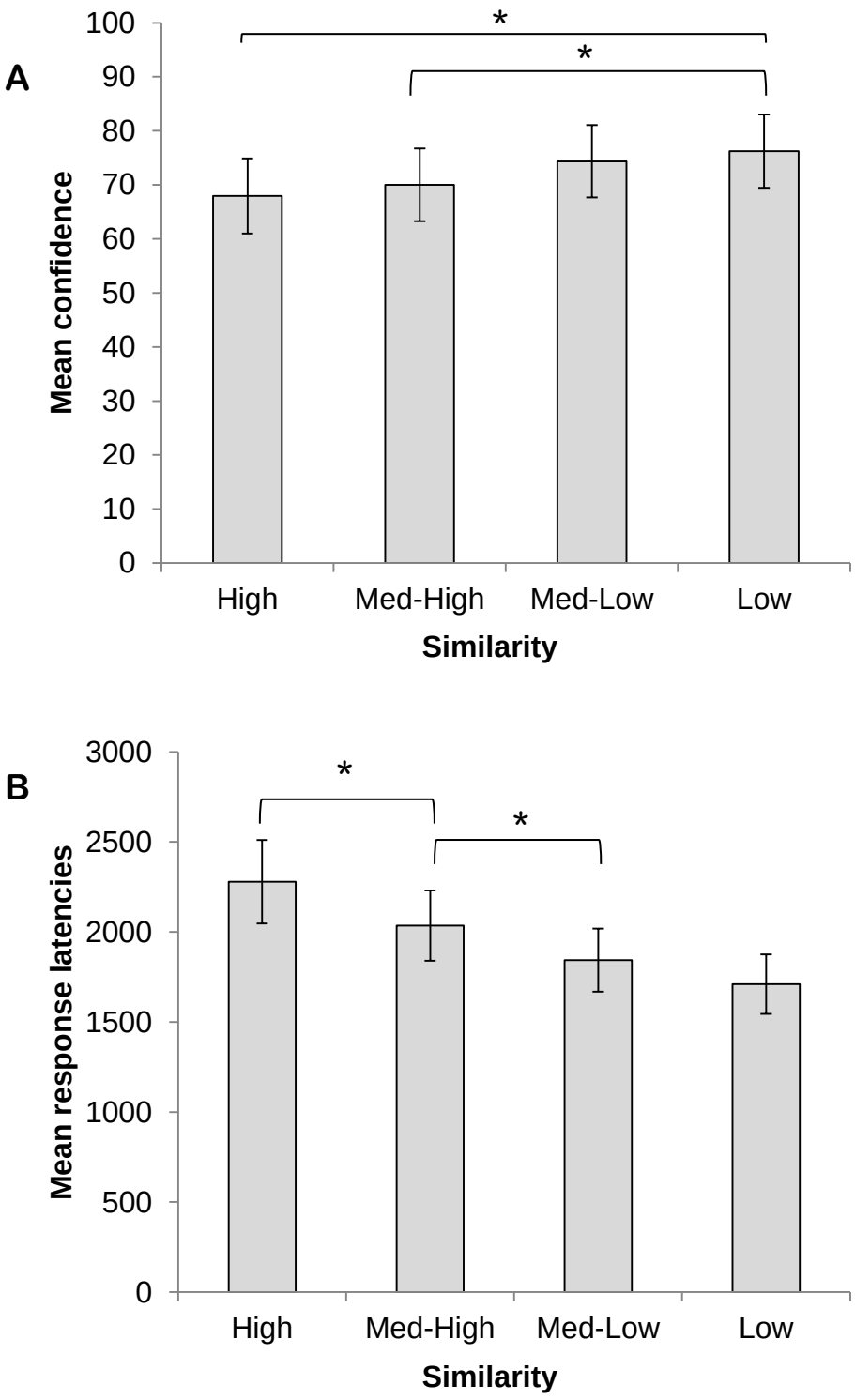


Figure 4. Mean confidence ratings (top panel) and response latency (bottom panel) for correct target identifications in Experiment 3. Error bars represent 95% confidence intervals. An asterisk indicates that the 95% CI of the difference excluded zero.

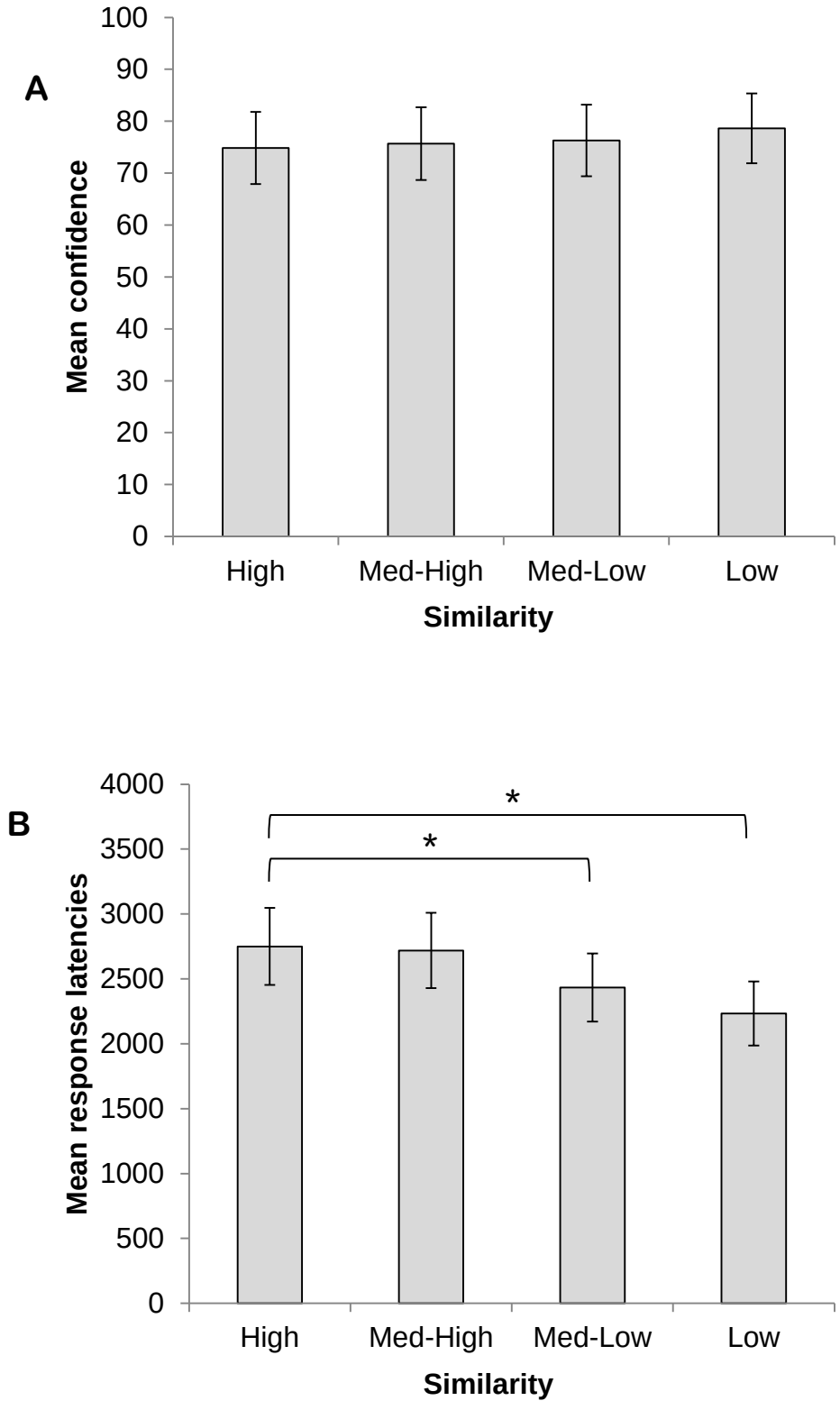


Figure 5. Mean confidence ratings (top panel) and response latency (bottom panel) for correct rejections in Experiment 3. Error bars represent 95% confidence intervals. An asterisk indicates that the 95% CI of the difference excluded zero.

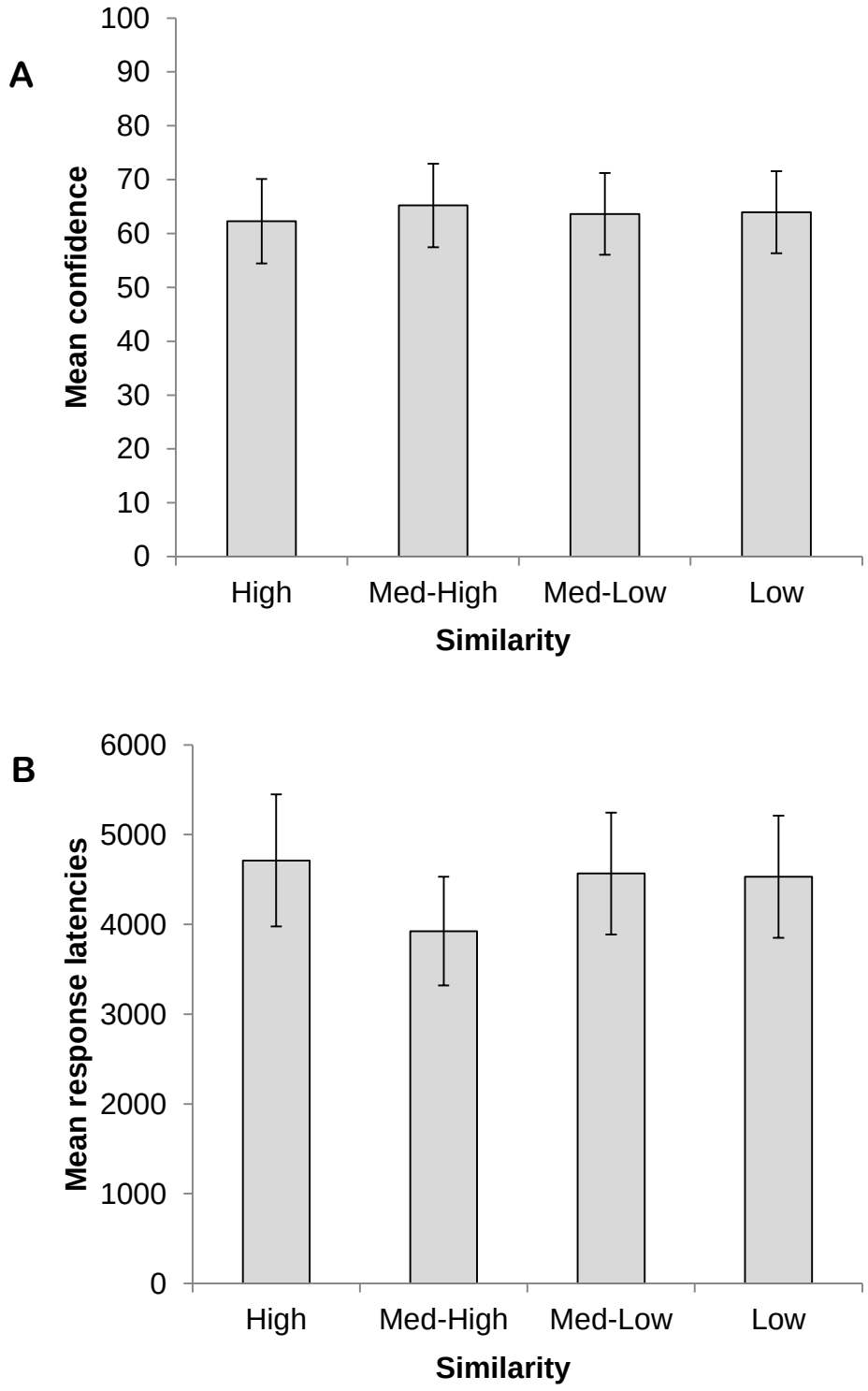


Figure 6. Mean confidence ratings (top panel) and response latency (bottom panel) for correct target identifications in Experiment 4. Error bars represent 95% confidence intervals.

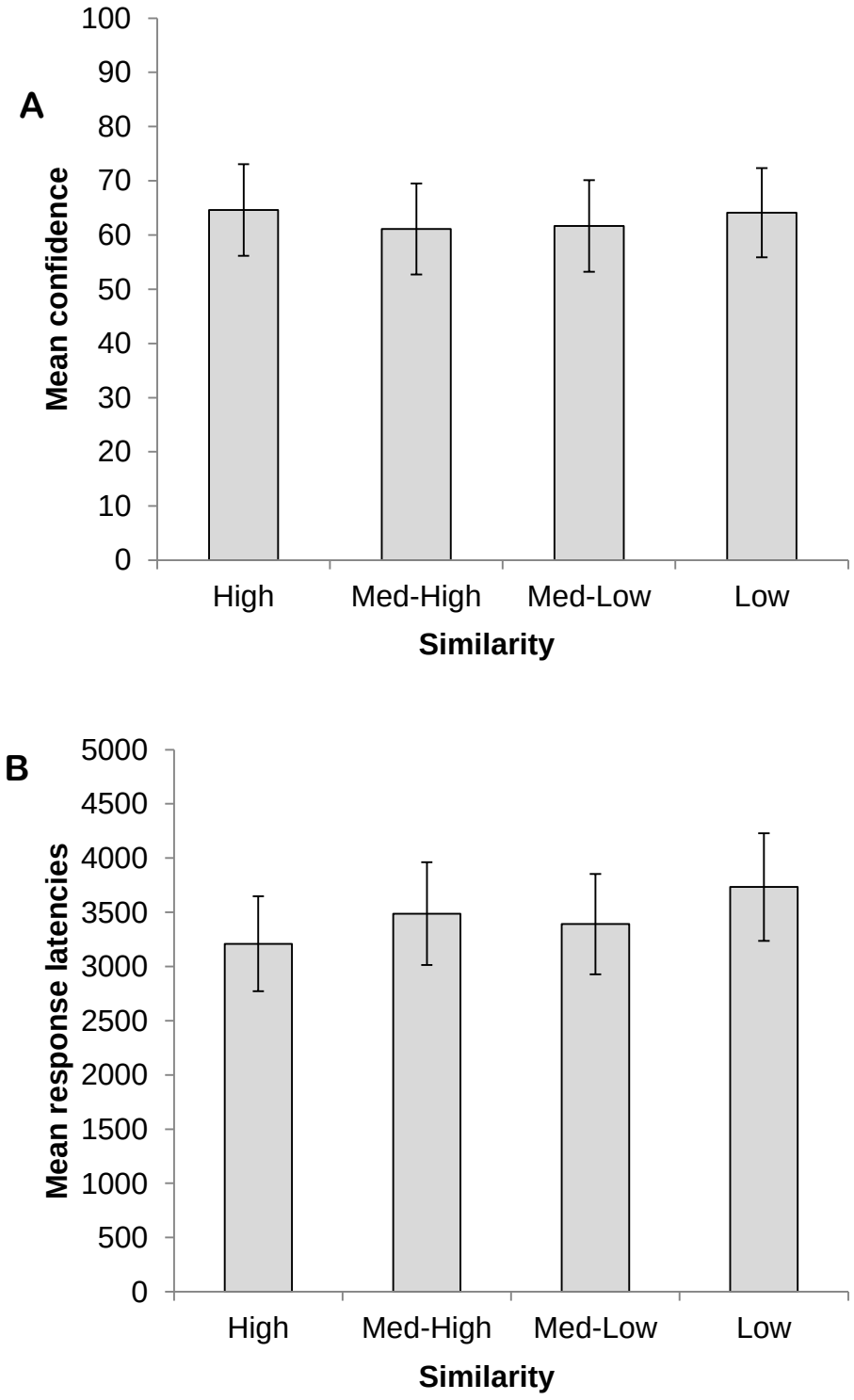


Figure 7. Mean confidence ratings (top panel) and response latency (bottom panel) for correct rejections of target-absent trials in Experiment 4. Error bars represent 95% confidence intervals.

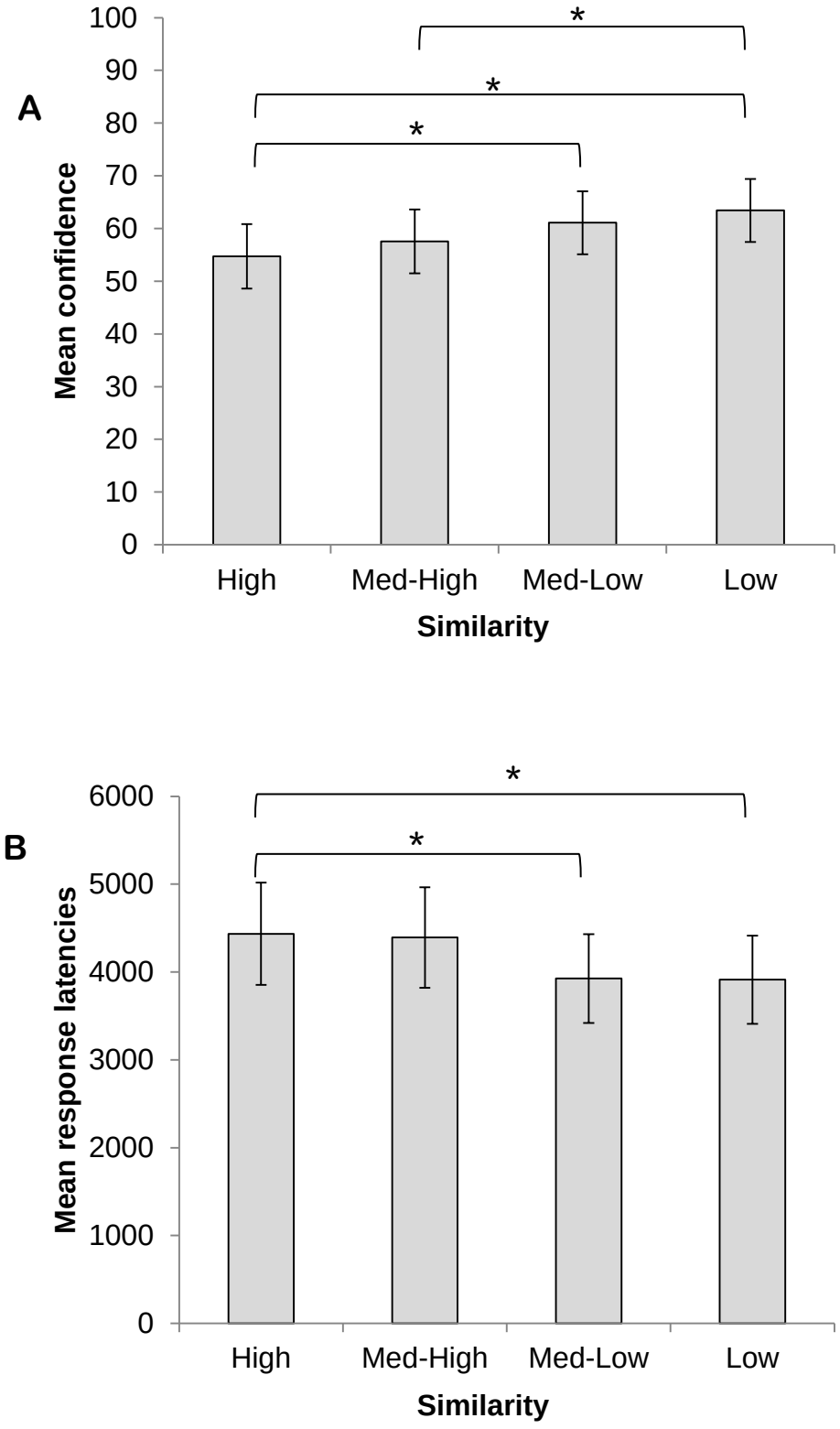


Figure 8. Mean confidence ratings (top panel) and response latency (bottom panel) for correct target identifications in Experiment 5. Error bars represent 95% confidence intervals. An asterisk indicates that the 95% CI of the difference excluded zero.

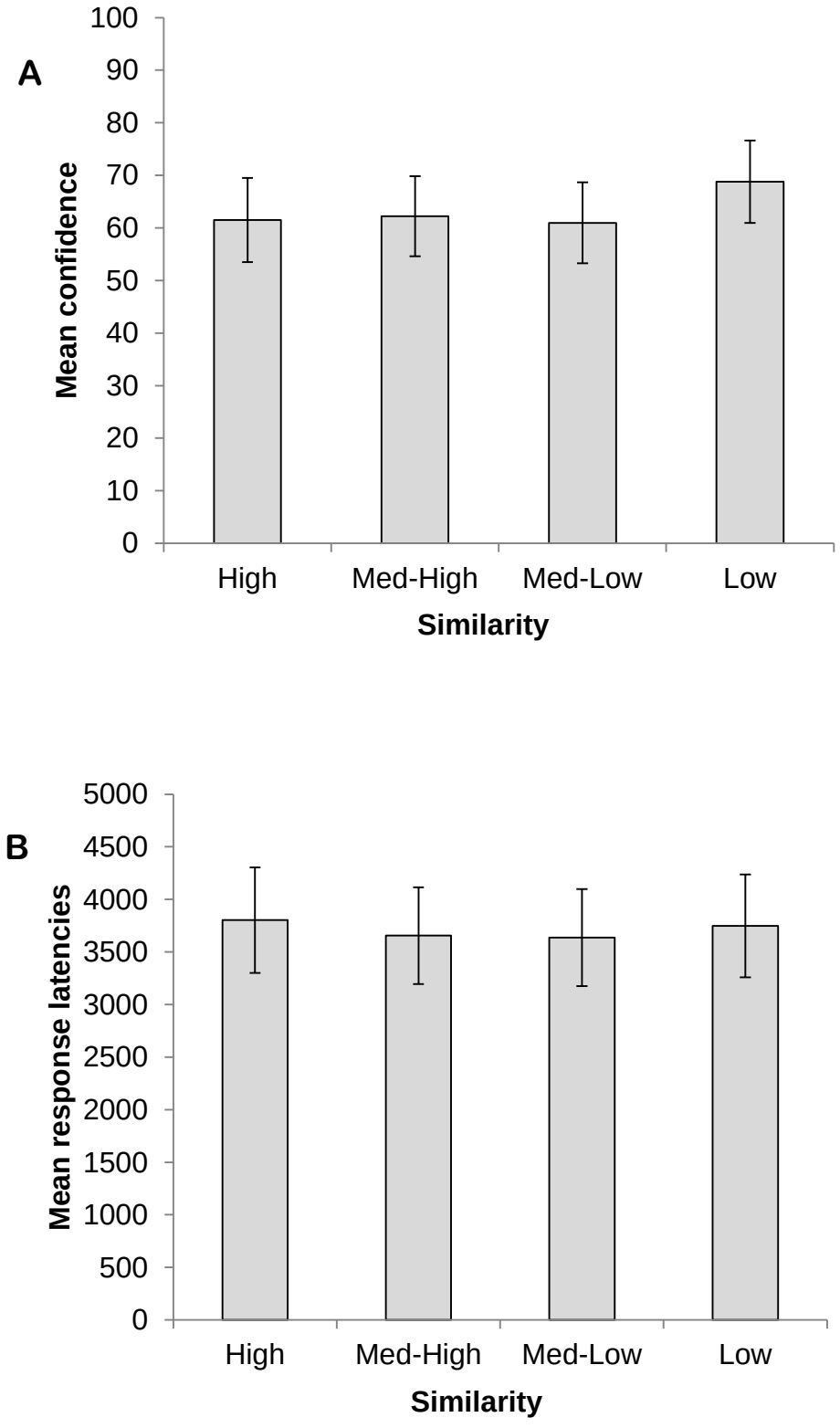


Figure 9. Mean confidence ratings (top panel) and response latency (bottom panel) for correct rejections in Experiment 5. Error bars represent 95% confidence intervals. An asterisk indicates that the 95% CI of the difference excluded zero.