© 2021, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/law0000329

Could precise and replicable manipulations of suspect-filler similarity optimize eyewitness identification performance?

Carmen A. Lucas and Neil Brewer

Flinders University

Author note.

Carmen A. Lucas, College of Education, Psychology and Social Work, Flinders University; Neil Brewer, College of Education, Psychology and Social Work, Flinders University.

This research was supported by an Australian Research Council Discovery Project Grant. The data are publicly available from: https://osf.io/5g87z/

Correspondence concerning this article should be addressed to Carmen Lucas, College of Education, Psychology and Social Work, Flinders University, GPO Box 2100, Adelaide, South Australia, Australia. E-mail: carmen.lucas@flinders.edu.au

Abstract

The optimal level of suspect-filler similarity in police lineups remains undefined. Difficulties inherent in pinpointing precise and replicable variations in face similarity create challenges for examining the effects of suspect-filler similarity on identification outcomes and providing decisive lineup construction recommendations. We tested the relationship between suspect-filler similarity and identification outcomes using stimuli developed with a combination of face matching and morphing software that could potentially be used by police investigators to standardize lineup composition. A group of fillers selected to be as low in similarity to the suspect as possible, while still matching a basic perpetrator description, were morphed to the suspects. Participants viewed lineups in which the fillers were either unmorphed, a 33% morph or a 50% morph to the suspect. Results showed a tendency for identification performance—including accuracy, discriminability between guilty and innocent suspects and the confidence-accuracy relationship—to suffer as similarity levels increased beyond matching a basic perpetrator description. Thus, despite the attraction of exploiting modern technology to standardize suspect-filler similarity relations, our findings across multiple sets of encoding and test materials broadly cohere with Wells and colleague's argument (Luus & Wells, 1991; Wells et al., 1993) that identification performance may be optimized when fillers are selected to be as low as possible in similarity to the suspect while still matching description. However, before unequivocally endorsing such a lineup construction approach, it is necessary to conduct a systematic examination of precisely what constitutes an adequate perpetrator description and how this might be consistently obtained.

Keywords: identification decisions; match-description; suspect-filler similarity; confidence-accuracy relationship

Could precise and replicable manipulations of suspect-filler similarity optimize eyewitness identification performance?

Although there have been many developments in understanding how eyewitness identification evidence should be obtained, there remain unanswered questions about the precise level of suspect-filler similarity that will maximize identification performance and about how lineup constructors could reliably target this level of similarity. Consequently, current lineup construction recommendations (e.g., Wells et al., 2020) can be adhered to while permitting suspect-filler similarity variations that could affect identification decision outcomes and undermine the interpretation of identification evidence (cf. Lucas et al., 2020). Our guiding premise in the current study was that the definition of optimal lineup member characteristics might be facilitated if suspect-filler similarity manipulations such as those developed for laboratory studies could—using currently available technology— be operationalized in a precise and objective manner, thereby allowing equivalent levels of similarity to be replicated across research studies and, ultimately, by "real world" lineup constructors. Accordingly, we investigated the effect of suspect-filler similarity on identification decision performance with a focus on maximizing the precision and replicability of the lineup conditions created.

Two pivotal papers (i.e., Luus & Wells, 1991; Wells et al., 1993) on the question of how to compose a lineup appeared in the early 1990s. Luus and Wells (1991) made the key distinction between the intuitive strategy of selecting fillers based on their similarity to the suspect and matching fillers to the witness's description of the perpetrator. The effect of increasing suspect-filler similarity on identification performance was proposed to take the form of an inverted U-shape: that is, to a certain extent, increasing suspect-filler similarity enhances identification performance, but at some point it becomes detrimental to identification outcomes. Luus and Wells (1991) argued that the selection of similar looking

fillers runs the risk of reducing guilty suspect identifications without offering any additional protection to innocent suspects. In contrast, matching fillers to description was purported to protect innocent suspects from standing out (i.e., as the only plausible candidate), while allowing guilty suspects to remain distinguishable from the other lineup members. In a notably sophisticated study (especially for its time) involving seven different targets and 252 lineups individually constructed with reference to each participant witness's perpetrator description, Wells et al. (1993) reported data supporting the propositions advanced by Luus and Wells (1991). Compared to a condition where each of the fillers did not fully match the perpetrator description, matching to description while minimizing suspect-filler similarity protected innocent suspects without being detrimental to guilty suspect identifications. Conversely, fillers selected to be high in similarity to the suspect suppressed both guilty and innocent suspect identifications. It is important to note that although a high similarity face may also match description, it is certainly possible for a high similarity face to mismatch the suspect on a key feature and cause the suspect to stand out. In this regard, Wells et al. (1993) did not specify whether any of the fillers in the high similarity condition failed to match description; however, the pattern of results (i.e., both guilty and innocent suspect identifications being suppressed by the high similarity fillers) suggests that the high similarity fillers generally tended to also match description.

The landmark experiment of Wells et al. (1993) provided a compelling indication that low similarity match-description fillers should be preferred over fillers selected based on their similarity to the suspect. However, their experiment did not capture a comprehensive range of variations in suspect-filler similarity, leaving open the possibility that there exists some specific level of similarity beyond what is achieved by matching to description that will enhance identification performance. Reinforcing this possibility, a meta-analysis of studies examining suspect-filler similarity and identification performance not only revealed a general

trade-off between guilty and innocent suspect picks across variations in similarity, but also yielded a smaller decrease in guilty compared with innocent suspect identifications between the medium and high similarity filler conditions (Fitzgerald, et al., 2013). It is important to bear in mind that the protection afforded to innocent suspects (i.e., across increases in filler similarity) may have been overestimated by the common practice in identification studies of manipulating suspect-filler similarity relative to the perpetrator across both target-present and -absent conditions (Colloff et al., 2021; Oriet & Fitzgerald, 2018; Tunnicliff & Clark, 2000). Nevertheless, rather than there being any conclusive evidence regarding the optimal level of suspect-filler similarity, the strongest argument against increasing filler similarity beyond matching to description remains that there is no clear stopping point for ensuring that filler similarity is not too high (Luus & Wells, 1991; see also Colloff et al., 2021, for a recent argument that Signal Detection Theory (Green & Swets, 1966) predicts a decrease in identification performance as suspect-filler similarity increases beyond a match to description—the somewhat inconclusive results of their study are discussed below). Consequently, lineup construction recommendations require fillers to match-description but are (admittedly and appropriately) vague regarding the ideal level of suspect-filler similarity (Wells et al., 2020).

A major impediment to clarifying optimal lineup member characteristics—and an important stimulus for our investigation—is that the commonly used methods for operationalizing suspect-filler similarity are imprecise, with the inevitable consequence that they do not provide the means for recreating equivalent levels of similarity across different contexts. It is common for pilot participants to provide subjective ratings of similarity between the target and a potential filler pool, with high and low similarity faces then being selected on the basis of the averaged similarity ratings (e.g., Brewer & Wells, 2006; Colloff et al., 2021; Oriet & Fitzgerald, 2018). Although this approach can produce ratings of

suspect-filler similarity that differ significantly across conditions, with those conditions significantly affecting identification outcomes, it is clearly an imprecise method of varying the likeness between the fillers and suspect and may well be preventing ideal lineup member characteristics from becoming apparent. Moreover, similarity ends up being defined relative to the other levels of similarity in the study, preventing comparisons of performance across studies or translation of the results into concrete recommendations for filler selection (Fitzgerald et al., 2013).

For many years, subjective ratings given to face stimuli provided researchers with the best tool for assessing and varying suspect-filler similarity. However, with modern technologies allowing digital manipulation and appraisal of faces, there is an opportunity to exercise more precise control over the manipulation of face similarity. Moreover, such technologies could provide a way for real police lineups to be reliably constructed to reflect optimal suspect-filler similarity, thereby potentially leading to a standardized approach to lineup construction within and across law enforcement jurisdictions. Hence, we set out to create precise and replicable levels of suspect-filler similarity, using a combination of face matching and morphing software. *Betaface* face matching software (www.betaface.com/wpa)—previously implemented by Bergold and Heaton (2018) to measure the similarity of best-matching faces obtained from different sized databases—was used to identify fillers from a match-description filler pool that were of as low similarity as possible to designated suspects. *Fantamorph* morphing software (www.fantamorph.com)—previously used by Fitzgerald et al. (2015) to create very high levels of suspect-filler similarity—was then employed to systematically (in conjunction with the face matching software) manipulate suspect-filler similarity. Compared to the average similarity rating method of manipulating suspect-filler similarity this approach (a) achieves more precise variation in face similarity than is permitted by selecting naturally occurring face stimuli

based on a set of subjective ratings and (b) enables the assignment of a meaningful and replicable suspect-filler similarity value to each face.

Within the framework described, we investigated the effect of variations in match-description, suspect-filler similarity on identification decision outcomes using four different sets of encoding and test stimulus materials. In other words, for each stimulus set all lineup fillers matched a general target description but varied in similarity to the suspect. We examined the influence of this suspect-filler similarity manipulation on overall identification patterns and confidence, as well as on identification accuracy, the discriminability between guilty and innocent suspects and the confidence-accuracy relationship.

Previous research has clearly linked an increase in suspect-filler similarity with an overall shift from suspect to filler identifications (Fitzgerald et al., 2013), attributed to a decrease in people's ability to discriminate the lineup members from one another (Lucas et al., 2020). Lucas et al. (2020) also posited that an increase in suspect-filler similarity causes people to accept less evidence to make a positive identification. While this might be expected to increase lineup choosing, a willingness to accept less evidence for making a positive identification could result in unchanged or even reduced choosing rates when pitted against an overwhelming decrease in discriminability (e.g., the witness cannot decide between two lineup members). Therefore, it is perhaps unsurprising that the effects of filler similarity on lineup choosing have been inconsistent (Fitzgerald, et al., 2013). The decrease in lineup member discriminability across increases in suspect-filler similarity has also been reflected in decreasing levels of post-identification confidence (e.g., Fitzgerald et al., 2015; Lucas et al., 2020; cf. Brewer & Wells, 2006), reinforcing the notion that confidence reflects the strength of evidence for chosen compared with unchosen items (Horry & Brewer, 2016). Accordingly, we expected that, across increasing levels of suspect-filler similarity, overall suspect

identifications and identification confidence would decrease while filler identifications and perhaps overall choosing would increase.

To the extent that increasing suspect-filler similarity undermines discriminability between lineup members, we expected that fewer accurate identification decisions (i.e., lineup rejections in target-absent cases and suspect identifications in target-present cases) would occur at higher levels of similarity. Indeed, Lucas et al. (2020) reported a decrease in identification accuracy across increases in the number of high similarity match-description fillers included in a lineup, suggesting that any benefits of increasing filler similarity would need to outweigh the increased risk of identification errors. Thus, when assessing the potential benefits of increasing suspect-filler similarity, an important consideration is how variations in suspect-filler similarity affect the trade-off between guilty and suspect identifications. As we have already established, research to date has fallen short of comprehensively delineating the effect of variations in filler similarity on guilty relative to innocent suspect identifications. In what follows, we focus on the results of two very recent studies that used ROC analysis to examine the effects of variations in the similarity of match-description fillers.

Examination of the trade-off between guilty and innocent suspect identifications has focused on the outcomes of ROC analysis to measure the discriminability between innocent and guilty suspects—note that guilty-innocent suspect discriminability is distinct from the more general concept of discriminability referred to earlier (see Lee & Penrod, 2019). Researchers have advocated for different approaches to the derivation of ROC curves. Some have constructed partial ROCs by cumulatively plotted guilty suspect identifications against innocent suspect identification across decreasing levels of confidence and compared partial area under the curve (pAUC) across conditions (Gronlund et al., 2014). Finding issue with the partial ROC approach, others have suggested plotting the various identification decisions

(i.e., suspect and filler identifications, as well as lineup rejections) across different levels of confidence to create full ROC curves—which measure the overall balance of evidence for and against suspect guilt—and comparing these using area under the curve (AUC; Smith et al., 2020).

Lucas et al. (2020) reported no significant differences in either AUC or pAUC across variations in the number of high similarity match-description fillers, suggesting at first glance that perhaps increasing filler similarity affects guilty and innocent suspect identifications to a similar extent. However, it is important to note that the (six) target replacements in target-absent cases were selected to represent the worst case scenario for innocent suspects: being high in similarity to the targets and likely in need of high similarity fillers to protect them from being mistakenly confused with the target. Under these circumstances, increases in suspect filler similarity would presumably be more likely to help protect an innocent suspect than if they were low in similarity to the target (i.e., both because a low similarity innocent suspect is less likely to be confused with the target and because increasing suspect-filler similarity would be unlikely to increase the fillers' similarity to the target)—a notion that was formalized in recent modelling conducted by Colloff et al. (2021). Furthermore, Lucas et al. (2020) flagged their manipulation of filler similarity as potentially protecting innocent suspects to a greater extent than would occur in real cases, as a result of filler similarity being manipulated relative to the targets in both target-present and -absent cases. Thus, their results may well have presented an overly optimistic estimate of discriminability between guilty and innocent suspects across increases in similarity.

In another recent study, Colloff et al. (2021) examined the effect of variations in match-description filler similarity on pAUC. In one of their two experiments they manipulated filler similarity relative to the suspect (i.e., using different fillers in target-present and -absent cases). Consistent with the original propositions of Luus and Wells

(1991) and Wells et al. (1993; see also Oriet & Fitzgerald, 2018)—that increasing filler similarity beyond matching to description risks affecting guilty but not innocent suspect identifications—Colloff et al., (2021) reported a decrease in pAUC as filler similarity increased. They contrasted these results with a second experiment where filler similarity was manipulated relative to the target. Under these circumstances, pAUC increased with filler similarity and supports the argument that manipulating filler similarity relative to the target in both target-present and -absent cases overestimates the protection high similarity fillers provide to innocent suspects (Oriet & Fitzgerald, 2018; Tunnicliff & Clark, 2000). Several features of this study should be borne in mind, however, when evaluating the likely generality of its conclusions. First, although by replicating each of the two experiments twice and pooling the data the researchers were able to achieve admirable sample sizes (more than 10,000 and 9,000 in Experiments 1 and 2, respectively), only a single encoding stimulus (i.e., crime video) was used. Second, the reported differences in pAUC were very small (.003 in both experiments), although they did achieve statistical significance using one-tailed tests.

In addition to interpretative difficulties caused by the filler similarity manipulation and stimulus sampling in Lucas et al. and Colloff et al., respectively, both studies took the standard approach of varying filler similarity by way of similarity ratings provided by pilot participants. With the use of a precise manipulation of suspect-filler similarity and multiple encoding and test stimuli, we sought to provide a more definitive statement about the likely generality of the relationship between suspect-filler similarity and guilty-innocent suspect discriminability. Following Lucas et al. (2020) we conducted this examination under circumstances where the innocent suspects were selected to be at least somewhat confusable with the targets, as this provides a more interesting contrast between the effects on guilty and innocent suspect identifications. As previously mentioned, if innocent suspects are low in similarity to the target there would be little scope for increases in suspect-filler similarity to

protect them. On the other hand, for innocent suspects who are confusable with the target it is of interest to ascertain whether high similarity fillers are effective at minimizing bias against them and what the trade-off between guilty and innocent suspect identifications is under this "worst case" scenario.

We also evaluated the effect of variations in suspect-filler similarity on the confidence-accuracy (CA) relationship. A common approach is to provide a conservative estimate of the CA relationship for suspect identifications by constructing calibration curves including target identifications and all positive identifications from target-absent lineups (e.g., Brewer & Wells, 2006). In cases where an innocent suspect is designated and the sample is sufficiently large, the CA relationship may be assessed for suspect identifications only with a confidence-accuracy characteristic (CAC) analysis (Mickes, 2015). Additionally, the CA relationship can be examined across all identifications, indicating people's ability to assign meaningful confidence to their decisions. In this regard, Lucas et al. (2020) reported calibration analyses characterized by more marked overconfidence with increasing numbers of high similarity fillers—a pattern exacerbated at longer retention intervals—indicating that people are less able to assign realistic confidence judgments to their decisions at high levels of filler similarity. This observation is consistent with the broader decision-making literature stipulating that an increase in task difficulty leads to increasing levels of overconfidence (e.g., Gigerenzer et al., 1991; Weber & Brewer, 2004). Contrary to the effects of filler similarity on the CA relationship for all positive identification decisions, CAC analyses showed that, at short retention intervals, high confidence in suspect identifications remained a good predictor of accuracy across variations in filler similarity (Lucas et al., 2020). These results are consistent with the notion that increasing the difficulty of the identification task decreases the number of high confidence suspect identifications but not necessarily their accuracy (Wixted & Wells, 2017). At long retention intervals, however, the inclusion of any high similarity

fillers undermined the reliability of high confidence suspect picks (Lucas et al., 2020). As our examination of identifications across variations in filler similarity involved a short retention interval, we anticipated an increase in overconfidence across increases in filler similarity for all choosers but expected that high confidence suspect identifications would remain highly accurate.

Finally, we examined the moderating influence of lineup size on the relationship between suspect-filler similarity and identification outcomes. Current recommendations for lineup construction emphasize the importance of effective fillers rather than a particular lineup size (Wells et al., 2020); however, the prospect of standardizing filler characteristics prompts the question of how lineup size impacts on the optimal level of suspect-filler similarity. Similar to the effect of increasing suspect-filler similarity, a larger number of fillers offers more potential for picks to be drawn away from the suspect (Juncu & Fitzgerald, 2021); therefore, we expected that the effects of filler similarity would be greater for larger lineups. To account for the prospect of diminishing returns (i.e., the difference to identification outcomes presented by an increase from 2 to 3 fillers being greater than the increase from 7 to 8 fillers) (Wells et al., 2020), we examined identification outcomes across 2- 3-, and 6-person lineups.

Method

Design and Participants

A 3 (filler similarity: low, medium, high) × 3 (lineup size: 2, 3, 6) × 2 (target presence: present, absent) × 4 (stimulus: 1–4) mixed design was used. Participants attempted to identify each target from separate lineups randomly assigned to be target-present or -absent and contain match-description fillers that were low, medium or high in similarity (i.e., unmorphed, a 33% morph or a 50% morph) to the suspect. Lineup size was manipulated between-subjects.

Data were collected online through Mechanical Turk. Although this approach could result in noisier data than a laboratory sample, our general experience of sourcing data online has been that sensible results are obtained when the sample is robust and stringent exclusion criteria are applied. We targeted 200 data points per cell (i.e., a sample of 3,600 participants making four identification decisions each), allowing discriminability between guilty and innocent suspects across filler similarity conditions to be examined separately for each stimulus set and at each different lineup size ($N \sim 800$). A total of 4,614 unique data files were obtained; 148 additional attempts to participate were blocked due to the use of a mobile device or an incorrect response to a screening question. Exclusions ($N = 1{,}018$) included people who failed more than one video attention check ($N = 501$; further details below), data files contaminated by previous attempts to participate ($N = 249$), incomplete datafiles ($N = 225$), and miscellaneous other reasons ($N = 43$). Participants included in the final sample ($N = 3{,}596$) were 2,160 male, 1,427 female and 9 other, aged 18 to 87 ($M = 36.0$; $SD = 11.1$). The sample was 70.8% White, 16.0% Black, 6.4% Asian, 4.2% Hispanic and 2.6% other. Average participation time was 9.3 minutes ($SD = 3.2$) and participants were paid an honorarium of $1.50 US. This research was approved by our Institutional Ethics Review Board.

Materials

Videos and Attention Checks

Four short videos, ranging from 39 to 58 s in length, served as the stimulus events. In each video the target person is shown committing a non-violent crime (stealing files from a computer, breaking into a house, purchasing drugs or stealing a handbag), with their face in clear view for 30 s at an apparent distance of approximately 5 m or less. As this experiment was conducted online, making it difficult to ensure the stimulus videos were watched, the following attention check procedure was implemented. A still photograph of an easily recognizable animal (a dolphin, dog, penguin or elephant) was displayed onscreen for 2 s at

the end of each video. Participants were then asked to select from a list of four options which animal had been shown. As previously mentioned, data from those participants who failed more than one video attention check were excluded from further analyses.

Lineups

Filler and innocent suspect selection. For each stimulus set, five filler faces (used in the low similarity filler condition) and an innocent suspect (used in target-absent lineups) were selected as follows. Ten people watched the stimulus videos and provided descriptions of the targets. From these, the most defining features of each target (i.e., descriptors mentioned by at least five people) were included in a "modal description" (cf. Horry et al., 2012). All of the modal descriptions specified the sex and race of the perpetrator, as well as an age range and some level of detail regarding hair. The age range spanned all estimates provided by those who described the perpetrator. Descriptions of the perpetrator's hair ranged from vague (i.e., dark hair) to quite specific (i.e., short brown hair). Build and eye color were mentioned once each. For the full descriptions see Supplemental Material (p. 1).

With reference to the modal descriptions, 20+ filler candidates were chosen for each stimulus set. In addition to matching description, fillers were selected to match the targets on features necessary for the images to be morphed to one another. Broadly, this meant the fillers had to not be showing their teeth, have both ears visible and the images needed to not be too blurry or over-exposed. Additionally, the filler candidates for each of the female stimulus sets were required to have their hair pulled back and the fillers for one of the male candidates were required to have light stubble. Approximately 15,000 database images (i.e., from online mugshot and other face databases, as well as locally sourced face databases provided to us by other researchers) were searched by a research assistant to obtain suitable filler pools, with roughly 400 match-description images deemed unsuitable for morphing.

The first author also reviewed the final filler pool and removed any potential fillers that did not meet the required criteria.

For each stimulus set, a match-description innocent suspect known to the researcher team (i.e., from the same pool of candidates as the targets) was selected on the basis that they matched-description and could potentially be somewhat at least somewhat confusable with the target. We also included a "backup" innocent suspect candidate in piloting.

Pilot participants (N = 305) viewed each target description and sorted the filler and innocent suspect candidates, as well as the target and one face clearly not matching description, into "matching description" and "not matching description" categories; participants who failed to classify correctly all four faces clearly not matching description (N = 151) were excluded from further analyses. The rate at which faces were categorized as matching description ranged from 27.9 to 87.0% *(*M = 65.6, SD = 15.4). Similarity between the fillers and both the targets and innocent suspects (as well as the innocent suspects to the targets) was indexed using *Betaface* face matching software. The range of similarity was from 54.0 to 80.0 (M = 64.8, SD = 4.4) on a scale from 0 to 100.

The similarity and match-description data were used to select fillers that were as low as possible in similarity to both the innocent and guilty suspects while also providing a reasonable match to description. Note that this procedure prevented the inclusion of fillers that were low in similarity to the innocent suspect but (by chance) a good likeness to the target, but avoided introducing different fillers across target-present and -absent cases. We were able to select five fillers for each stimulus set that were ranked in the bottom 50% of similarity to both the target and innocent suspect while being categorized as matching description by at least 50% of the pilot participants. Fifty percent might seem like a lax criterion for matching description; however, recall that all faces in the filler pool were initially judged as matching description by a research assistant and the first author. Therefore,

the low match-description percentages afforded to some fillers may well speak to the subjective nature of match-description judgements (e.g., what constitutes "light brown" hair). We also suspect that including only one face per stimulus in pilot testing that clearly did not match description supressed the match-description percentages for some faces: that is, the pilot participants likely felt compelled to "rule out" a substantial number of faces as matching description. Similarity of the fillers to the targets ranged from 56.9 to 64.8 (M = 61.3, SD = 2.3) and similarity of the fillers to the innocent suspects ranged from 54.0 to 67.4 (M = 61.5, SD = 3.3). Similarity between the innocent suspects and targets was 67.5, 69.0, 65.3 and 60.7 for Stimulus Sets 1 to 4, respectively, suggesting that apart from Stimulus 4 the innocent suspects were higher in similarity to the target than the fillers. Since the Stimulus 4 innocent suspect happened to have a similar hairstyle to the target we suspected they might also be more confusable with the target than the fillers. For a full breakdown of match-description percentages and suspect-filler similarity indices for each filler candidate, see Supplemental Material (pp. 2–5).

Operationalizing suspect-filler similarity. To increase suspect-filler similarity, fillers were morphed to the suspects using *Fantamorph*. That is, the fillers in each stimulus set were morphed to both the target and innocent suspect, with a different suspect photograph used for each morph. Prior to creating the morphs, the suspects were photographed wearing a variety of hairstyles (i.e., most importantly, hair parted in different places), so that each filler could be matched to a photograph of the suspect that was most likely to produce a seamless morph. Images in the same stimulus set were also adjusted within 10% brightness of each other, since the filler photographs tended to be overexposed relative to the suspect photographs. To hide evidence of the morphing procedure—often apparent around the neck, shoulders and outer edges of the hair—final lineup images were edited to show only the face and hair around the head on a white background. It is possible that some of the procedures

described above increased the lowest level of suspect-filler similarity beyond what would normally be achieved by matching to description; however, note that contemporary lineup construction recommendations suggest that background, lighting conditions etc. should be edited to be uniform across lineup members (Wells et al., 2020).

Addressing the potential concern that the suspects would stand out as the only non-digitally altered face in lineups where the fillers were morphed, we created morphed versions of the suspect faces (e.g., two images of Target 1 were morphed to each other). This process did not alter suspect appearance beyond giving the image quality the same degree of "softness" as the other morphed images. Additionally, a mock-witness test was used to examine whether the suspects were more likely to be chosen from morphed (vs. non-morphed) lineups. Pilot participants (N = 501) viewed the target description and guessed who the suspect was from the a 6-person version of either the target-present or -absent lineup for each stimulus set. The lineup fillers were either unmorphed, a 33% morph, or a 66% morph to the suspect. The frequencies with which each lineup member was selected, as well as lineup effective size calculations (Tredoux, 1998), are shown in Supplemental Material (pp. 6–9). There was no evidence of the suspects standing out in the morphed lineups compared with the unmorphed lineups. Instead, as would be expected for increases in similarity, there was some evidence of choosing being more evenly spread across lineup members as filler similarity increased. For one target-present lineup (Stimulus 2) and one target-absent lineup (Stimulus 3) the suspect was selected at higher than chance rates when the fillers were unmorphed; note that the Stimulus 3 innocent suspect was eventually replaced by the backup innocent suspect (further details below).

To determine which morph levels to use in the experiment, we first piloted whether a 33% morph—roughly a 7 to 8% increase in the *Betaface* similarity index—had any effect on identification patterns compared with unmorphed fillers. Participants (N = 760) made

identification decisions for each set of stimuli from 6-person lineups randomly allocated to be target-present or -absent and contain either morphed or unmorphed fillers. As suspect-filler similarity increased, in target-present cases suspect identifications decreased from .57, 95% CI [.53, .60] to .51, 95% CI [.48, .55], filler identifications increased from .11, 95% CI [.09, .14] to .26, 95% CI [.22, .29] and lineup choosing increased from .68, 95% CI [.65, .71] to .77, 95% CI [.74, .80]; in target-absent cases suspect identifications decreased from .21, 95% CI [.18, .24] to .18, 95% CI [.15, .21], filler identifications increased from .22, 95% CI [.19, .24] to .29, 95% CI [.26, .32] and lineup choosing increased from .43, 95% CI [.39, .46] to .48, 95% CI [.44, .51]. Of additional note was that some identification patterns diverged from what was suggested by their similarity indexes, showing that although such indices generally do a good job of measuring face similarity they are not perfect. In particular, despite having a high similarity rating, the innocent suspect for Stimulus 2 was rarely identified[1]—much less, in fact, than one of the supposedly low similarity fillers. As Stimulus 2 had notably higher baseline filler identifications than the other stimulus sets, we replaced the filler drawing a high number of picks with another low similarity face that 40.9% of participants rated as matching description. Also, as three of the four innocent suspects were drawing a low number of picks (7–12%), we took steps to create more scope for innocent suspect identifications to decrease. The innocent suspect for Stimulus 3—previously noted as being selected at higher than chance rates in the lineup fairness pilot—was replaced with its backup innocent suspect, which was higher in similarity to the target (with a *Betaface* index of 76.8 compared with a *Betaface* index of 65.3). Replacing the innocent suspect in Stimulus 3 also necessitated the replacement of two fillers to ensure that suspect-filler similarity in the low similarity condition was as low as possible in both target-present and -absent cases (note that match-description percentages for the replacement fillers were 49.4% and 67.5%).

Satisfied that the 33% morph caused a detectable increase in similarity from the unmorphed fillers, we aimed to further increase suspect-filler similarity by a comparable amount. *Betaface* similarity indexing suggested that the difference between a 33 and 50% morph (M = 7.3, SD = 2.2) was roughly comparable to the difference between a 0 and 33% morph (M = 7.8, SD = 2.0). Thus, we proceeded with 0%, 33% and 50% morphs in the low, medium and high similarity conditions. Average suspect-filler similarity indices in these conditions were 61.8 (SD = 3.1), 69.6 (SD = 3.7) and 76.9 (SD = 3.8), respectively. In target-absent cases, average target-filler similarity indices were 61.4 (SD = 2.8), 63.3 (SD = 3.1) and 64.5 (SD = 3.4) in the low, medium and high similarity filler conditions, suggesting only a marginal increase in similarity between the fillers and the target as innocent suspect-filler similarity increased. Individual similarity indices for each filler in each similarity condition are reported in Table 1.

Procedure

Participants were provided with an information sheet and consent statement to read prior to commencing the study. Those who passed the screening question to access the experiment then provided basic demographic information (i.e., indicated their age, sex, race and whether they had normal or corrected to normal vision) and completed four identification tasks. For each stimulus set, participants (a) viewed the stimulus video, (b) answered the multiple-choice attention check question, (c) made an identification decision, and (d) recorded confidence in their decision. Each stimulus video was automatically screened following onscreen instructions to advance the screen when the participant was ready to begin watching. After the video finished, the corresponding attention check question was displayed, with the response options randomized. Lineup instructions were then displayed, cautioning that the target may or may not be present and describing the procedure for recoding an identification decision. The 6-person lineups were displayed in two rows of

three, with lineup member position randomized and a "Not Present" option presented to the right of the top row. The 2- and 3-person lineups were displayed with the lineup members in one row, with the Not Present option on the right-hand end of the row. Suspect position was counterbalanced, with the filler places randomly populated from the pool of fillers used in the 6-person lineups. The images were cropped to 130 × 160 pixels, the largest size possible without causing the images to display in different sizes across lineup size conditions. After the identification decision was recorded, people were prompted to record confidence in their decision on an 11-point decile scale ranging from 0% (absolutely uncertain) to 100% (absolutely certain). On completing the study, participants were provided with study feedback and instruction on how to receive their honorarium.

Results

For each analysis the overall data were examined (i.e., treating each decision as an independent trial), as well as the first trial data and the overall data at stimulus level. This approach to analysing the data was taken after it became apparent that accounting for variation across participants and stimuli within the one analysis would not be possible; that is, mixed-effects models with the requisite random effects structure were unable to be fitted. Here we report the analyses on the overall data, noting whether results differed for the first trial data or at stimulus level.

Identification Patterns and Confidence

Overall identification decision patterns (i.e., suspect identifications, filler identifications and overall choosing) and confidence are reported in Table 2; a more detailed version of this table (i.e., showing the totals across filler similarity, target presence and lineup size), the first trial data and patterns at stimulus level appear in Supplemental Material (pp. 10–27). The effects on identification decision patterns of suspect-filler similarity, target presence, lineup size and all possible interactions were examined with loglinear analysis

using IBM SPSS for Windows (version 22). The confidence data violated normality assumptions and were thus examined with robust ANOVA using the WRS2 package (version 1.0-0; Mair & Wilcox, 2019) in R (version 3.6.1, R Development Core Team, 2019).

For the overall data, the 4-way association between suspect-filler similarity, target presence, lineup size and identification decision was non-significant, $k = 4$, LR $\chi^2(8) = 4.29$, $p = .831$, as were all 3-way associations between the variables, $k = 3$, LR $\chi^2(20) = 15.47$, $p = .749$. The test for 2-way associations was significant, $k = 2$, LR $\chi^2(18) = 2600.95$, $p < .001$, with partial associations indicating significant effects on identification decisions of suspect-filler similarity, *partial* $\chi^2(4) = 467.33$, $p < .001$, lineup size, *partial* $\chi^2(4) = 537.07$, $p < .001$, and target presence, *partial* $\chi^2(2) = 1620.01$, $p < .001$. Examinations of each identification decision type (i.e., suspect picks, filler picks and overall choosing) were then conducted for each of the significant main effects (adjusted alpha .05/3 = .017).

Suspect-filler similarity affected suspect identifications, $\chi^2(2) = 99.52$, $p < .001$, filler identifications, $\chi^2(2) = 432.48$, $p < .00$, and choosing, $\chi^2(2) = 55.58$, $p < .001$. Suspect identifications decreased from .49, 95% CI [.47, .50], in the low similarity condition to .44, 95% CI [.43, .46] in the medium similarity condition and .39, 95% CI [.37, .40], in the high similarity condition (OR low vs. high = 1.51; OR low vs. medium = 1.19; OR medium vs. high = 1.27). Filler identifications increased from .12, 95% CI [.11, .13], in the low similarity condition to .21, 95% CI [.20, .22], in the medium similarity condition and .29, 95% CI [.28, .30], in the high similarity condition (OR low vs. high = 3.06; OR low vs. medium = 2.02; OR medium vs. high = 1.51). Choosing increased from .60, 95% CI [.59, .62], in the low similarity condition to .65, 95% CI [.64, .67] in the medium similarity condition and .67, 95% CI [.66, .69], in the high similarity condition (OR low vs. high = 1.36; OR low vs. medium = 1.25; OR medium vs. high = 1.09).

Target presence also affected suspect identifications, $\chi^2(1) = 1512.53$, $p < .001$, filler identifications, $\chi^2(1) = 114.41$ $p < .001$, and choosing, $\chi^2(1) = 978.15$, $p < .001$. As one might expect, suspect identifications were higher in target-present, .60, 95% CI [.59, .61], than -absent, .28, 95% CI [.27, .29], cases (OR = 3.89). Overall choosing was also higher in target-present, .77, 95% CI [.76, .78], than -absent, .52, 95% CI [.51, .53] cases (OR = 3.09), while filler identifications were higher in target-absent, .24, 95% CI [.23, .25], than -present, .17, 95% CI [.16, .18], cases (OR = 1.56).

Finally, lineup size affected identifications of suspects, $\chi^2(2) = 109.34$, $p < .001$ and fillers, $\chi^2(2) = 362.94$, $p < .001$, as well as choosing, $\chi^2(2) = 18.55$, $p < .001$. As lineups expanded from two to three and then six members, suspect identifications decreased from, .49, 95% CI [.48, .51], to .46, 95% CI [.44, .47] and .37, 95% CI [.35, .38] (OR 2 vs. 6 = 1.70; OR 2 vs. 3 = 1.16; OR 3 vs. 6 = 1.46), filler identifications increased from .13, 95% CI [.12, .14], to .18, 95% CI [.17, .19] and .31, 95% CI [.29, .32] (OR 2 vs. 6 = 2.02; OR 2 vs. 3 = 1.53; OR 3 vs. 6 = 3.09), and choosing increased marginally from, .62, 95% CI [.61, .63], to .64, 95% CI [.62, .65] and .67, 95% CI [.66, .69] (OR 2 vs. 6 = 1.26; OR 2 vs. 3 = 1.07; OR 3 vs. 6 = 1.17).

Identification patterns were the same for the first trial data (see Supplemental Material, pp. 28–29), but varied somewhat at stimulus level. The most notable deviation from the overall pattern of results was an interaction between suspect-filler similarity and target presence on suspect identifications for Stimulus Set 4. In target-absent cases, increasing suspect-filler similarity decreased suspect identifications; however, in target-present cases, suspect identifications increased with suspect-filler similarity. Additionally, for Stimulus Set 1 suspect-filler similarity did not affect choosing patterns and for Stimulus Sets 1 and 3 lineup size did not affect choosing. For the full set of analyses at stimulus level refer to Supplemental Material (pp. 29–37).

Overall identification confidence was affected by suspect-filler similarity, test statistic = 79.60, $p < .001$, and lineup size, test statistic = 105.84, $p < .001$, but not target presence, test statistic = 0.17, $p = .677$, $d = 0.01$. As suspect-filler similarity increased, confidence decreased marginally from $M = 75.96$, $SD = 20.21$, in the low similarity condition to $M = 73.89$, $SD = 21.04$, in the medium similarity condition and $M = 72.03$, $SD = 21.24$, in the high similarity condition (d low vs. high = 0.19; d low vs. medium = 0.10; d medium vs. high = 0.09). Confidence decreased as lineup size increased from two ($M = 76.12$, $SD = 19.95$) to three ($M = 74.68$, $SD = 20.57$) to six ($M = 71.22$, $SD = 21.80$) (d 2 vs. 6 = 0.23; d 2 vs. 3 = 0.07; d 3 vs. 6 = 0.16). None of the interactions were significant (test statistics < 8.78, ps > .067).

Confidence results or the first trial data, summarized in Supplemental Material (p. 29), were the same as the overall data. At stimulus level, in addition to main effects in line with the overall data, there was a 3-way interaction between suspect-filler similarity, target presence and lineup size on confidence for Stimulus Set 1, and a 2-way interaction on confidence between similarity and target presence for Stimulus Set 2; for details see Supplemental Material (pp. 31, 33, 34–35, & 36–37).

Accuracy

For the overall data, the 4-way association between the independent variables and accuracy was non-significant, $k = 4$, LR $\chi^2(4) = 4.44$, $p = .350$. The test for 3-way associations was significant, $k = 3$, LR $\chi^2(12) = 24.42$, $p = .018$, with partial associations indicating an interaction between lineup size and target presence on accuracy, $\chi^2(2) = 11.44$, $p = .003$. As lineup size increased, accuracy decreased to a greater extent in target-present than -absent cases. In target-absent cases, accuracy decreased from 51.60%, 95% CI [49.57, 53.59], in 2-person lineups to 49.29%, 95% CI [47.27, 51.27], in 3-person lineups and 43.44%, 95% CI [41.45, 45.38], in 6-person lineups, $\chi^2(2) = 34.33$, $p < .001$(OR 2 vs. 6 =

1.39; OR 2 vs. 3 = 1.10; OR 3 vs. 6 = 127). In target-present lineups, accuracy decreased from 66.21%, 95% CI [64.29, 68.09], in 2-person lineups to 62.18%, 95% CI [60.20, 64.12], in 3-person lineups and 51.76%, 95% CI [49.75, 53.73], in 6-person lineups, $\chi^2(2) = 111.48$, $p < .001$ (OR 2 vs. 6 = 1.83; OR 2 vs. 3 = 1.19; OR 3 vs. 6 = 1.53). The test for 2-way associations was also significant, $k = 2$, LR $\chi^2(13) = 482.74$, $p < .001$, with partial associations indicating an effect of suspect-filler similarity on accuracy, $\chi^2(2) = 140.26$, $p < .001$. Accuracy decreased from 59.84%, 95% CI [58.44, 61.22], in the low similarity condition to 54.28%, 95% CI [52.86, 55.68] in the medium similarity condition and 47.90%, 95% CI [46.48, 49.30], in the high similarity condition (OR low vs. high = 1.62; OR low vs. medium = 1.26; OR medium vs. high = 1.29).

Differences in accuracy patterns for the first trial and stimulus level data were as follows. For Stimulus Set 1, accuracy patterns were characterized by a 3-way interaction between filler similarity, target presence and lineup size. For the first trial data, as well as the overall data for Stimulus Sets 2 and 3, accuracy decreased as lineup size increased without interacting with target presence. For Stimulus Set 4, there was an interaction between similarity and target presence on accuracy. In target-absent cases, accuracy trended downwards as filler similarity increased, but in target-present cases accuracy trended upwards. For the full first trial and stimulus level accuracy analyses see Supplemental Material (pp. 37–40).

Discriminability Between Guilty and Innocent Suspects

Full ROCs were constructed by cumulatively plotting identification decisions in target-present against -absent cases across different levels of confidence (Smith et al., 2020). To maximize cell sizes in each condition we collapsed data into 90–100%, 80–70% and 60–0% confidence categories. With the objective of ordering identification evidence from strongest to weakest, we plotted suspect identifications across descending levels of

confidence, followed by filler identifications and lineup rejections across ascending levels of confidence. Low confidence filler identifications were followed by low confidence lineup rejections and so forth. Our decision to order filler identifications before lineup rejections was based on the premise that although filler identifications and lineup rejections have been indicated to provide equivalently strong exonerating evidence, the latter is likely given more weight by investigators (Wells et al., 2015). The relevant ROC curves appear in Figure 1. For each ROC curve, both overall AUC and pAUC spanning suspect identifications were computed, and pairwise comparisons (adjusted alpha .05/3 = .017) were conducted using the pROC package (Robin et al., 2011) in R (version 3.6.1; R Development Core Team, 2019).

Across all trials, AUC was lower in the high similarity filler condition, AUC = .665, 95% CI [.649, .681], than the low similarity filler condition, AUC = .701, 95% CI [.686, .715], p = .001, D = 3.35. AUC in the medium similarity filler condition, AUC = .685, 95% CI [.670, .700], fell partway between the low and high conditions but did not differ significantly from either, p = .136, D = 1.49 and p = .065, D = 1.85, respectively. The pattern of AUC decreasing as filler similarity increased was evident for the 3- and 6-person lineups, but not for the 2-person lineups (see Supplemental Material, pp. 41–43). A decrease in AUC as filler similarity increased was evident for Stimulus Sets 1 to 3, but for Stimulus Set 4 AUC instead increased with filler similarity (see Supplemental Material, pp. 45–48). Further, when examining only the first trial data filler similarity did not affect AUC (see Supplemental Material, p. 44). Lineup size did not affect AUC (see Supplemental Material, p. 49).

Across all trials, pAUC = .121, 95% CI [.113, .128], .119, 95% CI [.111, .126], and .114, 95% CI [.106, .121], in the low, medium and high similarity conditions, respectively, with no significant differences between conditions (ps > .213, Ds < 1.25). These results did not differ for lineup size or the first trial data (see Supplemental Materials, pp. 41–44). At stimulus level, as filler similarity increased there was evidence of pAUC decreasing for

Stimulus Sets 1, 2 and 3 and increasing for Stimulus Set 4 (see Supplemental Material, pp. 45–48). However, pAUC did not vary depending on lineup size (see Supplemental Material, p. 49). In other words, the patterns were similar for both the AUC and pAUC approaches, although only the AUC revealed statistically significant differences across similarity conditions (when examining all trials).

The CA Relationship

To examine the CA relationship, calibration curves were constructed by plotting confidence against accuracy for all cases (Juslin et al., 1996; Lucas et al., 2020). All curves are presented with the 11 confidence categories collapsed into five (0–20%, 30–40%, 50–60%, 70–80%, 90–100%). Separate curves were constructed for all choosers (i.e., fillers and suspect identifications in both target-present and -absent cases) and non-choosers (i.e., lineup rejections); the non-chooser curves, showing the typical lack of a CA relationship, are reported in Supplemental Material (p. 50). In this instance, the chooser curves were also largely absent a CA relationship (i.e., the curves were notably flat). Resolution was not improved when plotting confidence against accuracy for target-present trials only (see Supplemental Material, p. 51), shedding doubt on the possibility that innocent suspect characteristics undermined the CA relationship.

Despite the flatness of the chooser curves, they were characterized by more marked overconfidence as filler similarity increased for the overall dataset (see Figure 2), and this pattern was present across all lineup sizes (see Supplemental Material, pp. 52–54). Evidence of overconfidence increasing with filler similarity was also evident for the first trial data, as well as at stimulus level (with some variation in the degree of overconfidence across stimuli; see Supplemental Material, pp. 55–59). Similar to the effect of filler similarity, increasing lineup size was associated with more marked overconfidence, again for notably flat calibration curves (see Supplemental Material, p. 60).

CAC analyses were conducted by plotting confidence against accuracy for suspect identifications (Mickes, 2015). For the overall data (see Figure 3) the accuracy of suspect identifications made with 100% confidence did not vary depending on filler similarity. However, high confidence accuracy was fairly low (~73%) and not much higher than accuracy at lower levels of confidence. At medium and high levels of similarity in particular, the curves flattened across the top half. The pattern of high confidence accuracy remaining similar across variations in filler similarity was stable across different lineup sizes but varied somewhat for the first trial data and across stimuli (see Supplemental Material, pp. 61–68). Finally, lineup size did not lead to variation in CACs (see Supplemental Materials, p. 69).

Discussion

Based on extant research concerning the relationship between suspect-filler characteristics and identification performance, current lineup construction recommendations require lineup members to match the witness's description of the perpetrator but are necessarily vague regarding the optimal level of suspect-filler similarity (Wells et al., 2020). A key contributor to the difficulty in defining ideal lineup member characteristics may be the imprecision of the typically used approaches to operationalizing suspect-filler similarity in laboratory studies. Hence, the present study focused on examining identification outcomes across precise and replicable variations in suspect-filler similarity using multiple sets of encoding and test materials.

With some exceptions, the observed identification and confidence patterns tended to be consistent with what we expected to see based on previous research (e.g., Fitzgerald et al. 2013; Lucas et al. 2020). Across all identification decisions in the sample, increasing filler similarity beyond matching a basic description of the perpetrator caused the expected shift from suspect to filler identifications and increased overall choosing from lineups. However, although the predicted decrease in confidence as filler similarity increased was observed, this

effect was notably small. Further, there was no significant variation in the effects of filler similarity depending on target presence or lineup size. It is possible that our variations in lineup size (i.e., 2, 3 and 6-person lineups) were simply not large enough to influence the effects of suspect-filler similarity. With regard to target presence, we suspect that the innocent suspects selected to be at least somewhat confusable to the targets—which in some cases ended up being highly confusable with the targets—caused increases in suspect-filler similarity to be more protective than if the innocent suspects had been low in similarity to the targets and, therefore, similar in magnitude to the (unwanted) protection afforded to the guilty suspects. Of particular note is that while increases in suspect-filler similarity tended to offer innocent suspects some protection, these still stood out (sometimes a lot) in the high similarity filler condition, suggesting that increasing the similarity of fillers to the innocent suspect is a largely ineffective means for addressing bias against innocent suspects who are confusable with the target. The identification decision and confidence results were the same for the overall and first trial (experienced by witnesses) data; however, analyses at stimulus level revealed some variation in the effects of filler similarity. The most notable deviation from the overall pattern of results was that, for Stimulus 4, increased suspect-similarity decreased innocent suspect identifications but inexplicably increased guilty suspect identifications.

Regarding the other identification outcome measures examined, overall decision accuracy decreased as filler similarity increased, while overconfidence increased for all choosers but not suspect identifications. Apart from the unexplained flatness of the CA curves, which we discuss in further detail below, these results are consistent with what has been reported in previous research (Lucas et al., 2020). While the interaction between suspect-filler similarity and target presence on suspect identifications was non-significant, there was some evidence of variations in suspect-filler similarity affecting the more sensitive

measures of discriminability between guilty and innocent suspects. Across all trials, pAUC did not differ significantly depending on filler similarity. There were, however, differences in pAUC at stimulus level, underscoring the critical importance of testing effects across multiple stimuli. As filler similarity increased, pAUC decreased for Stimuli 1–3 and increased for Stimulus 4. Further, across all identification decisions—as well as for three of the four individual stimuli (i.e., again Stimuli 1–3)—AUC was significantly lower in the high than the low similarity condition. There were, however, no meaningful differences in AUC when examining only the first trial data and, for Stimulus 4, AUC was higher in the high compared with the low similarity condition. Overall, the ROC analyses suggest that increasing suspect-filler similarity beyond a match to description undermines discriminability between guilty and innocent suspects—more strongly than previous research using this measure of identification performance has indicated (Colloff et al., 2021; Lucas et al., 2020). Recall that Lucas et al. found no evidence of discriminability being affected by variations in suspect-filler similarity, but that protection to innocent suspects may have been overestimated due to the use of high similarity innocent suspects and filler similarity being manipulated relative to the targets (i.e., in both target present and -absent cases). Colloff et al. (2021), on the other hand, reported a decrease in discriminability across increases in suspect-filler similarity; however, in addition to detecting only very small effect (significant using a one-tailed test) using a measure of discriminability that other researchers have criticized (e.g., Smith et al., 2020; Starns et al., 2021), only one stimulus set was examined.

Returning briefly to the unexplained lack of a CA relationship for choosers, there are several important things to note. Since the innocent suspects tended to stand out from the other lineup members, even in the high similarity fillers conditions, our lineups ended up not meeting the Wixted & Wells (2017) criteria for expecting high confidence suspect identifications to be highly accurate. However, since the innocent suspects tended to not

stand out in the *a priori* tests of lineup fairness (i.e., the criterion on which lineup bias can be judged in real cases) the lack of a CA relationship under these circumstances poses a problem for those looking to draw firm conclusions about the likely guilt of a suspect identified with high confidence (Brewer et al., 2021; Smith et al., in press). Further, the absence of a CA relationship was evident in the calibration curves that included only data from target-present cases, suggesting that it was not necessarily the innocent suspect characteristics (or these alone) causing the issue. Thus, our results reinforce previous statements that although confidence has generally been observed to be a good predictor of identification accuracy (e.g., Brewer & Wells, 2006), there are conditions—which are often undefined—where this relationship breaks down and investigators need to be cautious about placing too much value on the confidence with which a suspect identification is made (Sauer et al., 2019).

On balance, identification performance in the present study was clearly best at the lowest level of filler similarity, a condition created by selecting fillers that were as dissimilar to the suspect as possible while still matching a basic perpetrator description. Of particular importance is that discriminability between guilty and innocent suspect identifications declined with an increase in filler similarity despite innocent suspects being selected to be at least somewhat confusable with the target. As previously outlined, increasing filler similarity under these circumstances would be expected to be more likely to protect the innocent suspects than if they were low in similarity to the targets; therefore, it seems possible (likely even) that the present study underestimates the negative impact of increases in suspect-filler similarity above matching to description. Other studies have also drawn attention to the likely variation in the relationship between suspect-filler similarity and identification outcomes depending on target-innocent suspect similarity. We followed Lucas et al.'s (2020) lead in choosing to focus on cases where the innocent suspects were high in similarity to the targets and thus be more likely to be protected by increases in suspect-filler similarity. Colloff et al.

(2021) also noted that the impact of suspect-filler similarity likely varies depending on target-innocent suspect similarity and chose a "median" similarity innocent suspect to represent the overall impact of increases in filler similarity across a range of innocent suspect characteristics. However, the moderating influence of innocent suspect-target similarity on the relationship between suspect-filler similarity and identifications has not been formally tested and should be priority for future research.

The implication of our results is consistent with what researchers (Luus & Wells, 1991) first proposed and then demonstrated in their landmark study (Wells et al., 1993) almost three decades ago: that perhaps lineup constructors should be cautioned against increasing the general similarity between lineup members beyond matching them on a few key descriptors. While the original argument for a "match-description only" approach to lineup construction left open the possibility that there may be higher (but as yet undefined) levels of suspect-filler similarity that could improve identification outcomes, the present research, with its broad stimulus sampling and carefully operationalized similarity manipulations, would appear to provide the most convincing evidence to date that it may not be possible to secure any additional benefits by increasing filler similarity beyond a match to description. Even when manipulating suspect-filler similarity as precisely as possible across a variety of encoding and test stimuli—and in a way that could facilitate lineup constructors consistently targeting a desirable level of similarity—we observed no benefits to increasing filler similarity beyond matching to description. Some might wonder whether suspect-filler similarity could be measured at even smaller increments than was done in the present study and whether this might be further illuminating (e.g., a smaller increase in filler similarity than a 30% morph might show an increase in identification performance). While we do not have a definitive answer to this possibility, the following seems worthy of note. Although a 30% morph may sound like a substantial increase in filler similarity, our subjective impression of

the stimuli—reinforced by the *Betaface* values and demonstrated differences in identification patterns—was that our increases in filler similarity were modest.

The conclusion that suspect-filler similarity ought to be minimized is based on our evaluation of identification performance. Also worthy of consideration is the plight of innocent suspects who are unlucky enough to resemble the perpetrator; our results indicate they are at an alarmingly high risk of being misidentified from lineups containing low similarity fillers. The dire consequences associated with the misidentification of innocent suspects could form the basis for arguing that suspect-filler similarity should be increased despite the resultant reduction in overall identification performance and despite it appearing to fall short of adequately reigning in bias against innocent suspects. Important to remember, however, is that cases where innocent suspects resemble the target are likely often distinguished by characteristics which prompt a different, more protective, lineup construction approach. For example, if someone becomes a suspect because they resemble CCTV footage or a composite sketch of the perpetrator, lineup construction recommendations clearly specify that fillers should be selected using the same criteria (Wells et al., 2020). Reinforcing this recommendation, research where fillers are selected relative to their similarity to the target suggests that high similarity fillers significantly reduce bias against high similarity innocent suspects (e.g., Lucas et al. 2020). In this context, the general use of low similarity match-description fillers seems less problematic.

If increasing suspect filler similarity beyond a match to description is indeed counterproductive, then the quality of the perpetrator description is of utmost importance. In this regard, it is important to note that the present study—along with most studies examining the effects of variations in filler similarity beyond a match to description (though note Wells et al., 1993)—approximated matching to description by matching the fillers to a general description of the perpetrator rather than to each participant's individual description. That is,

perpetrator descriptions were compiled using the most common descriptors from 10 different descriptions. In another example, Colloff et al. (2021) reported creating a perpetrator description "consistent with the descriptions" of 91 Mechanical Turk workers (Supplemental Material, p.22). While these methodologies could be argued to result in key features of each target (i.e., important to match the fillers on) being included in the perpetrator description—a possibility that remains unconfirmed—they do not clarify what constitutes an adequate witness's description or, importantly from a practical perspective, how this could be reliably obtained. Therefore, although advocating for a match-description-only approach to lineup construction seems appropriate based on current data, it is vital that researchers first give further consideration to the appropriate contents of a perpetrator description.

Traditionally, a perpetrator description constitutes details spontaneously reported by the witness, with no specific criteria regarding what should be included. A key assumption here is that the details that are clear in the witness's memory will be those that are reported by the witness and then used as the basis for selecting lineup members (Wells et al., 1998). However, while it seems likely that witnesses will volunteer distinctive details (e.g., a bald head or a beard), general details may often fail to be reported (e.g., approximate age or build). Indeed, as part of a recent examination of real police lineups—which revealed a concerning proportion that were unfairly biased against the suspect—Steblay and Wells (2020) noted that many extremely impoverished witness descriptions (e.g., stating only the sex and race of the perpetrator) likely omitted important details that the witness could have easily reported but did not think to mention without prompting. Therefore, although interviewers are strongly advised against seeking out specific information by leading the witness, as this might prompt the reporting of incorrect details (Wells et al., 2020), the importance of a complete perpetrator description has led to the recent suggestion that it may be important for a pre-specified minimum set of descriptors to be elicited (Steblay & Wells, 2020). In this vein,

interviewing techniques that increase the quality of perpetrator descriptions have been developed (e.g., Demarchi & Py, 2009; Satin & Fisher, 2019) and the most recent current lineup construction recommendations strongly urge investigators to use such techniques to flesh out perpetrator descriptions (Wells et al. 2020). However, the links between a more detailed description (or the reporting of particular details), the optimal approach to obtaining that description without leading the witness, the quality of the lineup construction and identification performance are yet to be explored and may reveal additional important considerations for obtaining perpetrator descriptions that can guide selection of lineup fillers.

In addition to investigating the essential contents of a perpetrator description, future research should seek to determine whether lineup constructors' subjective judgements are sufficient for minimizing suspect-filler similarity. Instead, perhaps face-matching technology—such as what was used in the present study—could be used to select the lowest similarity faces from a match-description filler pool.

Although there are several important avenues of research left to follow in the pursuit of defining optimal filler characteristics, lineup constructors are faced with the task of selecting fillers based on the current knowledge base. In this regard, the tech-driven approach we used to manipulate suspect-filler similarity and envisaged as a potential means for investigators to reliably target an optimal level of suspect-filler similarity, appears to have limited applied value—although face matching software may yet turn out to be useful for identifying the lowest similarity faces in a match description pool. Instead, our results reinforce the importance of current lineup constructions recommendations to probe witnesses for detailed perpetrator descriptions and then match fillers to these descriptions (Wells et al. 2020) and caution against additional increases in suspect-filler similarity.

References

Bergold, A. N., & Heaton, P. (2018). Does filler database size influence identification accuracy? *Law and Human Behavior, 42*(3), 227–243. https://doi.org/10.1037/lhb0000289

Brewer, N., Lucas, C. A., Sauer, J. D., & Palmer, M. (2021). Measuring the relationship between eyewitness identification confidence and accuracy. In A. Smith, M. Toglia, and J. M. Lampinen (Eds.), *Methods, Measures, and Theories in Eyewitness Identification Tasks.* Taylor and Francis Group.

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*(1), 11–30. https://doi.org/10.1037/1076-898X.12.1.11

Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. *Proceedings of the National Academy of Sciences*, *118*(8). https://doi.org/10.1073/pnas.2017292118

Demarchi, S., & Py, J. (2009). A method to enhance person description: A field study. In R. Bull, T. Valentine, & T. Williamson (Eds.), *Handbook of psychology of investigative interviewing: Current developments and future directions* (pp. 241–256). Wiley. https://doi.org/10.1002/9780470747599.ch14

Fitzgerald, R. J., Oriet, C., & Price, H. L. (2015). Suspect-filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law and Human Behavior, 39*(1), 62–74. https://doi.org/10.1037/lhb0000095

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19*(2), 151–164. https://doi.org/10.1037/a0030618

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

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

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*(1), 3–10. https://doi.org/10.1177/0963721413498891

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

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*(4), 346–360. https://doi.org/10.1037/a0029779

Juncu, S., & Fitzgerald, R. J. (2021). A meta-analysis of lineup size effects on eyewitness identification. *Advance online publication.* https://doi.org/10.1037/law0000311

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1304–1316. https://doi.org/10.1037/0278-7393.22.5.1304

Lee, J., & Penrod, S. D. (2019). New signal detection theory-based framework for eyewitness performance in lineups. *Law and Human Behavior, 43*(5), 436–454. https://doi.org/10.1037/lhb0000343

Lucas, C. A., Brewer, N., & Palmer, M. A. (2020). Eyewitness identification: The complex issue of suspect-filler similarity. *Psychology, Public Policy, and Law*. Advance online publication. https://doi.org/10.1037/law0000243

Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15*(1), 43–57. http://dx.doi.org/10.1007/BF01044829

Mair, P., & Wilcox, R. (2019). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

Mickes, L. (2015). Receiver operating analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*(2), 93–102. https://doi.org/10.1016/j.jarmac.2015.01.003

Oriet, C., & Fitzgerald, R. J. (2018). The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law and Human Behavior*, *42*(1), 1–12. https://doi.org/10.1037/lhb0000272

R Development Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77–84. https://doi.org/10.1186/1471-2105-12-77

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision. *Psychology, Public Policy, and Law, 25*, 147–165. https://doi.org/10.1037/law0000203

Smith, A. M., Smalarz, L., Ditchfield, R. Ayala, N. T. (in press). Evaluating the claim that high confidence implies high accuracy in eyewitness identification. *Psychology, Public Policy, and Law.*

Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver operating characteristic curves of lineup identification performance. *Perspectives on Psychological Science, 15*(3), 589–607. https://doi.org/10.1177/1745691620902426

Starns, J. J., Cohen, A. L., & Rotello, C. M. (2021). A complete method for assessing the effectiveness of eyewitness identification procedures: Expected information gain. https://doi.org/10.31234/osf.io/syhzn

Steblay, N. K., & Wells, G. L. (2020). Assessment of bias in police lineups. *Psychology, Public Policy, and Law, 26*(4), 393–412. https://doi.org/10.1037/law0000287

Tredoux, C. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22*(2), 217–237. https://doi.org/10.1023/A:1025746220886

Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: Matching suspects or descriptions? *Law and Human Behavior, 24*(2), 231–258. https://doi.org/10.1023/A:1005463020252

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. https://doi.org/10.1037/1076-898X.10.3.156

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*(1), 3–36. http://dx.doi.org/10.1037/lhb0000359

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). On the selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*(5)*,* 835–844. https://doi.org/10.1037/0021-9010.78.5.835

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603–647. https://doi.org/10.1023/A:1025750605807

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior, 39*(2)*,* 99–122. https://doi.org/10.1037/lhb0000125

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

Footnotes

[1] Although the innocent suspect did not draw many picks, increasing the similarity of the fillers to this innocent suspect nevertheless increased filler identifications (the largest increase in filler identification across all stimulus sets). Since morphing the fillers to be more similar to the innocent suspect caused them to be more likely to be mistakenly identified as the target, it stands to reason that the innocent suspect did share similar features with the target. However, why the innocent suspect herself was not confused with the target remains unclear.

Table 1

Betaface Suspect-Filler Similarity Indices for Each Filler Across Stimulus Sets, Similarity Conditions and Target Presence

Filler	Suspect-Filler Similarity		
	Low (Unmorphed)	Medium (33% Morph)	High (50% Morph)
		Stimulus Set 1	
Target-Present			
1	63.0	71.2	77.2
2	57.8	61.0	68.9
3	65.8	72.6	82.8
4	56.6	66.2	74.1
5	63.8	71.6	76.2
Target-Absent			
1	63.8 (63.0)	73.9 (65.3)	83.8 (65.7)
2	59.2 (57.8)	64.8 (58.2)	72.2 (60.3)
3	62.5 (65.8)	72.7 (67.7)	80.6 (67.2)
4	55.1 (56.6)	66.0 (60.1)	74.9 (63.0)
5	58.7 (63.8)	66.2 (65.2)	74.9 (67.1)
		Stimulus Set 2	
Target-Present			
1	65.6	74.4	79.6
2	62.5	70.8	83.7
3	62.5	72.2	78.6
4	55.0	62.3	71.7
5	64.7	76.0	82.2
Target-Absent			
1	57.8 (65.6)	62.5 (67.0)	68.5 (68.6)
2	67.9 (62.5)	74.8 (65.2)	79.6 (67.7)
3	66.6 (62.5)	72.5 (65.2)	78.2 (66.6)
4	62.5 (55.0)	68.0 (58.3)	73.6 (60.6)
5	67.4 (64.7)	73.5 (67.0)	77.3 (68.5)
		Stimulus Set 3	
Target-Present			
1	61.2	69.9	75.3
2	59.2	70.2	77.1
3	61.9	70.4	78.7
4	64.7	74.1	80.0
5	59.7	69.9	80.6
Target-Absent			

1	58.9 (61.2)	64.6 (64.8)	70.5 (65.8)
2	62.3 (59.2)	70.8 (62.0)	78.1 (64.4)
3	61.9 (61.9)	70.5 (64.4)	73.9 (67.1)
4	60.6 (64.7)	67.1 (68.5)	73.0 (70.2)
5	65.6 (59.7)	73.9 (62.5)	76.9 (64.7)
		Stimulus Set 4	
Target-Present			
1	60.0	70.0	81.0
2	62.5	70.3	78.8
3	60.6	66.2	73.9
4	59.7	68.1	76.3
5	60.6	65.7	72.6
Target-Absent			
1	62.4 (60.0)	69.9 (60.6)	79.6 (60.3)
2	63.7 (62.5)	68.4 (62.1)	75.0 (61.5)
3	60.9 (60.6)	67.6 (61.2)	77.2 (60.0)
4	65.1 (59.7)	75.4 (60.5)	81.1 (59.7)
5	62.5 (60.6)	68.5 (59.7)	77.3 (60.1)

Note: Values in parentheses index similarity between the filler and the target in target-absent cases

Table 2

All Trials. Identification Decision Patterns and Confidence Across Filler Similarity, Target Presence and Lineup Size

		Suspect Pick		Filler Pick		Choosing		Confidence	
Suspect-Filler Similarity	*Trials*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*M*	*SD*
					2-Person Lineups				
Target-Present Cases									
Low	807	.70	[.66, .73]	.04	[.03, .05]	.73	[.70, .76]	78.24	(18.94)
Medium	792	.68	[.65, .71]	.10	[.08, .12]	.78	[.75, .81]	75.20	(20.42)
High	789	.61	[.57, .64]	.14	[.12, .17]	.75	[.72, .78]	73.60	(20.54)
Target-Absent Cases									
Low	812	.35	[.31, .38]	.10	[.08, .12]	.44	[.41, .48]	77.62	(19.80)
Medium	786	.33	[.29, .36]	.17	[.14, .19]	.49	[.46, .53]	77.29	(19.81)
High	778	.30	[.27, .34]	.21	[.18, .24]	.52	[.48, .55]	74.67	(19.78)
					3-Person Lineups				
Target-Present Cases									
Low	779	.69	[.66, .73]	.06	[.05, .08]	.76	[.73, .79]	75.55	(20.47)
Medium	790	.63	[.60, .67]	.14	[.12, .16]	.77	[.74, .80]	75.04	(21.18)
High	787	.54	[.50, .57]	.24	[.21, .27]	.78	[.75, .80]	73.75	(20.05)
Target-Absent Cases									
Low	804	.35	[.31, .38]	.11	[.09, .14]	.46	[.43, .50]	77.18	(19.53)
Medium	785	.29	[.26, .32]	.22	[.19, .25]	.51	[.47, .54]	73.75	(21.01)
High	811	.25	[.22, .28]	.30	[.27, .34]	.55	[.52, .59]	72.33	(20.84)
					6-Person Lineups				
Target-Present Cases									
Low	796	.59	[.55, .62]	.16	[.14, .19]	.75	[.72, .78]	73.23	(21.13)
Medium	812	.52	[.49, .56]	.26	[.23, .29]	.79	[.76, .82]	72.34	(20.85)

High	811	.44	[.41, .48]	.36	[.33, .39]	.81	[.78, .83]	69.73	(22.69)
Target-Absent Cases									
Low	810	.26	[.23, .29]	.22	[.19, .25]	.48	[.44, .51]	73.88	(20.82)
Medium	816	.21	[.18, .24]	.37	[.34, .40]	.58	[.55, .61]	69.91	(22.13)
High	819	.18	[.15, .21]	.46	[.42, .49]	.64	[.61, .67]	68.32	(22.57)

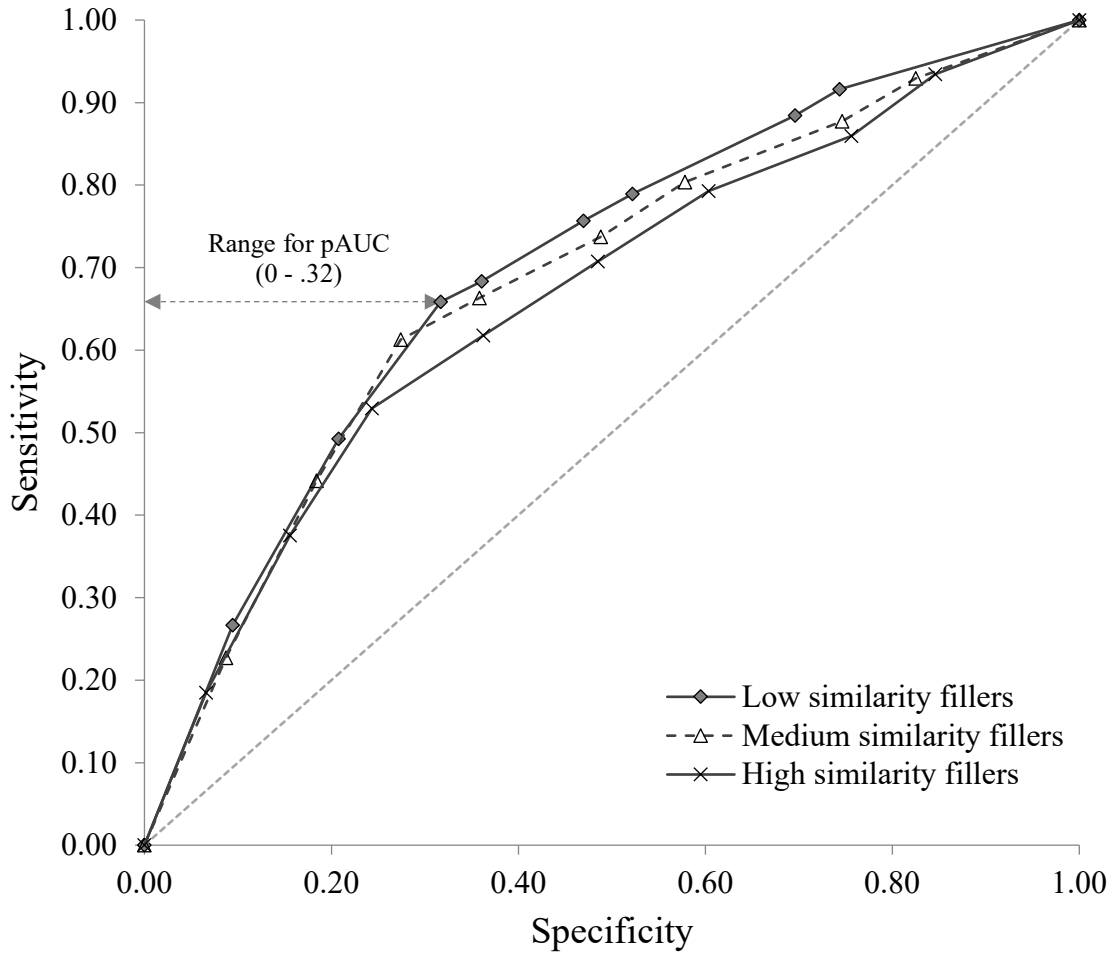

Figure 1. Full ROC curves in the low, medium and high similarity filler conditions

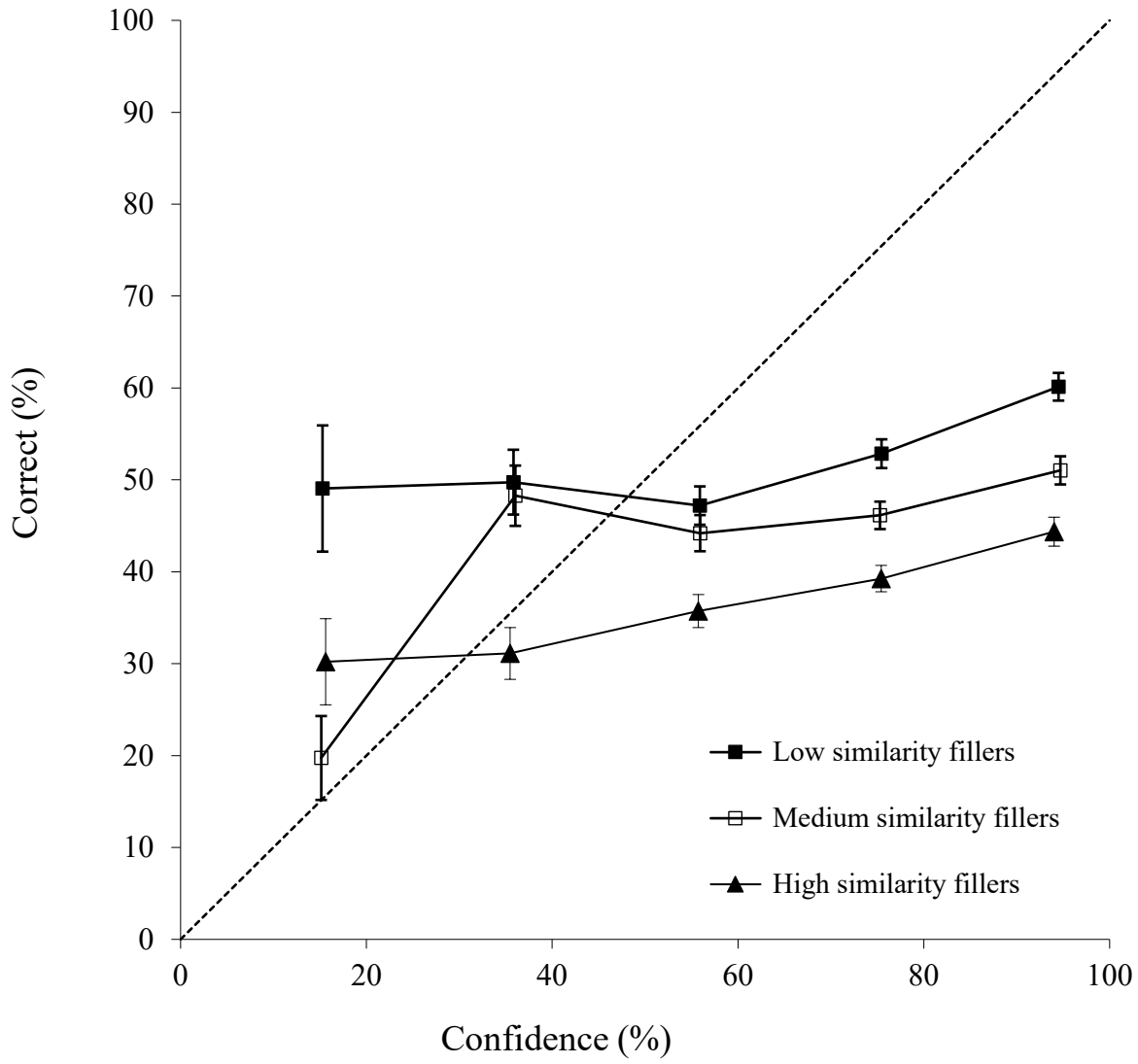

Figure 2. Calibration curves for all choosers in the low, medium and high similarity filler conditions

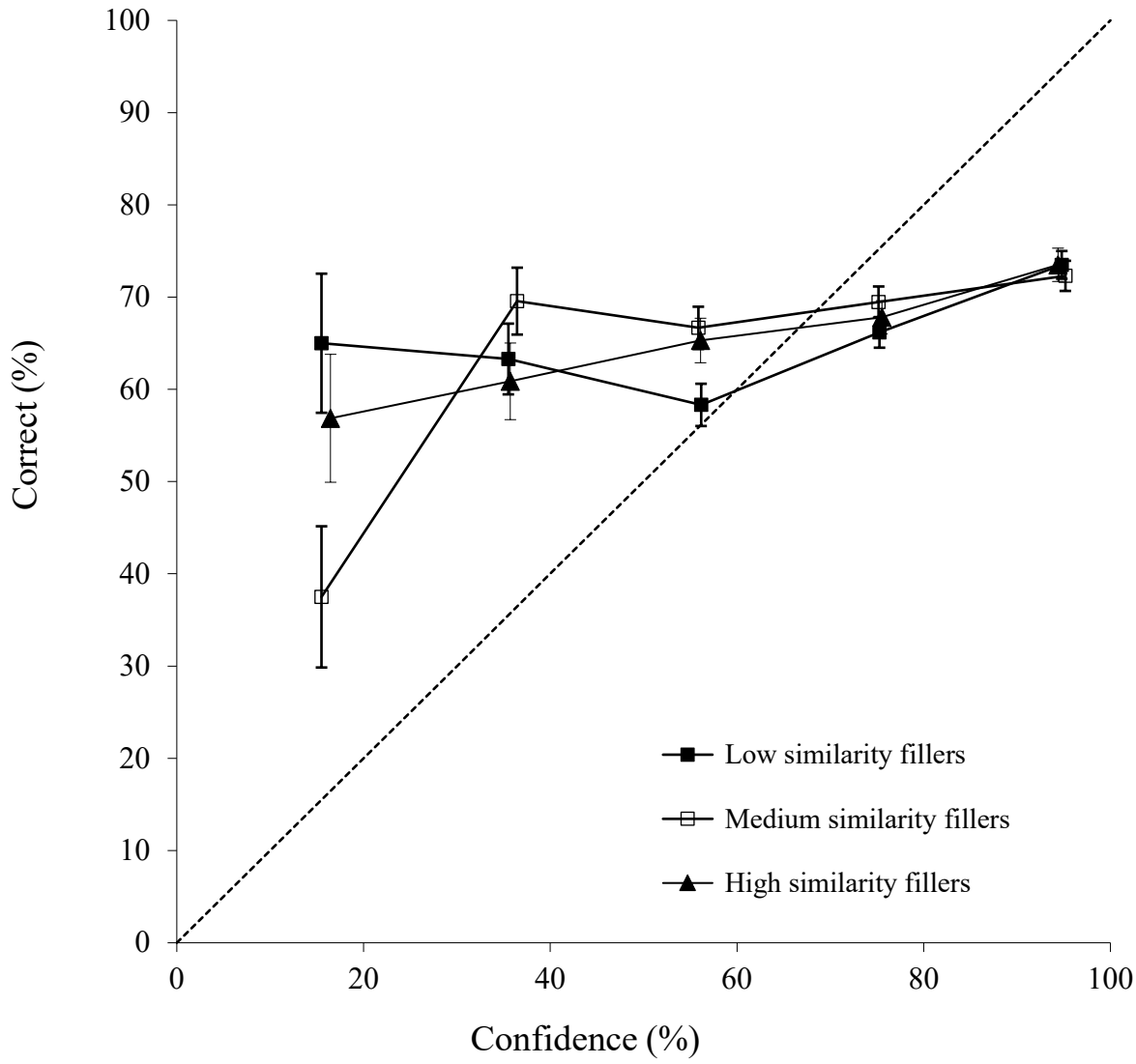

Figure 3. CAC curves in the low, medium and high similarity filler conditions

Supplemental Material

Target Descriptions

Stimulus 1

Male, Caucasian, 20s – 30s, dark hair, short beard

Stimulus 2

Female, Caucasian, Early 20s, slim build, light-brown hair

Stimulus 3

Female, Caucasian, early 20s to early 30s, dark hair, dark eyes

Stimulus 4

Male, Caucasian, mid-20s to early 30s, short brown hair

Match to Description and Suspect-Filler Similarity

Table S1

Stimulus 1. Match-Description and Similarity Indices for the Target, Innocent Suspect (IS), Backup IS and Filler Candidates

	Similarity		
Face	Target	IS (Backup IS)	Match-Description
Target	-	67.5 (63.9)	81.2
IS	67.5	- (70.3)	80.5
Backup IS	63.9	70.3 (-)	60.4
Filler Candidate 1	69.4	62.8 (61.6)	51.3
Filler Candidate 2	68.3	58.9 (62.1)	79.2
Filler Candidate 3	66.2	67.4 (66.3)	85.7
Filler Candidate 4	65.1	63.4 (61.4)	86.4
Filler Candidate 5	70.2	62.5 (62.0)	61.0
Filler Candidate 6	61.1	63.4 (66.1)	81.2
Filler Candidate 7	60.1	67.0 (65.3)	63.6
Filler Candidate 8*	61.4	62.4 (62.5)	61.0
Filler Candidate 9	64.3	63.4 (61.9)	84.4
Filler Candidate 10*	58.7	59.1 (61.6)	72.1
Filler Candidate 11	73.5	66.9 (64.0)	66.2
Filler Candidate 12	63.9	62.4 (67.7)	43.5
Filler Candidate 13	66.3	62.6 (62.4)	79.9
Filler Candidate 14	62.5	70.7 (70.2)	72.7
Filler Candidate 15	65.1	62.5 (62.9)	84.4
Filler Candidate 16*	63.8	62.9 (62.9)	61.0
Filler Candidate 17	64.8	65.6 (66.7)	64.9
Filler Candidate 18*	58.2	56.4 (64.6)	76.6
Filler Candidate 19*	61.1	59.2 (61.7)	86.4
Filler Candidate 20	64.4	63.8 (61.7)	84.4
Filler Candidate 21	66.5	61.0 (61.7)	77.3
Filler Candidate 22	80.0	67.2 (62.0)	51.9
Filler Candidate 23	60.2	57.8 (65.2)	39.6

*filler used in study

Table S2

Stimulus 2. Match-Description and Similarity Indices for the Target, Innocent Suspect (IS), Backup IS and Filler Candidates

	Similarity		
Face	Target	IS (Backup IS)	Match-Description
Target	-	69.0 (63.0)	82.5
IS	69.0	- (68.0)	68.2
Backup IS	63.0	68.0 (-)	77.3
Filler Candidate 1	62.9	68.5 (59.8)	76.0
Filler Candidate 2	59.2	73.2 (70.5)	35.1
Filler Candidate 3**	63.8	67.6 (61.1)	40.9
Filler Candidate 4	61.1	66.1 (57.4)	27.9
Filler Candidate 5	69.9	73.4 (66.9)	42.2
Filler Candidate 6	67.7	67.0 (59.3)	41.6
Filler Candidate 7*	64.2	57.5 (62.8)	50.0
Filler Candidate 8*(r)	62.0	66.7 (60.7)	68.8
Filler Candidate 9	66.7	64.4 (63.8)	61.0
Filler Candidate 10	68.0	68.0 (60.6)	37.0
Filler Candidate 11	59.2	68.5 (66.2)	75.3
Filler Candidate 12*	63.8	67.1 (64.4)	76.6
Filler Candidate 13*	56.9	62.1 (61.7)	71.4
Filler Candidate 14	64.2	74.0 (76.4)	35.7
Filler Candidate 15	70.9	65.1 (59.7)	53.2
Filler Candidate 16	67.5	67.9 (70.7)	31.2
Filler Candidate 17	68.1	71.7 (61.5)	63.0
Filler Candidate 18	75.9	62.9 (62.8)	55.8
Filler Candidate 19	66.6	72.9 (66.0)	60.4
Filler Candidate 20	62.4	68.5 (62.0)	40.9
Filler Candidate 21*	63.2	67.6 (68.1)	53.9

*used in study; (r) replaced; **replacement

Table S3

Stimulus 3. Match-Description and Similarity Indices for the Target, Innocent Suspect (IS), Backup IS and Filler Candidates

	Similarity		
Face	Target	IS (Backup IS)	Match-Description
Target	-	65.3 (76.8)	83.8
IS	65.3	- (65.6)	82.5
Backup IS	76.8	65.6 (-)	87.0
Filler Candidate 1	68.5	73.1 (68.0)	80.5
Filler Candidate 2	66.1	65.4 (61.9)	43.5
Filler Candidate 3**	59.7	65.7 (62.1)	67.5
Filler Candidate 4	68.3	65.3 (68.5)	80.5
Filler Candidate 5	66.1	63.0 (66.0)	72.7
Filler Candidate 6	65.6	63.2 (62.0)	74.0
Filler Candidate 7	69.9	59.5 (66.1)	72.7
Filler Candidate 8	61.9	60.3 (64.6)	44.8
Filler Candidate 9**	62.3	61.4 (61.5)	49.4
Filler Candidate 10	67.2	64.4 (65.8)	67.5
Filler Candidate 11*	61.5	62.9 (59.1)	79.9
Filler Candidate 12	64.7	66.5 (69.1)	61.7
Filler Candidate 13	66.2	78.6 (67.1)	65.6
Filler Candidate 14	61.4	63.8 (65.2)	63.6
Filler Candidate 15	65.6	61.1 (63.2)	65.6
Filler Candidate 16*(r)	64.8	60.3 (70.2)	69.5
Filler Candidate 17	65.6	60.0 (67.4)	77.9
Filler Candidate 18*(r)	59.1	62.5 (66.6)	71.4
Filler Candidate 19	71.6	73.5 (72.6)	74.7
Filler Candidate 20*	64.3	57.9 (60.6)	66.9
Filler Candidate 21*	61.0	57.8 (65.2)	78.6
Filler Candidate 22	61.1	68.3 (58.3)	35.1
Filler Candidate 23	67.4	64.4 (71.7)	65.6

*used in study; (r) replaced; **replacement

Table S4

Stimulus 4. Match-Description and Similarity Indices for the Target, Innocent Suspect (IS), Backup IS and Filler Candidates

	Similarity		
Face	Target	IS (Backup IS)	Match-Description
Target	-	60.87 (69.9)	66.9
IS	60.7	- (59.8)	62.3
Backup IS	69.9	59.8 (-)	67.5
Filler Candidate 1*	59.2	61.5 (61.5)	82.5
Filler Candidate 2	65.6	67.0 (59.7)	62.3
Filler Candidate 3	65.3	61.1 (69.4)	87.7
Filler Candidate 4	63.4	63.4 (59.3)	35.7
Filler Candidate 5*	62.8	62.9 (63.4)	63.6
Filler Candidate 6*	59.1	60.6 (58.3)	81.8
Filler Candidate 7	63.4	65.4 (63.4)	67.5
Filler Candidate 8	57.8	67.2 (61.6)	70.8
Filler Candidate 9	64.3	65.7 (62.5)	46.1
Filler Candidate 10*	61.5	63.3 (59.1)	68.2
Filler Candidate 11	61.5	67.2 (66.6)	70.1
Filler Candidate 12	69.3	64.6 (70.7)	54.5
Filler Candidate 13	64.6	67.0 (67.4)	33.8
Filler Candidate 14	63.0	59.6 (66.9)	59.1
Filler Candidate 15	67.7	68.8 (64.9)	72.1
Filler Candidate 16*	59.5	63.4 (62.6)	89.0
Filler Candidate 17	57.5	63.4 (63.7)	83.1
Filler Candidate 18	62.3	67.1 (63.2)	72.1
Filler Candidate 19	66.1	61.2 (62.3)	82.5
Filler Candidate 20	58.9	67.0 (61.4)	64.9
Filler Candidate 21	62.8	70.7 (62.8)	72.7

*used in study

Lineup Fairness

Table S5

Lineup Fairness Pilot, Stimulus 1. Frequency (and Proportion) of Lineup Member Selections, and Lineup Functional Size, at Each Level of Target Presence and Morph to the Suspect

		Lineup Member						
Morph to Suspect	*N*	1	2	3	4	5	6	Functional Size
Target-Present								
0%	82	10 (.12)	13 (.16)	10 (.12)	15 (.18)	11 (.13)	23 (.28)	5.41 [4.71, 6.34]
33%	79	6 (.08)	6 (.08)	4 (.05)	21 (.27)	22 (.28)	20 (.25)	4.42 [3.82, 5.23]
66%	84	7 (.08)	6 (.07)	7 (.08)	18 (.21)	10 (.12)	36 (.43)	3.81 [3.06, 5.04]
Target-Absent								
0%	83	10 (.12)	7 (.08)	7 (.08)	15 (.18)	15 (.18)	29 (.35)	4.63 [3.84, 5.83]
33%	84	9 (.11)	16 (.19)	7 (.08)	16 (.19)	15 (.18)	21 (.25)	5.39 [4.83, 6.11]
66%	83	15 (.18)	10 (.12)	15 (.18)	15 (.18)	13 (.16)	15 (.18)	5.89 [5.61, 6.21]

*denotes selection of the suspect (i.e., lineup member 1) at a rate greater than chance (p <.05)

Table S6

Lineup Fairness Pilot, Stimulus 2. Frequency (and Proportion) of Lineup Member Selections, and Lineup Functional Size, at Each Level of Target Presence and Morph to the Suspect

		Lineup Member						
Morph to Suspect	*N*	1	2	3	4	5	6	Functional Size
Target-Present								
0%	84	30*	9	5	22	16	2	4.03
		(.36)	(.11)	(.06)	(.26)	(.19)	(.02)	[3.43, 4.89]
33%	82	7	25	9	22	15	4	4.54
		(.09)	(.30)	(.11)	(.27)	(.18)	(.05)	[3.89, 5.47]
66%	82	6	27	8	24	12	5	4.27
		(.07)	(.33)	(.24)	(.29)	(.15)	(.06)	[3.59, 5.28]
Target-Absent								
0%	83	11	15	10	19	24	4	4.92
		(.13)	(.18)	(.12)	(.23)	(.29)	(.05)	[4.28, 5.79]
33%	83	7	16	14	20	24	2	4.65
		(.08)	(.19)	(.17)	(.24)	(.29)	(.02)	[4.10, 5.38]
66%	81	4	18	17	15	18	9	5.21
		(.05)	(.22)	(.21)	(.19)	(.22)	(.11)	[4.71, 5.83]

*denotes selection of the suspect (i.e., lineup member 1) at a rate greater than chance (p <.05)

Table S7

Lineup Fairness Pilot, Stimulus 3. Frequency (and Proportion) of Lineup Member Selections, and Lineup Functional Size, at Each Level of Target Presence and Morph to the Suspect

		Lineup Members						
Morph to Suspect	N	1	2	3	4	5	6	Functional Size
Target-Present								
0%	83	11 (.13)	5 (.06)	11 (.13)	25 (.30)	10 (.12)	21 (.25)	4.81 [4.10, 5.80]
33%	84	11 (.13)	4 (.05)	13 (.15)	19 (.23)	16 (.19)	21 (.25)	5.17 [4.63, 5.86]
66%	82	13 (.16)	10 (.12)	12 (.15)	23 (.28)	12 (.15)	12 (.15)	5.47 [4.77, 6.40]
Target-Absent								
0%	83	22* (.27)	6 (.07)	10 (.12)	19 (.23)	3 (.04)	23 (.28)	4.54 [3.96, 5.30]
33%	83	13 (.16)	3 (.04)	14 (.17)	30 (.36)	4 (.05)	19 (.23)	4.17 [3.52, 5.12]
66%	80	20 (.25)	9 (.11)	8 (.10)	22 (.28)	11 (.14)	10 (.13)	5.12 [4.40, 6.11]

*denotes selection of the suspect (i.e., lineup member 1) at a rate greater than chance (p <.05)

Table S8

Lineup Fairness Pilot, Stimulus 4. Frequency and Proportion of Lineup Member Selections, and Lineup Functional Size, at Each Level of Target Presence and Morph to the Suspect

		Lineup Member						
Morph to Suspect	N	1	2	3	4	5	6	Functional Size
Target-Present								
0%	84	16	21	23	5	0	19	4.38
		(.19)	(.25)	(.27)	(.06)	(.00)	(.23)	[3.99, 4.84]
33%	82	5	20	27	6	2	22	4.01
		(.06)	(.24)	(.33)	(.07)	(.02)	(.27)	[3.47, 4.74]
66%	81	6	21	17	11	9	17	5.22
		(.07)	(.26)	(.21)	(.14)	(.11)	(.21)	[4.60, 6.03]
Target-Absent								
0%	83	10	14	33	7	4	15	4.11
		(.12)	(.17)	(.40)	(.08)	(.05)	(.18)	[3.34, 5.34]
33%	82	11	10	27	14	7	13	4.93
		(.13)	(.12)	(.33)	(.17)	(.09)	(.16)	[4.12, 6.12]
66%	83	14	24	12	9	10	14	5.33
		(.17)	(.29)	(.14)	(.11)	(.12)	(.17)	[4.61, 6.31]

*denotes selection of the suspect (i.e., lineup member 1) at a rate greater than chance (p <.05)

Identification Patterns and Confidence

Table S9

All Trials. Identification Decision Patterns and Confidence Across Filler Similarity, Target Presence and Lineup Size

		Suspect Pick		Filler Pick		Choosing		Confidence	
Suspect-Filler Similarity	*Trials*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*M*	*SD*
					2-Person Lineups				
Target-Present Cases									
Low	807	.70	[.66, .73]	.04	[.03, .05]	.73	[.70, .76]	78.24	(18.94)
Medium	792	.68	[.65, .71]	.10	[.08, .12]	.78	[.75, .81]	75.20	(20.42)
High	789	.61	[.57, .64]	.14	[.12, .17]	.75	[.72, .78]	73.60	(20.54)
Overall	2,388	.66	[.64, .68]	.09	[.08, .11]	.76	[.74, .77]	75.70	(20.06)
Target-Absent Cases									
Low	812	.35	[.31, .38]	.10	[.08, .12]	.44	[.41, .48]	77.62	(19.80)
Medium	786	.33	[.29, .36]	.17	[.14, .19]	.49	[.46, .53]	77.29	(19.81)
High	778	.30	[.27, .34]	.21	[.18, .24]	.52	[.48, .55]	74.67	(19.78)
Overall	2,376	.33	[.31, .34]	.16	[.14, .17]	.48	[.46, .50]	76.54	(19.83)
All Cases									
Low	1,619	.52	[.50, .54]	.07	[.06, .08]	.59	[.56, .61]	77.93	(19.37)
Medium	1,578	.51	[.48, .53]	.13	[.12, .15]	.64	[.61, .66]	76.24	(20.14)
High	1,567	.46	[.43, .48]	.18	[.16, .20]	.63	[.61, .66]	74.13	(20.17)
Overall	4,764	.49	[.48, .51]	.13	[.12, .14]	.62	[.61, .63]	76.12	(19.95)
					3-Person Lineups				
Target-Present Cases									
Low	779	.69	[.66, .73]	.06	[.05, .08]	.76	[.73, .79]	75.55	(20.47)
Medium	790	.63	[.60, .67]	.14	[.12, .16]	.77	[.74, .80]	75.04	(21.18)
High	787	.54	[.50, .57]	.24	[.21, .27]	.78	[.75, .80]	73.75	(20.05)
Overall	2,356	.62	[.60, .64]	.15	[.13, .16]	.77	[.75, .79]	74.78	(20.58)
Target-Absent Cases									
Low	804	.35	[.31, .38]	.11	[.09, .14]	.46	[.43, .50]	77.18	(19.53)

Medium	785	.29	[.26, .32]	.22	[.19, .25]	.51	[.47, .54]	73.75	(21.01)
High	811	.25	[.22, .28]	.30	[.27, .34]	.55	[.52, .59]	72.33	(20.84)
Overall	2,400	.29	[.28, .31]	.21	[.20, .23]	.51	[.49, .53]	74.42	(20.56)
All Cases									
Low	1,583	.52	[.49, .54]	.09	[.08, .10]	.61	[.58, .63]	76.37	(20.01)
Medium	1,575	.46	[.44, .49]	.18	[.16, .20]	.64	[.62, .66]	74.41	(21.10)
High	1,598	.39	[.37, .41]	.27	[.25, .29]	.66	[.64, .69]	73.03	(20.46)
Overall	4,756	.46	[.44, .47]	.18	[.17, .19]	.64	[.62, .65]	74.60	(20.57)
					6-Person Lineups				
Target-Present Cases									
Low	796	.59	[.55, .62]	.16	[.14, .19]	.75	[.72, .78]	73.23	(21.13)
Medium	812	.52	[.49, .56]	.26	[.23, .29]	.79	[.76, .82]	72.34	(20.85)
High	811	.44	[.41, .48]	.36	[.33, .39]	.81	[.78, .83]	69.73	(22.69)
Overall	2,419	.52	[.50, .54]	.26	[.25, .28]	.78	[.77, .80]	71.76	(21.62)
Target-Absent Cases									
Low	810	.26	[.23, .29]	.22	[.19, .25]	.48	[.44, .51]	73.88	(20.82)
Medium	816	.21	[.18, .24]	.37	[.34, .40]	.58	[.55, .61]	69.91	(22.13)
High	819	.18	[.15, .21]	.46	[.42, .49]	.64	[.61, .67]	68.32	(22.57)
Overall	2,445	.22	[.20, .23]	.35	[.33, .37]	.57	[.55, .59]	70.69	(21.97)
All Cases									
Low	1,606	.42	[.40, .44]	.19	[.17, .21]	.61	[.59, .64]	73.56	(20.97)
Medium	1,628	.37	[.34, .39]	.32	[.33, .41]	.68	[.66, .71]	71.12	(21.53)
High	1,630	.31	[.29, .33]	.41	[.39, .43]	.72	[.70, .74]	69.02	(22.64)
Overall	4,864	.37	[.35, .38]	.31	[.29, .32]	.67	[.66, .69]	71.22	(21.80)
					All Lineup Sizes				
Target-Present Cases									
Low	2,382	.66	[.64, .68]	.09	[.08, .10]	.75	[.73, .77]	75.68	(20.29)
Medium	2,394	.61	[.59, .63]	.17	[.15, .18]	.78	[.77, .80]	74.19	(20.85)
High	2,387	.53	[.51, .55]	.25	[.23, .27]	.78	[.76, .79]	72.33	(21.21)
Overall	7,163	.60	[.59, .61]	.17	[.16, .18]	.77	[.76, .78]	74.07	(20.83)
Target-Absent Cases									
Low	2,426	.32	[.30, .34]	.14	[.13, .16]	.46	[.44, .48]	76.22	(20.12)

Medium	2,387	.27	[.26, .29]	.25	[.24, .27]	.53	[.51, .55]	73.60	(21.23)
High	2,408	.24	[.23, .26]	.33	[.31, .35]	.57	[.55, .59]	71.72	(21.27)
Overall	7,221	.28	[.27, .29]	.24	[.23, .25]	.52	[.51, .53]	73.86	(20.96)
All Cases									
Low	4,808	.49	[.47, .50]	.12	[.11, .13]	.60	[.59, .62]	75.96	(20.20)
Medium	4,781	.44	[.43, .46]	.21	[.20, .22]	.65	[.64, .67]	73.89	(21.04)
High	4,795	.39	[.37, .40]	.29	[.28, .30]	.67	[.66, .69]	72.03	(21.24)
Overall	14,384	.44	[.43, .45]	.21	[.20, .21]	.64	[.64, .65]	73.96	(20.89)

Table S10

First Trial. Identification Decision Patterns and Confidence Across Filler Similarity, Target Presence and Lineup Size

		Suspect Pick		Filler Pick		Choosing		Confidence	
Suspect-Filler Similarity	*Trials*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*M*	*SD*
					2-Person Lineups				
Target-Present Cases									
Low	193	.68	[.62, .75]	.04	[.01, .06]	.72	[.65, .78]	76.37	(18.82)
Medium	220	.65	[.58, .71]	.11	[.07, .15]	.76	[.70, .81]	75.14	(18.81)
High	196	.63	[.56, .70]	.11	[.06, .15]	.74	[.68, .80]	72.65	(19.43)
Overall	609	.65	[.61, .69]	.09	[.06, .11]	.74	[.70, .77]	74.73	(19.05)
Target-Absent Cases									
Low	179	.31	[.24, .38]	.09	[.04, .13]	.40	[.33, .47]	76.93	(19.34)
Medium	193	.32	[.25, .38]	.16	[.10, .20]	.48	[.40, .54]	77.62	(17.72)
High	210	.26	[.20, .32]	.22	[.17, .28]	.49	[.42, .55]	75.10	(20.41)
Overall	582	.30	[.26, .33]	.16	[.13, .19]	.46	[.42, .50]	76.49	(19.22)
All Cases									
Low	372	.51	[.45, .55]	.06	[.04, .08]	.57	[.52, .62]	76.64	(19.05)
Medium	413	.49	[.44, .54]	.13	[.10, .16]	.63	[.58, .67]	76.30	(18.33)
High	406	.44	[.39, .49]	.17	[.13, .20]	.61	[.56, .65]	73.92	(19.96)
Overall	1191	.48	[.45, .51]	.12	[.10, .14]	.60	[.57, .63]	75.59	(19.14)
					3-Person Lineups				
Target-Present Cases									
Low	168	.68	[.61, .75]	.10	[.05, .14]	.78	[.71, .84]	73.15	(20.03)
Medium	199	.64	[.57, .70]	.15	[.09, .19]	.78	[.72, .84]	75.48	(18.90)
High	192	.50	[.43, .57]	.26	[.20, .32]	.76	[.70, .82]	74.84	(18.90)
Overall	559	.60	[.56, .64]	.17	[.14, .20]	.77	[.74, .81]	74.56	(19.24)
Target-Absent Cases									
Low	190	.33	[.26, .40]	.13	[.08, .17]	.46	[.38, .53]	76.47	(18.51)
Medium	223	.25	[.19, .31]	.20	[.15, .25]	.45	[.39, .52]	75.15	(18.15)
High	217	.24	[.18, .29]	.34	[.28, .40]	.58	[.51, .64]	72.77	(18.58)

Overall	630	.27	[.24, .31]	.23	[.19, .26]	.50	[.46, .54]	74.73	(18.44)
All Cases									
Low	358	.50	[.44, .55]	.11	[.08, .14]	.61	[.56, .66]	74.92	(19.28)
Medium	422	.43	[.39, .48]	.18	[.14, .21]	.61	[.56, .65]	75.31	(18.49)
High	409	.36	[.31, .41]	.30	[.26, .35]	.67	[.62, .71]	73.74	(18.73)
Overall	1189	.43	[.40, .46]	.20	[.18, .22]	.63	[.60, .66]	74.65	(18.81)
					6-Person Lineups				
Target-Present Cases									
Low	203	.54	[.47, .60]	.19	[.13, .24]	.72	[.66, .78]	72.17	(20.54)
Medium	188	.49	[.42, .56]	.28	[.21, .34]	.78	[.71, .83]	70.00	(20.42)
High	212	.44	[.37, .50]	.36	[.29, .42]	.80	[.74, .85]	69.39	(22.52)
Overall	603	.49	[.45, .53]	.28	[.24, .31]	.77	[.73, .80]	70.51	(21.22)
Target-Absent Cases									
Low	218	.27	[.21, .32]	.22	[.16, .27]	.49	[.42, .55]	73.72	(18.87)
Medium	177	.19	[.13, .25]	.38	[.30, .45]	.57	[.49, .64]	70.51	(21.11)
High	218	.17	[.12, .22]	.48	[.41, .55]	.66	[.59, .72]	69.13	(20.63)
Overall	613	.21	[.18, .24]	.36	[.32, .40]	.57	[.53, .61]	71.16	(20.23)
All Cases									
Low	421	.40	[.35, .44]	.20	[.16, .24]	.60	[.55, .65]	72.97	(19.69)
Medium	365	.35	[.30, .40]	.33	[.28, .37]	.68	[.63, .72]	70.25	(20.73)
High	430	.30	[.26, .35]	.42	[.37, .47]	.73	[.68, .77]	69.26	(21.56)
Overall	1216	.35	[.32, .38]	.32	[.29, .34]	.67	[.64, .69]	70.84	(20.72)
					All Lineup Sizes				
Target-Present Cases									
Low	564	.63	[.59, .67]	.11	[.08, .13]	.74	[.70, .77]	73.90	(19.87)
Medium	607	.60	[.56, .63]	.17	[.14, .20]	.77	[.74, .81]	73.66	(19.48)
High	600	.52	[.48, .56]	.25	[.21, .28]	.77	[.73, .80]	72.20	(20.51)
Overall	1771	.58	[.56, .60]	.18	[.16, .19]	.76	[.74, .78]	73.24	(19.96)
Target-Absent Cases									
Low	587	.30	[.26, .34]	.15	[.12, .18]	.45	[.41, .49]	75.59	(18.92)
Medium	593	.26	[.22, .29]	.24	[.20, .27]	.50	[.45, .54]	74.57	(19.13)
High	645	.22	[.19, .26]	.35	[.31, .39]	.58	[.54, .61]	72.29	(20.01)

Overall	1825	.26	[.24, .28]	.25	[.23, .27]	.51	[.49, .53]	74.09	(19.42)
All Cases									
Low	1151	.46	[.43, .49]	.13	[.11, .15]	.59	[.56, .62]	74.76	(19.40)
Medium	1200	.43	[.40, .46]	.21	[.18, .23]	.64	[.61, .66]	74.11	(19.30)
High	1245	.37	[.34, .39]	.30	[.27, .32]	.67	[.64, .69]	72.25	(20.24)
Overall	3596	.42	[.40, .43]	.21	[.20, .23]	.63	[.62, .65]	73.67	(19.69)

Table S11

All Trials, Stimulus Set 1. Identification Decision Patterns and Confidence Across Filler Similarity, Target Presence and Lineup Size

		Suspect Pick		Filler Pick		Choosing		Confidence	
Suspect-Filler Similarity	*Trials*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*M*	*SD*
					2-Person Lineups				
Target-Present Cases									
Low	196	.86	[.81, .90]	.03	[.00, .05]	.88	[.84, .93]	78.88	(18.99)
Medium	200	.74	[.68, .80]	.10	[.05, .13]	.84	[.78, .88]	75.30	(19.95)
High	205	.63	[.57, .70]	.13	[.08, .17]	.76	[.70, .82]	73.02	(20.76)
Overall	601	.74	[.71, .78]	.08	[.06, .10]	.83	[.79, .85]	75.69	(20.04)
Target-Absent Cases									
Low	209	.58	[.51, .65]	.08	[.04, .12]	.67	[.60, .73]	72.30	(20.60)
Medium	187	.58	[.50, .65]	.11	[.06, .15]	.69	[.62, .75]	76.42	(21.31)
High	194	.45	[.38, .52]	.17	[.11, .22]	.62	[.55, .68]	73.14	(20.51)
Overall	590	.54	[.50, .58]	.12	[.09, .15]	.66	[.62, .70]	73.88	(20.84)
All Cases									
Low	405	.72	[.67, .76]	.05	[.03, .08]	.77	[.73, .81]	75.48	(20.09)
Medium	387	.66	[.61, .71]	.10	[.07, .13]	.76	[.72, .81]	75.84	(20.60)
High	399	.54	[.49, .59]	.15	[.11, .18]	.69	[.65, .74]	73.08	(20.61)
Overall	1191	.64	[.61, .67]	.10	[.08, .12]	.74	[.72, .76]	74.79	(20.45)
					3-Person Lineups				
Target-Present Cases									
Low	203	.81	[.76, .86]	.06	[.03, .10]	.88	[.83, .92]	77.64	(18.76)
Medium	197	.73	[.67, .79]	.11	[.06, .15]	.84	[.78, .89]	75.13	(20.19)
High	200	.57	[.49, .63]	.31	[.24, .37]	.87	[.82, .91]	72.05	(19.06)
Overall	600	.70	[.67, .74]	.16	[.13, .19]	.86	[.83, .89]	74.95	(19.44)
Target-Absent Cases									
Low	195	.57	[.50, .64]	.09	[.05, .13]	.67	[.60, .73]	75.38	(19.62)
Medium	197	.51	[.44, .57]	.18	[.12, .23]	.69	[.62, .75]	71.22	(22.80)
High	197	.37	[.30, .43]	.33	[.26, .39]	.70	[.63, .76]	71.68	(21.94)

Overall	589	.48	[.44, .52]	.20	[.17, .23]	.68	[.64, .72]	72.75	(21.55)
All Cases									
Low	398	.70	[.65, .74]	.08	[.05, .10]	.77	[.73, .81]	76.53	(19.19)
Medium	394	.62	[.57, .67]	.14	[.11, .18]	.76	[.72, .80]	73.17	(21.60)
High	397	.47	[.42, .51]	.32	[.27, .36]	.78	[.74, .82]	71.86	(20.51)
Overall	1189	.59	[.57, .62]	.18	[.16, .20]	.77	[.75, .80]	73.86	(20.53)
					6-Person Lineups				
Target-Present Cases									
Low	200	.73	[.66, .78]	.14	[.09, .18]	.86	[.81, .91]	74.15	(21.15)
Medium	195	.65	[.58, .71]	.24	[.18, .30]	.89	[.84, .93]	71.44	(20.76)
High	204	.52	[.45, .59]	.32	[.26, .39]	.85	[.80, .89]	71.27	(23.41)
Overall	599	.63	[.59, .67]	.23	[.20, .27]	.86	[.84, .89]	72.29	(21.83)
Target-Absent Cases									
Low	208	.48	[.41, .55]	.16	[.11, .21]	.64	[.58, .71]	72.69	(21.30)
Medium	209	.38	[.31, .44]	.31	[.24, .37]	.68	[.62, .74]	69.33	(21.54)
High	200	.31	[.24, .37]	.49	[.41, .55]	.79	[.73, .84]	65.15	(23.98)
Overall	617	.39	[.35, .43]	.32	[.28, .35]	.71	[.67, .74]	69.11	(22.46)
All Cases									
Low	408	.60	[.55, .65]	.15	[.11, .18]	.75	[.71, .79]	73.41	(21.21)
Medium	404	.51	[.46, .55]	.27	[.23, .32]	.78	[.74, .82]	70.35	(21.17)
High	404	.42	[.37, .46]	.40	[.35, .45]	.82	[.78, .86]	68.24	(23.86)
Overall	1216	.51	[.48, .54]	.28	[.25, .30]	.78	[.76, .81]	70.67	(22.20)
					All Lineup Sizes				
Target-Present Cases									
Low	599	.80	[.77, .83]	.08	[.05, .10]	.87	[.85, .90]	76.88	(19.73)
Medium	592	.71	[.67, .74]	.15	[.12, .17]	.85	[.82, .88]	73.97	(20.34)
High	609	.57	[.53, .61]	.25	[.22, .28]	.83	[.79, .86]	72.12	(21.14)
Overall	1800	.69	[.67, .71]	.16	[.14, .17]	.85	[.83, .87]	74.31	(20.50
Target-Absent Cases									
Low	612	.55	[.51, .58]	.11	[.09, .14]	.66	[.62, .70]	73.42	(20.55)
Medium	593	.48	[.44, .52]	.20	[.17, .23]	.69	[.65, .72]	72.19	(22.06)
High	591	.37	[.33, .41]	.33	[.29, .37]	.70	[.66, .74]	69.95	(22.44)

Overall	1796	.47	[.44, .49]	.21	[.19, .23]	.68	[.66, .70]	71.87	(21.72)
All Cases									
Low	1211	.67	[.64, .70]	.09	[.08, .11]	.76	[.74, .79]	75.13	(20.21)
Medium	1185	.59	[.57, .62]	.17	[.15, .20]	.77	[.75, .79]	73.08	(21.23)
High	1200	.48	[.45, .50]	.29	[.26, .32]	.77	[.74, .79]	71.05	(21.81)
Overall	3596	.58	[.56, .60]	.19	[.17, .20]	.77	[.75, .78]	73.09	(21.15)

Table S12

All Trials, Stimulus Set 2. Identification Decision Patterns and Confidence Across Filler Similarity, Target Presence and Lineup Size

		Suspect Pick		Filler Pick		Choosing		Confidence	
Suspect-Filler Similarity	*Trials*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*M*	*SD*
					2-Person Lineups				
Target-Present Cases									
Low	200	.62	[.55, .68]	.05	[.02, .08]	.67	[.60, .73]	76.20	(19.84)
Medium	195	.57	[.50, .64]	.12	[.07, .17]	.69	[.62, .75]	74.26	(21.68)
High	194	.40	[.33, .47]	.24	[.18, .30]	.64	[.57, .71]	72.06	(21.08)
Overall	589	.53	[.49, .57]	.14	[.11, .16]	.67	[.63, .70]	74.19	(20.90)
Target-Absent Cases									
Low	200	.17	[.11, .21]	.11	[.06, .14]	.27	[.21, .33]	81.05	(18.95)
Medium	209	.15	[.10, .19]	.26	[.20, .32]	.41	[.34, .47]	76.75	(18.71)
High	193	.16	[.10, .20]	.26	[.19, .32]	.41	[.34, .48]	74.97	(19.26)
Overall	602	.16	[.13, .18]	.21	[.17, .24]	.36	[.32, .40]	77.61	(19.10)
All Cases									
Low	400	.39	[.34, .44]	.08	[.05, .10]	.47	[.42, .52]	78.63	(19.52)
Medium	404	.35	[.30, .40]	.19	[.15, .23]	.54	[.49, .59]	75.54	(20.21)
High	387	.28	[.23, .32]	.25	[.21, .29]	.53	[.48, .58]	73.51	(20.22)
Overall	1191	.34	[.31, .37]	.17	[.15, .19]	.51	[.49, .54]	75.92	(20.08)
					3-Person Lineups				
Target-Present Cases									
Low	190	.62	[.54, .68]	.06	[.03, .10]	.68	[.61., 74]	73.21	(21.32)
Medium	203	.46	[.39, .52]	.20	[.14, .25]	.66	[.59, .72]	73.35	(22.17)
High	201	.38	[.31, .45]	.28	[.22, .34]	.67	[.60, .73]	73.68	(20.70)
Overall	594	.48	[.44, .52]	.18	[.15, .21]	.67	[.63, .70]	73.42	(21.38)
Target-Absent Cases									
Low	203	.18	[.13, .23]	.11	[.06, .15]	.29	[.23, .35]	80.00	(18.98)
Medium	198	.10	[.05, .13]	.27	[.20, .33]	.36	[.29, .43]	74.04	(20.89)
High	194	.08	[.04, .12]	.41	[.34, .47]	.49	[.42, .56]	71.49	(21.53)

Overall	595	.12	[.09, .15]	.26	[.22, .29]	.38	[.34, .42]	75.24	(20.75)
All Cases									
Low	393	.39	[.34, .44]	.09	[.06, .11]	.48	[.43, .53]	76.72	(20.41)
Medium	401	.28	[.23, .32]	.23	[.19, .27]	.51	[.46, .56]	73.69	(21.53)
High	395	.24	[.19, .28]	.34	[.30, .39]	.58	[.53, .63]	72.61	(21.11)
Overall	1189	.30	[.28, .33]	.22	[.20, .24]	.52	[.49, .55]	74.33	(21.08)
					6-Person Lineups				
Target-Present Cases									
Low	204	.46	[.39, .52]	.19	[.13, .24]	.65	[.58, .71]	72.75	(21.23)
Medium	200	.38	[.31, .44]	.35	[.28, .41]	.73	[.67, .79]	71.40	(20.10)
High	209	.20	[.14, .25]	.54	[.47, .61]	.74	[.67, .79]	64.74	(23.31)
Overall	613	.34	[.30, .38]	.36	[.32, .40]	.70	[.67, .74]	69.58	(21.86)
Target-Absent Cases									
Low	198	.12	[.07, .16]	.22	[.16, .27]	.34	[.27, .40]	75.25	(20.74)
Medium	198	.05	[.01, .07]	.49	[.42, .56]	.54	[.46, .60]	69.95	(21.46)
High	207	.06	[.03, .09]	.54	[.47, .60]	.60	[.53, .66]	69.32	(22.24)
Overall	603	.08	[.05, .10]	.42	[.38, .45]	.49	[.45, .53]	71.48	(21.63)
All Cases									
Low	402	.29	[.25, .33]	.20	[.16, .24]	.50	[.44, .54]	73.98	(21.00)
Medium	398	.21	[.17, .25]	.42	[.37, .47]	.63	[.58, .68]	70.68	(20.77)
High	416	.13	[.10, .16]	.54	[.49, .59]	.67	[.62, .71]	67.02	(22.87)
Overall	1216	.21	[.19, .23]	.39	[.36, .42]	.60	[.57, .63]	70.52	(21.76)
					All Lineup Sizes				
Target-Present Cases									
Low	594	.56	[.52, .60]	.10	[.08, .13]	.66	[.62, .70]	74.06	(20.82)
Medium	598	.47	[.43, .51]	.22	[.19, .26]	.69	[.65, .73]	72.99	(21.33)
High	604	.32	[.29, .36]	.36	[.32, .40]	.68	[.65, .72]	70.07	(22.07)
Overall	1796	.45	[.43, .47]	.23	[.21, .25]	.68	[.66, .70]	72.36	(21.47)
Target-Absent Cases									
Low	601	.16	[.13, .18]	.14	[.11, .17]	.30	[.26, .34]	78.79	(19.70)
Medium	605	.10	[.07, .12]	.34	[.30, .37]	.43	[.39, .47]	73.64	(20.51)
High	594	.10	[.07, .12]	.40	[.36, .44]	.50	[.46, .54]	71.87	(21.17)

Overall	1800	.12	[.10, .13]	.29	[.27, .32]	.41	[.39, .43]	74.77	(20.67)
All Cases									
Low	1195	.36	[.33, .38]	.12	[.10, .14]	.48	[.45, .51]	76.44	(20.39)
Medium	1203	.28	[.26, .31]	.28	[.26, .31]	.56	[.53, .59]	73.32	(20.92)
High	1198	.21	[.19, .24]	.38	[.35, .41]	.59	[.57, .62]	70.96	(21.64)
Overall	3596	.28	[.27, .30]	.26	[.25, .28]	.55	[.53, .56]	73.57	(21.11)

Table S13

All Trials, Stimulus Set 3. Identification Decision Patterns and Confidence Across Filler Similarity, Target Presence and Lineup Size

		Suspect Pick		Filler Pick		Choosing		Confidence	
Suspect-Filler Similarity	*Trials*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*M*	*SD*
					2-Person Lineups				
Target-Present Cases									
Low	214	.76	[.70, .81]	.01	[.00, .03]	.77	[.71, .82]	80.75	(17.67)
Medium	202	.78	[.72, .83]	.07	[.04, .11]	.85	[.80, .90]	77.33	(18.79)
High	203	.70	[.63, .76]	.12	[.07, .16]	.82	[.76, .87]	76.75	(18.44)
Overall	619	.74	[.71, .78]	.07	[.05, .09]	.81	[.78, .84]	78.32	(18.35)
Target-Absent Cases									
Low	194	.40	[.33, .46]	.06	[.03, .09]	.46	[.39, .53]	78.40	(20.16)
Medium	193	.42	[.35, .49]	.07	[.03, .10]	.49	[.41, .55]	77.41	(19.30)
High	185	.42	[.35, .49]	.15	[.09, .19]	.57	[.49, .64]	76.00	(20.01)
Overall	572	.41	[.37, .45]	.09	[.07, .11]	.50	[.46, .54]	77.29	(19.81)
All Cases									
Low	408	.59	[.54, .63]	.04	[.02, .05]	.62	[.57, .69]	79.63	(18.90)
Medium	395	.60	[.55, .65]	.07	[.04, .09]	.67	[.63, .72]	77.37	(19.02)
High	388	.57	[.52, .62]	.13	[.10, .16]	.70	[.65, .74]	79.39	(19.18)
Overall	1191	.59	[.56, .61]	.08	[.06, .09]	.66	[.64, .69]	77.83	(19.06)
					3-Person Lineups				
Target-Present Cases									
Low	191	.76	[.70, .82]	.04	[.01, .07]	.80	[.74, .86]	77.49	(19.87)
Medium	196	.70	[.63, .76]	.11	[.07, .15]	.81	[.75, .86]	76.94	(20.60)
High	194	.65	[.58, .71]	.16	[.11, .21]	.81	[.75, .86]	76.80	(20.96)
Overall	581	.70	[.66, .74]	.11	[.08, .13]	.81	[.77, .84]	77.07	(20.45)
Target-Absent Cases									
Low	206	.40	[.33, .47]	.08	[.04, .12]	.49	[.41, .55]	75.39	(17.90)
Medium	191	.39	[.32, .46]	.21	[.15, .26]	.60	[.53, .67]	74.40	(20.58)
High	211	.37	[.30, .43]	.19	[.13, .24]	.56	[.49, .62]	73.93	(19.20)

Overall	608	.39	[.35, .43]	.16	[.13, .19]	.55	[.51, .59]	74.57	(19.21)
All Cases									
Low	397	.57	[.52, .62]	.06	[.04, .09]	.64	[.59, .68]	76.40	(18.88)
Medium	387	.55	[.50, .60]	.16	[.12, .20]	.71	[.66, .75]	75.68	(20.61)
High	405	.50	[.45, .55]	.18	[.14, .21]	.68	[.63, .72]	75.31	(20.09)
Overall	1189	.54	[.51, .57]	.13	[.11, .15]	.67	[.65, .70]	75.79	(19.86)
					6-Person Lineups				
Target-Present Cases									
Low	197	.72	[.66, .78]	.10	[.05, .14]	.82	[.76, .87]	76.29	(20.85)
Medium	211	.62	[.55, .68]	.17	[.12, .22]	.80	[.74, .85]	75.31	(21.19)
High	200	.56	[.49, .63]	.25	[.19, .31]	.81	[.75, .86]	72.80	(20.23)
Overall	608	.63	[.59, .67]	.17	[.14, .20]	.81	[.78, .84]	74.80	(20.78)
Target-Absent Cases									
Low	205	.26	[.20, .32]	.21	[.16, .27]	.47	[.40, .54]	74.00	(20.45)
Medium	200	.32	[.25, .38]	.27	[.21, .33]	.59	[.52, .66]	71.05	(22.11)
High	203	.25	[.18, .30]	.37	[.31, .44]	.62	[.55, .68]	71.87	(19.18)
Overall	608	.27	[.23, .30]	.29	[.25, .32]	.56	[.52, .60]	72.32	(20.61)
All Cases									
Low	402	.49	[.44, .53]	.16	[.12, .19]	.64	[.59, .69]	75.12	(20.65)
Medium	411	.47	[.43, .52]	.22	[.18, .26]	.70	[.65, .75]	73.24	(21.72)
High	403	.40	[.35, .45]	.31	[.27, .36]	.71	[.67, .76]	72.33	(19.69)
Overall	1216	.45	[.43, .48]	.23	[.21, .25]	.68	[.66, .71]	73.56	(20.73)
					All Lineup Sizes				
Target-Present Cases									
Low	602	.75	[.71, .78]	.05	[.03, .07]	.80	[.76, .83]	78.26	(19.51)
Medium	609	.70	[.66, .74]	.12	[.09, .14]	.82	[.79, .85]	76.50	(20.21)
High	597	.64	[.60, .67]	.18	[.14, .21]	.81	[.78, .84]	75.44	(19.94)
Overall	1808	.69	[.67, .71]	.12	[.10, .13]	.81	[.79, .83]	76.74	(19.92)
Target-Absent Cases									
Low	605	.35	[.31, .39]	.12	[.09, .15]	.47	[.43, .51]	75.88	(19.58)
Medium	584	.38	[.34, .42]	.18	[.15, .21]	.56	[.52, .60]	74.25	(20.84)
High	599	.34	[.31, .38]	.24	[.20, .27]	.58	[.54, .62]	73.87	(19.48)

Overall	1788	.36	[.33, .38]	.18	[.16, .20]	.54	[.51, .56]	74.68	(19.98)
All Cases									
Low	1207	.55	[.52, .58]	.09	[.07, .10]	.63	[.61, .66]	77.07	(19.57)
Medium	1193	.54	[.51, .57]	.15	[.13, .17]	.69	[.67, .72]	75.40	(20.55)
High	1196	.49	[.46, .52]	.21	[.18, .23]	.70	[.67, .72]	74.66	(19.72)
Overall	3596	.53	[.51, .54]	.15	[.14, .16]	.67	[.66, .69]	75.71	(19.97)

Table S14

All Trials, Stimulus Set 4. Identification Decision Patterns and Confidence Across Filler Similarity, Target Presence and Lineup Size

		Suspect Pick		Filler Pick		Choosing		Confidence	
Suspect-Filler Similarity	*Trials*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*Prop*	*[95% CI]*	*M*	*SD*
					2-Person Lineups				
Target-Present Cases									
Low	197	.55	[.48, .62]	.07	[.03, .10]	.62	[.55, .68]	76.95	(19.08)
Medium	195	.64	[.57, .71]	.11	[.06, .15]	.75	[.69, .81]	73.85	(21.20)
High	187	.69	[.62, .75]	.09	[.04, .12]	.78	[.72, .84]	72.41	(21.68)
Overall	579	.63	[.58, .66]	.09	[.06, .11]	.71	[.68, .75]	74.44	(20.72)
Target-Absent Cases									
Low	209	.23	[.17, .29]	.14	[.19, .18]	.37	[.31, .44]	78.95	(18.47)
Medium	197	.19	[.13, .24]	.22	[.16, .27]	.41	[.33, .47]	78.58	(20.03)
High	206	.20	[.15, .26]	.27	[.20, .32]	.47	[.40, .54]	74.61	(19.40)
Overall	612	.21	[.18, .24]	.21	[.17, .24]	.42	[.38, .45]	77.37	(19.36)
All Cases									
Low	406	.39	[.34, .43]	.11	[.07, .13]	.49	[.44, .54]	77.98	(76.22)
Medium	392	.41	[.36, .46]	.16	[.13, .20]	.58	[.53, .62]	76.22	(20.73)
High	393	.44	[.38, .48]	.18	[.14, .22]	.62	[.57, .66]	73.56	(20.52)
Overall	1191	.41	[.38, .44]	.15	[.13, .17]	.56	[.53, .59]	75.94	(20.08)
					3-Person Lineups				
Target-Present Cases									
Low	195	.58	[.51, .65]	.09	[.05, .12]	.67	[.60, .74]	73.74	(21.61)
Medium	194	.65	[.58, .71]	.14	[.09, .19]	.79	[.73, .85]	74.90	(21.67)
High	192	.56	[.49, .63]	.20	[.14, .25]	.76	[.70, .82]	72.50	(19.20)
Overall	581	.60	[.56, .64]	.14	[.11, .17]	.74	[.71, .78]	73.72	(20.85)
Target-Absent Cases									
Low	200	.24	[.18, .30]	.18	[.12, .23]	.42	[.34, .48]	77.90	(21.28)
Medium	199	.16	[.11, .21]	.22	[.16, .27]	.38	[.31, .44]	75.33	(19.59)
High	209	.17	[.11, .22]	.30	[.24, .36]	.47	[.40, .53]	72.11	(20.79)

Overall	608	.19	[.16, .22]	.23	[.20, .26]	.42	[.38, .46]	75.07	(20.67)
All Cases									
Low	395	.41	[.36, .46]	.13	[.10, .16]	.54	[.49, .59]	75.85	(21.52)
Medium	393	.40	[.35, .45]	.18	[.14, .22]	.58	[.53, .63]	75.11	(20.62)
High	401	.36	[.31, .40]	.25	[.21, .29]	.61	[.56, .66]	72.29	(20.02)
Overall	1189	.39	[.36, .42]	.19	[.17, .21]	.58	[.55, .61]	74.41	(20.76)
					6-Person Lineups				
Target-Present Cases									
Low	195	.45	[.37, .51]	.24	[.17, .29]	.68	[.61, .74]	69.69	(20.90)
Medium	206	.45	[.38, .58]	.30	[.24, .36]	.75	[.69, .80]	71.07	(21.16)
High	198	.51	[.43, .57]	.32	[.26, .39]	.83	[.77, .88]	70.30	(22.93)
Overall	599	.47	[.43, .50]	.29	[.25, .32]	.75	[72, .79]	70.37	(21.65)
Target-Absent Cases									
Low	199	.16	[.10, .20]	.29	[.22, .35]	.44	[.37, .51]	73.62	(20.84)
Medium	209	.09	[.05, .13]	.42	[.35, .48]	.51	[.44, .57]	69.38	(23.43)
High	209	.11	[.07, .16]	.44	[.37, .51]	.56	[.48, .62]	66.89	(24.13)
Overall	617	.12	[.09, .14]	.38	[.34,.42]	.50	[.46, .54]	69.90	(23.00)
All Cases									
Low	394	.30	[.25, .34]	.26	[.22, .30]	.56	[.51, .61]	71.68	(20.94)
Medium	415	.27	[.22, .31]	.36	[.31, .40]	.62	[.58, .67]	70.22	(22.32)
High	407	.30	[.26, .35]	.38	[.33, .43]	.69	[.64, .73]	68.55	(23.59)
Overall	1216	.29	[.26, .32]	.34	[.31, .36]	.63	[.60, .65]	70.13	(22.34)
					All Lineup Sizes				
Target-Present Cases									
Low	587	.53	[.49, .57]	.13	[.10, .16]	.66	[.62, .70]	73.48	(20.73)
Medium	595	.58	[.54, .62]	.19	[.15, .22]	.76	[.73, .80]	72.23	(21.37)
High	577	.58	[.54, .62]	.20	[.17, .24]	.79	[.75, .82]	71.72	(21.33)
Overall	1759	.56	[.54, .59]	.17	[.16, .19]	.74	[.72, .76]	72.81	(21.15)
Target-Absent Cases									
Low	608	.21	[.18, 24]	.20	[.17, .23]	.41	[.37, .45]	76.86	(20.31)
Medium	605	.15	[.12, .17]	.27	[.25, .32]	.43	[.39, .47]	74.33	(21.44)
High	624	.16	[.13, .19]	.34	[.30, .37]	.50	[.46, .54]	71.18	(21.74)

Overall	1837	.17	[.16, .19]	.27	[.25, .29]	.45	[.42, .47]	74.10	(21.29)
All Cases									
Low	1195	.37	[.34, .39]	.17	[.14, .19]	.53	[.50, .56]	75.20	(20.58)
Medium	1200	.36	[.33, .39]	.24	[.21, .26]	.60	[.57, .62]	73.78	(21.40)
High	1201	.36	[.34, .39]	.27	[.25, .30]	.64	[.61, .66]	71.44	(21.54)
Overall	3596	.36	[.35, .38]	.23	[.21, .24]	.59	[.57, .60]	73.47	(21.23)

First Trial Data

Effects on the first trial data were the same as for the overall data. The 4-way association, as well as the test for 3-way associations, were again non-significant, $k = 4$, LR $\chi^2(8) = 4.76$, $p = .783$ and $k = 3$, LR $\chi^2(20) = 12.60$, $p = .894$, respectively. The test for 2-way associations was significant, $k = 2$, LR $\chi^2(18) = 680.05$, $p < .001$, with partial associations indicating significant effects on identification decisions of filler similarity, *partial* $\chi^2(4) = 112.90$, $p < .001$, lineup size, *partial* $\chi^2(4) = 150.84$, $p < .001$, and target presence, *partial* $\chi^2(2) = 408.74$, $p < .001$.

Filler similarity affected suspect identifications, $\chi^2(2) = 22.99$, $p < .001$, filler identifications, $\chi^2(2) = 103.30$, $p < .00$, and choosing, $\chi^2(2) = 14.52$, $p = .001$. Suspect identifications decreased from .46, 95% CI [.43, .49], in the low similarity condition to .43, 95% CI [.41, .47] in the medium similarity condition and .37, 95% CI [.34, .39], in the high similarity condition (OR low vs. high = 1.48; OR low vs. medium = 1.15; OR medium vs. high = 1.29). Filler identifications increased from .13, 95% CI [.11, .15], in the low similarity condition to .21, 95% CI [.18, .23], in the medium similarity condition and .30, 95% CI [.27, .32], in the high similarity condition, $\chi^2(2) = 154.20$, $p < .001$ (OR low vs. high = 2.88; OR low vs. medium = 1.75; OR medium vs. high = 1.64). Choosing increased from .59, 95% CI [.56, .62], in the low similarity condition to .64, 95% CI [.61, .66] in the medium similarity condition and .67, 95% CI [.64, .69], in the high similarity condition (OR low vs. high = 1.38; OR low vs. medium = 1.20; OR medium vs. high = 1.15).

Target presence also affected suspect identifications, $\chi^2(1) = 383.96$, $p < .001$, filler identifications, $\chi^2(1) = 27.66$, $p < .001$, and choosing, $\chi^2(1) = 242.63$, $p < .001$. Suspect identifications were higher in target-present, .58, 95% CI [.56, .60], than -absent, .26, 95% CI [.24, .28], cases (OR = 3.97); choosing was also higher in target-present, .76, 95% CI [.74, .78], than -absent, .51, 95% CI [.49, .53] cases (OR = 3.05). Filler identifications were higher

in target-absent, .25, 95% CI [.23, .27], than -present, .18, 95% CI [.16, .19], cases (OR = 1.55).

Lineup size affected identifications of suspects, $\chi^2(2) = 42.40$, $p < .001$ and fillers, $\chi^2(2) = 138.91$, $p < .001$, as well as choosing, $\chi^2(2) = 11.36$, $p = .003$. As lineups expanded from containing two to three and six members, suspect identifications decreased from, .48, 95% CI [.45, .51], to .43, 95% CI [.40, .46] and .35, 95% CI [.32, .38] (OR 2 vs. 6 = 1.71; OR 2 vs. 3 = 1.23; OR 3 vs. 6 = 1.39) filler identifications increased from .12, 95% CI [.10, .14], to .20, 95% CI [.18, .22] and .32, 95% CI [.29, .34] (OR 2 vs. 6 = 3.32; OR 2 vs. 3 = 1.79; OR 3 vs. 6 = 1.86), and overall choosing increased marginally from, .60, 95% CI [.57, .63], to .63, 95% CI [.60, .66] and .67, 95% CI [.64, .69] (OR 2 vs. 6 = 1.33; OR 2 vs. 3 = 1.12; OR 3 vs. 6 = 1.19).

As with the overall data, confidence was affected by filler similarity, test statistic = 18.77, $p < .001$, and lineup size, test statistic = 19.85, $p < .001$, but not target presence, test statistic = 0.31, $p = .577$, $d = 0.04$. As filler similarity increased confidence decreased from $M = 74.76$, $SD = 19.40$, in the low similarity condition to $M = 74.11$, $SD = 19.30$, in the medium similarity condition and $M = 72.25$, $SD = 20.24$, in the high similarity condition (d low vs. high = 0.13; d low vs. medium = 0.03; d medium vs. high = 0.09). Confidence also decreased as lineup size increased from two ($M = 75.59$, $SD = 19.14$) to three ($M = 74.65$, $SD = 18.81$) to six ($M = 70.84$, $SD = 20.72$) (d 2 vs. 6 = 0.24; d 2 vs. 3 = 0.05; d 3 vs. 6 = 0.19). None of the interactions were significant (test statistic < 6.31, p > .179).

By Stimulus

Stimulus 1

The 4-way and 3-way associations between variables were non-significant, $k = 4$, LR $\chi^2(8) = 10.81$, $p = .213$ and $k = 3$, LR $\chi^2(20) = 19.16$, $p = .512$, respectively. The test for 2-way associations was significant, $k = 2$, LR $\chi^2(18) = 509.79$, $p < .001$, with partial

associations indicating significant effects on identification decisions of filler similarity, *partial* $\chi^2(4) = 180.03$, $p < .001$, lineup size, *partial* $\chi^2(4) = 129.72$, $p < .001$, and target presence, *partial* $\chi^2(2) = 210.02$, $p < .001$.

Filler similarity significantly affected suspect identifications, $\chi^2(2) = 96.15$, $p < .001$, which decreased from .67, 95% CI [.64, .70], in the low similarity condition to .59, 95% CI [.57, .62] in the medium similarity condition to .47, 95% CI [.45, .50], in the high similarity condition (OR low vs. high = 2.25; OR low vs. medium = 1.39; OR medium vs. high = 1.62). Filler identifications increased from .09, 95% CI [.08, .11], in the low similarity condition to .17, 95% CI [.15, .20], in the medium similarity condition and .29, 95% CI [.26, .32], in the high similarity condition, $\chi^2(2) = 154.20$, $p < .001$ (OR low vs. high = 3.93; OR low vs. medium = 204; OR medium vs. high = 1.93). Choosing was not affected by filler similarity, $\chi^2(2) = 0.10$, $p = .950$ (OR low vs. high = 1.00; OR low vs. medium = 1.03; OR medium vs. high = 1.03).

Target presence affected suspect identifications, $\chi^2(1) = 185.15$, $p < .001$, filler identifications, $\chi^2(1) = 18.27$, $p < .001$, and choosing, $\chi^2(1) = 142.55$, $p < .001$. As one might expect, suspect identifications were higher in target-present, .69, 95% CI [.67, .71], than -absent, .47, 95% CI [.44, .49], cases (OR = 2.55) and choosing was lower in target-absent, .68, 95% CI [.66, .70], than -present, .85, 95% CI [.83, .87] cases (OR = 2.65). Filler identifications were higher in target-absent, .21, 95% CI [.19, .23], than -present, .16, 95% CI [.14, .17], cases (OR = 1.45).

Lineup size affected identifications of suspects, $\chi^2(2) = 44.63$, $p < .001$ and fillers, $\chi^2(2) = 120.72$, $p < .001$, but not overall lineup choosing, $\chi^2(2) = 6.20$, $p = .045$ (OR 2 vs. 6 = 1.13; OR 2 vs. 3 = 1.18; OR 3 vs. 6 = 1.05). Suspect identifications decreased as lineups increased from 2-person, .64, 95% CI [.61, .67], to 3-person, .59, 95% CI [.57, .62] to 6-person, .51, 95% CI [.48, .54], lineups (OR 2 vs. 6 = 1.73; OR 2 vs. 3 = 1.22; OR 3 vs. 6 =

1.41). Filler identifications increased as lineups increased from 2-person, .10, 95% CI [.08, .12], to 3-person, .18, 95% CI [.16, .20] to 6-person, .28, 95% CI [.25, .30], lineups (OR 2 vs. 6 = 3.36; OR 2 vs. 3 = 1.93; OR 3 vs. 6 =1.74).

Confidence was affected by filler similarity, test statistic = 16.95, $p < .001$, lineup size, test statistic = 19.31, $p < .001$, and target presence, test statistic = 9.06, $p = .003$. As filler similarity increased confidence decreased from $M = 75.13$, $SD = 20.21$, in the low similarity condition to $M = 73.08$, $SD = 21.23$, in the medium similarity condition and $M = 71.05$, $SD = 21.81$, in the high similarity condition (d low vs. high = 0.19; d low vs. medium = 0.10; d medium vs. high = 0.09). Confidence also decreased as lineup size increased from two ($M = 74.79$, $SD = 20.45$) to three ($M = 73.86$, $SD = 20.53$) to six ($M = 70.67$, $SD = 22.20$) (d 2 vs. 6 = 0.19; d 2 vs. 3 = 0.05; d 3 vs. 6 = 0.15). Confidence was higher in target-present ($M = 74.31$, $SD = 20.50$) than -absent ($M = 71.87$, $SD = 21.72$) cases ($d = 0.12$). None of the 2-way interactions were significant (test statistics < 4.85, p > .304); however, there was a 3-way interaction between filler similarity, lineup size and target presence on confidence (test statistic = 11.81, $p = .020$). Refer to Table S11 for descriptive statistics.

Stimulus 2

As for Stimulus 1, both the 4-way association and the 3-way associations were non-significant, $k = 4$, LR $\chi^2(8) = 7.57$, $p = .477$ and $k = 3$, LR $\chi^2(20) = 17.84$, $p = .598$, respectively. The test for 2-way associations was significant, $k = 2$, LR $\chi^2(18) = 951.81$, $p < .001$, with partial associations showing significant effects of filler similarity, *partial* $\chi^2(4) = 246.08$, $p < .001$, lineup size, *partial* $\chi^2(4) = 129.72$, $p < .001$, and target presence, *partial* $\chi^2(2) = 180.49$, $p < .001$, on identifications.

As filler similarity increased, suspect identifications decreased from .36, 95% CI [.33, .38] in low similarity condition, to .28, 95% CI [.26, .31] in the medium similarity condition, and .21, 95% CI [.19, .24], in the high similarity condition, $\chi^2(2) = 61.45$, $p < .001$ (OR low

vs. high = 2.06; OR low vs. medium = 1.42; OR medium vs. high = 1.45). Filler identifications increased from .12, 95% CI [.10, .14], in the low similarity condition, to .28, 95% CI [.26, .31], in the medium similarity condition and .38, 95% CI [.35, .41], in the high similarity condition, $\chi^2(2) = 210.08$, $p < .001$(OR low vs. high = 4.40; OR low vs. medium = 2.79; OR medium vs. high = 1.58). Lineup choosing increased across the low, .48, 95% CI [.45, .51], medium .56, 95% CI [.53, .59], and high .59, CI [.57, .62], similarity conditions, $\chi^2(2) = 96.15$, $p < .001$ (OR low vs. high = 1.58; OR low vs. medium = 1.39; OR medium vs. high = 1.14).

Target presence affected suspect identifications, $\chi^2(1) = 489.35$, $p < .001$, filler identifications, $\chi^2(1) = 19.67$, $p < .001$, and lineup choosing, $\chi^2(1) = 259.74$, $p < .001$. Again, suspect identifications were higher in target-present, .45, 95% CI [.43, .47], than -absent, .12, 95% CI [.10, .13], cases (OR = 6.14), choosing was lower in target-absent, .41, 95% CI [.39, .43], than -present, .68, 95% CI [.66, .70], cases (OR = 3.03) and filler identifications were higher in target-absent, .29, 95% CI [.27, .32], than -present, .23, 95% CI [.21, .25], cases (OR = 1.40).

Lineup size affected identifications of suspects, $\chi^2(2) = 53.13$, $p < .001$ and fillers, $\chi^2(2) = 160.49$, $p < .001$, as well as lineup choosing, $\chi^2(2) = 21.52$, $p < .001$. Suspect identifications decreased as lineups increased from 2-person, .34, 95% CI [.31, .37], to 3-person, .30, 95% CI [.28, .33] to 6-person, .21, 95% CI [.19, .23], lineups (OR 2 vs. 6 = 1.94; OR 2 vs. 3 = 1.20; OR 3 vs. 6 = 1.62). Filler identifications increased as lineup size expanded from 2-person, .17, 95% CI [.15, .19], to 3-person, .22, 95% CI [.20, .24] to 6-person, .39, 95% CI [.36, .42], lineups (OR 2 vs. 6 = 3.04; OR 2 vs. 3 = 1.36; OR 3 vs. 6 = 2.24). Choosing was lower from 2- and 3-person lineups, .51, 95% CI [.49, .54], and .52, 95% CI [.49, .55], respectively, than 6-person lineups, .60, 95% CI [.57, .63] (OR 2 vs. 6 = 1.42; OR 2 vs. 3 = 1.04; OR 3 vs. 6 = 1.36).

Confidence was affected by filler similarity, test statistic = 42.24, $p < .001$, lineup size, test statistic = 31.76, $p < .001$, and target presence, test statistic = 8.14, $p = .003$. As filler similarity increased confidence decreased from $M = 76.44$, $SD = 20.39$, in the low similarity condition to $M = 73.32$, $SD = 20.92$, in the medium similarity condition and $M = 70.96$, $SD = 21.64$, in the high similarity condition (d low vs. high = 0.26; d low vs. medium = 0.15; d medium vs. high = 0.11). Confidence also decreased as lineup size increased from two ($M = 75.92$, $SD = 20.08$) to three ($M = 74.33$, $SD = 21.08$) to six ($M = 70.52$, $SD = 21.76$) (d 2 vs. 6 = 0.26; d 2 vs. 3 = 0.08; d 3 vs. 6 = 0.18). Confidence was higher in target-absent ($M = 74.77$, $SD = 20.67$) than -present ($M = 72.36$, $SD = 21.47$) cases ($d = 0.11$). There was a 2-way interaction between filler similarity and target presence, test statistic = 11.12, $p = .004$, with a larger decrease in confidence across increases in filler similarity for target-absent cases (see Table S12). None of the other interactions were significant (test statistics < 6.72, $p > .153$).

Stimulus 3

Again the 4- and 3-way associations were non-significant, $k = 4$, LR $\chi^2(8) = 9.93$, $p = .270$ and $k = 3$, LR $\chi^2(20) = 20.71$, $p = .414$, respectively, while the test for 2-way associations was significant, $k = 2$, LR $\chi^2(18) = 630.58$, $p < .001$. Partial associations showed significant effects of filler similarity, *partial* $\chi^2(4) = 77.74$, $p < .001$, lineup size, *partial* $\chi^2(4) = 118.08$, $p < .001$, and target presence, *partial* $\chi^2(2) = 432.76$, $p < .001$, on identifications.

As with the previous stimuli, filler similarity affected suspect identifications, $\chi^2(2) = 9.70$, $p = .008$. Suspect identifications were similar in the low similarity condition, .55, 95% CI [.52, .58], and the medium similarity condition, .54, 95% CI [.51, .57], but decreased in the high similarity, condition, .49, 95% CI [.46, .52] (OR low vs. high = 1.26; OR low vs. medium = 1.03; OR medium vs. high = 1.23). Increasing filler similarity increased filler identifications across the low, .09, 95% CI [.07, .10], medium, .15, 95% CI [.13, .17], and

high .21, 95% CI [.18, .23], conditions, $\chi^2(2) = 71.22$, $p < .001$ (OR low vs. high = 2.80; OR low vs. medium = 1.90; OR medium vs. high = 1.47). Overall choosing was also affected by filler similarity, $\chi^2(2) = 13.51$, $p = .001$, increasing from .63, 95% CI [.61, .66], in the low similarity condition to .69, 95% CI [.67, .72], in the medium similarity condition, but not decreasing meaningfully further in the high similarity condition, .70, 95% CI [.67, .72] (OR low vs. high = 1.33; OR low vs. medium = 1.30; OR medium vs. high = 1.02).

Target presence affected suspect identifications, $\chi^2(1) = 407.59$, $p < .001$, filler identifications, $\chi^2(1) = 30.74$, $p < .001$, and lineup choosing, $\chi^2(1) = 299.62$, $p < .001$. Suspect identifications were higher in target-present, .69, 95% CI [.67, .71], than -absent, .36, 95% CI [.33, .38], cases (OR = 4.07), Choosing was lower in target-absent, .54, 95% CI [.51, .56], than -present, .81, 95% CI [.79, .83], cases (OR = 3.64), and filler identifications were higher in target-absent, .18, 95% CI [.16, .20], than -present, .12, 95% CI [.10, .13], cases (OR = 1.70).

Lineup size affected identifications of suspects, $\chi^2(2) = 43.24$, $p < .001$ and fillers, $\chi^2(2) = 111.39$, $p < .001$, but not lineup choosing, $\chi^2(2) = 1.02$, $p = .602$ (OR 2 vs. 6 = 1.10; OR 2 vs. 3 = 1.05; OR 3 vs. 6 = 1.05). Suspect identifications decreased as lineups increased from 2-person, .59, 95% CI [.56, .61], to 3-person, .54, 95% CI [.51, .57] to 6-person, .45, 95% CI [.43, .48], lineups (OR 2 vs. 6 = 1.70; OR 2 vs. 3 = 1.19; OR 3 vs. 6 = 1.42). Filler identifications increased as lineup size expanded from 2-person, .08, 95% CI [.06, .09], to 3-person, .13, 95% CI [.11, .15] to 6-person, .23, 95% CI [.21, .25], lineups (OR 2 vs. 6 = 3.47; OR 2 vs. 3 = 1.79; OR 3 vs. 6 = 1.94).

Confidence was affected by filler similarity, test statistic = 10.06, $p = .007$, lineup size, test statistic = 24.59, $p < .001$, and target presence, test statistic = 10.88, $p = .001$. As filler similarity increased confidence decreased from $M = 77.07$, $SD = 19.57$, in the low similarity condition to $M = 75.40$, $SD = 20.55$, in the medium similarity condition and $M =$

74.66, *SD* = 19.72, in the high similarity condition (*d* low vs. high = 0.12; *d* low vs. medium = 0.08; *d* medium vs. high = 0.04). Confidence also decreased as lineup size increased from two (*M* = 77.83, *SD* = 19.06) to three (*M* = 75.79, *SD* = 19.86) to six (*M* = 73.56, *SD* = 20.73) (*d* 2 vs. 6 = 0.21; *d* 2 vs. 3 = 0.10; *d* 3 vs. 6 =0.11). Confidence was higher in target-present (*M* = 76.74, *SD* = 19.92) than -absent (*M* = 74.68, *SD* = 19.98) cases (*d* = 0.10). None of the interactions were significant (test statistics < 4.67, *p*s > .300).

Stimulus 4

As for all the other stimuli, the 4-way association between variables was non-significant, $k = 4$, LR $\chi^2(8) = 4.70$, $p = .789$; however, the test for 3-way associations was significant, $k = 3$, LR $\chi^2(20) = 31.76$, $p = .046$. Partial associations indicated an interaction between filler similarity and target presence on identifications, *partial* $\chi^2(4) = 17.98$, $p = .001$.

Separate examinations of the each identification decision type indicated variation in the effect of filler similarity on suspect identifications depending on target presence, $\chi^2(2) = 13.80$, $p = .001$. In target-absent cases, suspect identifications were higher in the low similarity condition, .21, 95% CI [.18, .24], than the medium and high similarity conditions, .15, 95% CI [.12, .17] and .16, 95% CI [.13, .19], respectively (OR low vs. high = 1.38; OR low vs. medium = 1.57; OR medium vs. high = 1.13). In target-present cases, suspect identifications were lower in the low similarity condition, .53, 95% CI [.49, .57], than the medium and high similarity conditions, .58, 95% CI [.54, .62] and .58, 95% CI [.54, .62], respectively (OR low vs. high = 1.26; OR low vs. medium = 1.22; OR medium vs. high = 1.03). With the adjusted alpha level, there was no interaction between filler similarity and target presence on filler identifications or overall choosing, $\chi^2(2) = 0.84$, $p = .657$ and $\chi^2(2) = 6.42$, $p = .040$, respectively. Filler identifications increased across the low, .17, 95% CI [.14, .19], medium, .24, 95% CI [.21, .26] and high, .27, 95% CI [.25, .30] similarity conditions, *partial* $\chi^2(2) = 41.98$, $p < .001$ (OR low vs. high = 1.89; OR low vs. medium = 1.56; OR

medium vs. high = 1.21). Overall choosing also increased across the low, .53, 95% CI [.50, .56], medium, .60, 95% CI [.57, .62], and high .64, 95% CI [.61, .66], similarity conditions, *partial* $\chi^2(2) = 32.75$, $p < .00$ (OR low vs. high = 1.55; OR low vs. medium = 1.30; OR medium vs. high = 1.19). Filler identifications were lower in target-present, .17, 95% CI [.16, .19], than -absent, .27, 95% CI [.25, .29], cases, *partial* $\chi^2(1) = 52.33$, $p < .001$, OR = 1.80; lineup choosing was higher in target-present, .74, 95% CI [.72, .76], than -absent, .45, 95% CI [.42, .47] cases, *partial* $\chi^2(2) = 320.68$, $p < .001$, OR = 3.45.

The test for 2-way associations was also significant, $k = 2$, LR $\chi^2(18) = 814.86$, $p < .001$, with partial associations showing a significant effect of lineup size on identifications, *partial* $\chi^2(2) = 141.82$, $p < .001$. Separate analyses of each identification decision showed effects of lineup size on suspect picks, $\chi^2(2) = 43.45$, $p < .001$, filler picks, $\chi^2(2) = 133.20$, $p < .001$ and lineup choosing, $\chi^2(2) = 11.31$, $p < .001$. Suspect identifications decreased as lineups increased from 2-person, .41, 95% CI [.38, .44], to 3-person, .39, 95% CI [.36, .42] to 6-person, .29, 95% CI [.26, .32], lineups (OR 2 vs. 6 = 1.71; OR 2 vs. 3 = 1.10; OR 3 vs. 6 = 1.56). Filler identifications increased as lineup size expanded from 2-person, .15, 95% CI [.13, .17], to 3-person, .19, 95% CI [.17, .21] to 6-person, .34, 95% CI [.31, .36], lineups (OR 2 vs. 6 = 2.87; OR 2 vs. 3 = 1.32; OR 3 vs. 6 = 2.17). Overall choosing from 2- and 3-person lineups, .56, 95% CI [.53, .59], and .58, 95% CI [.55, .61], respectively, was lower than from 6-person lineups, .63, 95% CI [.60, .65] (OR 2 vs. 6 = 1.31; OR 2 vs. 3 = 1.07; OR 3 vs. 6 = 1.22).

Confidence was affected by filler similarity, test statistic = 20.38, $p < .001$, lineup size, test statistic = 34.74, $p < .001$, and target presence, test statistic = 6.31, $p = .013$. As filler similarity increased confidence decreased from $M = 75.20$, $SD = 20.58$, in the low similarity condition to $M = 73.78$, $SD = 21.40$, in the medium similarity condition and $M = 71.44$, $SD = 21.54$, in the high similarity condition (*d* low vs. high = 0.18; *d* low vs. medium

= 0.07; *d* medium vs. high = 0.11). Confidence also decreased as lineup size increased from two (*M* = 75.94, *SD* = 20.08) to three (*M* = 74.41, *SD* = 20.76) to six (*M* = 70.13, *SD* = 22.34) (*d* 2 vs. 6 = 0.27; *d* 2 vs. 3 = 0.07; *d* 3 vs. 6 = 0.20). Confidence was higher in target-absent (*M* = 74.10, *SD* = 21.29) than -present (*M* = 72.81, *SD* = 21.15) cases (*d* = 0.06). None of the interactions were significant (test statistics < 5.86, *p*s > .078).

Accuracy

For the first trial data, both the 4-way and 3-way associations between variables were non-significant, $k = 4$, LR $\chi^2(4) = 2.29$, $p = .683$ and $k = 3$, LR $\chi^2(12) = 10.18$, $p = .600$, respectively. The test for 2-way associations was significant, $k = 2$, LR $\chi^2(13) = 131.81$, $p < .001$, with partial associations indicating effects on accuracy of filler similarity, $\chi^2(2) = 35.57$, $p < .001$, lineup size, $\chi^2(2) = 50.32$, $p < .001$ and target presence, $\chi^2(1) = 30.07$, $p < .001$. Accuracy decreased from 58.91%, 95% CI [56.02, 61.71], in the low similarity filler condition to 55.08%, 95% CI [52.22, 57.85], and 47.15%, 95% CI [44.34, 49.88], in the medium and high similarity conditions, respectively (OR low vs. high = 1.61; OR low vs. medium = 1.17; OR medium vs. high = 1.37). As lineup size increased, accuracy decreased from 59.95%, 95% CI [57.13, 62.69], in the 2-person lineups to 55.00%, 95% CI [52.13, 57.79], in the 3-person lineups and 45.89%, 95% CI [43.05, 48.65], in the 6-person lineups (OR 2 vs. 6 = 1.77; OR 2 vs. 3 = 1.22; OR 3 vs. 6 = 1.44). Finally, accuracy was higher in target-present, 58.22%, 95% CI [55.89, 60.49], than -absent 49.04, 95% CI [46.72, 51.31], lineups (OR = 1.45).

For Stimulus 1, there was a significant 4-way association between variables, $k = 4$, LR $\chi^2(4) = 9.83$, $p = .043$, characterized by differences in the effect of filler similarity on accuracy across different lineup sizes in target-absent cases (note in target-present cases, increased filler similarity consistently decreased identification accuracy; see target-present suspect identification rates in Table S11). For 2-person lineups, target-absent accuracy

increased in the high similarity condition, 38.14%, 95 % CI [31.05, 44.72], compared with the medium and low similarity conditions, 31.02%, 95% CI [24.12, 37.38] and 33.49%, 95% CI [26.85, 39.65], respectively (OR low vs. high = 1.22; OR low vs. medium = 1.12; OR medium vs. high = 1.37). For 3-person lineups, target-absent accuracy did not differ across filler similarity conditions (33.33, 95% CI [26.46, 39.69], 31.47, 95% CI [24.73, 37.70] and 30.46, 95% CI [23.78, 36.63] in the low, medium and high similarity conditions, respectively) (OR low vs. high = 1.14; OR low vs. medium = 1.09; OR medium vs. high = 1.05). For 6-person lineups, target-absent accuracy decreased as filler similarity increased, from 35.58%, 95% CI [28.83, 41.85 in the low similarity condition, to 31.58%, 95% CI [25.04, 37.64] in the medium similarity condition and 21.00%, 95% CI [15.11, 26.40] in the high similarity condition (OR low vs. high = 2.08; OR low vs. medium = 1.20; OR medium vs. high = 1.72).

For Stimulus 2, both the 4-way and 3-way associations between variables were non-significant, $k = 4$, LR $\chi^2(4) = 7.12$, $p = .129$ and $k = 3$, LR $\chi^2(12) = 10.80$, $p = .546$, respectively. The test for 2-way associations was significant, $k = 2$, LR $\chi^2(13) = 258.52$, $p < .001$, with partial associations indicating effects on accuracy of filler similarity, $\chi^2(2) = 119.98$, $p < .001$, lineup size, $\chi^2(2) = 69.64$, $p < .001$ and target presence, $\chi^2(1) = 69.75$, $p < .001$. Accuracy decreased from 63.10%, 95% CI [60.32, 65.79], in the low similarity filler condition to 51.70%, 95% CI [48.83, 54.48], and 40.98%, 95% CI [38.15, 43.72], in the medium and high similarity conditions, respectively (OR low vs. high = 2.46; OR low vs. medium = 1.60; OR medium vs. high = 1.54). As lineup size increased, accuracy decreased from 58.35%, 95% CI [55.51, 61.11], in the 2-person lineups to 55.17%, 95% CI [52.30, 57.95], in the 3-person lineups and 42.43%, 95% CI [39.61, 45.17], in the 6-person lineups (OR 2 vs. 6 = 1.90; OR 2 vs. 3 = 1.14; OR 3 vs. 6 = 1.67). Finally, accuracy was lower in

target-present, 45.04%, 95% CI [42.71, 47.31], than -absent 58.78%, 95% CI [56.48, 61.03], lineups (OR = 1.74).

For Stimulus 3, both the 4-way and 3-way associations between variables were non-significant, $k = 4$, LR $\chi^2(4) = 1.20$, $p = .878$ and $k = 3$, LR $\chi^2(12) = 12.53$, $p = .404$, respectively. The test for 2-way associations was significant, $k = 2$, LR $\chi^2(13) = 254.06$, $p < .001$, with partial associations indicating effects on accuracy of filler similarity, $\chi^2(2) = 31.74$, $p < .001$, lineup size, $\chi^2(2) = 18.53$, $p < .001$ and target presence, $\chi^2(1) = 200.75$, $p < .001$. Accuracy decreased from 63.63%, 95% CI [60.87, 66.30], in the low similarity filler condition to 57.17%, 95% CI [54.32, 59.94], and 52.68%, 95% CI [49.81, 55.47], in the medium and high similarity conditions, respectively (OR low vs. high = 1.57; OR low vs. medium = 1.31; OR medium vs. high = 1.20). As lineup size increased, accuracy decreased from 62.55%, 95% CI [59.76, 65.26], in the 2-person lineups to 57.44%, 95% CI [54.59, 60.21], in the 3-person lineups and 53.62%, 95% CI [50.78, 56.38], in the 6-person lineups (OR 2 vs. 6 = 1.44; OR 2 vs. 3 = 1.24; OR 3 vs. 6 = 1.17). Finally, accuracy was higher in target-present, 69.36%, 95% CI [67.21, 71.46], than -absent 46.20%, 95% CI [43.86, 48.48], lineups (OR = 2.64).

For Stimulus 4, the 4-way association between variables was non-significant, $k = 4$, LR $\chi^2(4) = 4.94$, $p = .294$; the test for 3-way associations was significant, $k = 3$, LR $\chi^2(12) = 22.88$, $p = .029$. Partial associations indicated an interaction between filler similarity and target presence on accuracy. In target-present cases, as filler similarity increased, accuracy increased from 52.64%, 95% CI [48.52, 56.59] in the low similarity condition to 57.65%, 95% CI [53.60, 61.54] and 58.41%, 95% CI [54.30, 62.35] in the medium and high similarity conditions, respectively (OR low vs. high = 1.26; OR low vs. medium = 1.22; OR medium vs. high = 1.03). In target absent cases, as filler similarity increased, accuracy decreased from 59.05%, 95% CI [55.06, 62.88] in the low similarity condition, to 56.86%, 95% CI

[52.83, 60.72] in the medium similarity condition and 50.16%, 95% CI [46.16, 54.00] in the high similarity condition (OR low vs. high = 1.43; OR low vs. medium = 1.09; OR medium vs. high = 1.31). The test for 2-way associations was also significant, $k = 2$, LR $\chi^2(13) = 47.55$, $p < .001$, with partial associations indicating a significant effect of lineup size on accuracy, $\chi^2(2) = 43.24$, $p < .001$. As lineup size increased, accuracy decreased from 60.37%, 95% CI [57.55, 63.11], in the 2-person lineups to 58.87%, 95% CI [56.03, 61.62], in the 3-person lineups and 48.19%, 95% CI [45.34, 50.96], in the 6-person lineups (OR 2 vs. 6 = 1.64; OR 2 vs. 3 = 1.06; OR 3 vs. 6 = 1.54).

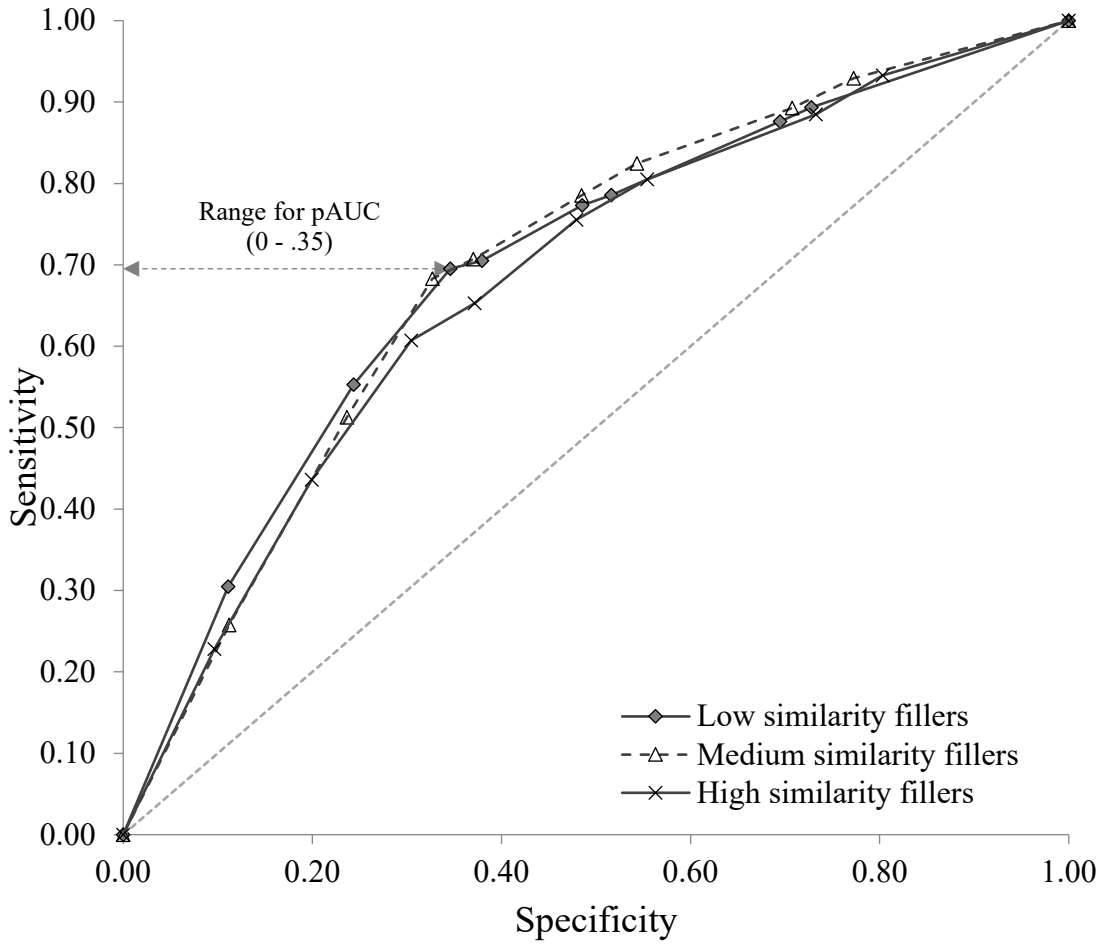

Figure S1. ROC curves for the 2-person low, medium and high similarity filler lineups. AUC = .699, 95% CI [.673, .724], .699, 95% CI [.673, .725], and .682, 95% CI [.655, .707] in the low, medium and high similarity conditions (*p*s > .355, Ds < 0.94); pAUC = .140, 95% CI [.125, .156], .132, 95% CI [.118, .148], and .128, 95% CI [.114, .143] in the low, medium and high similarity conditions (*p*s > .251, Ds < 1.15).

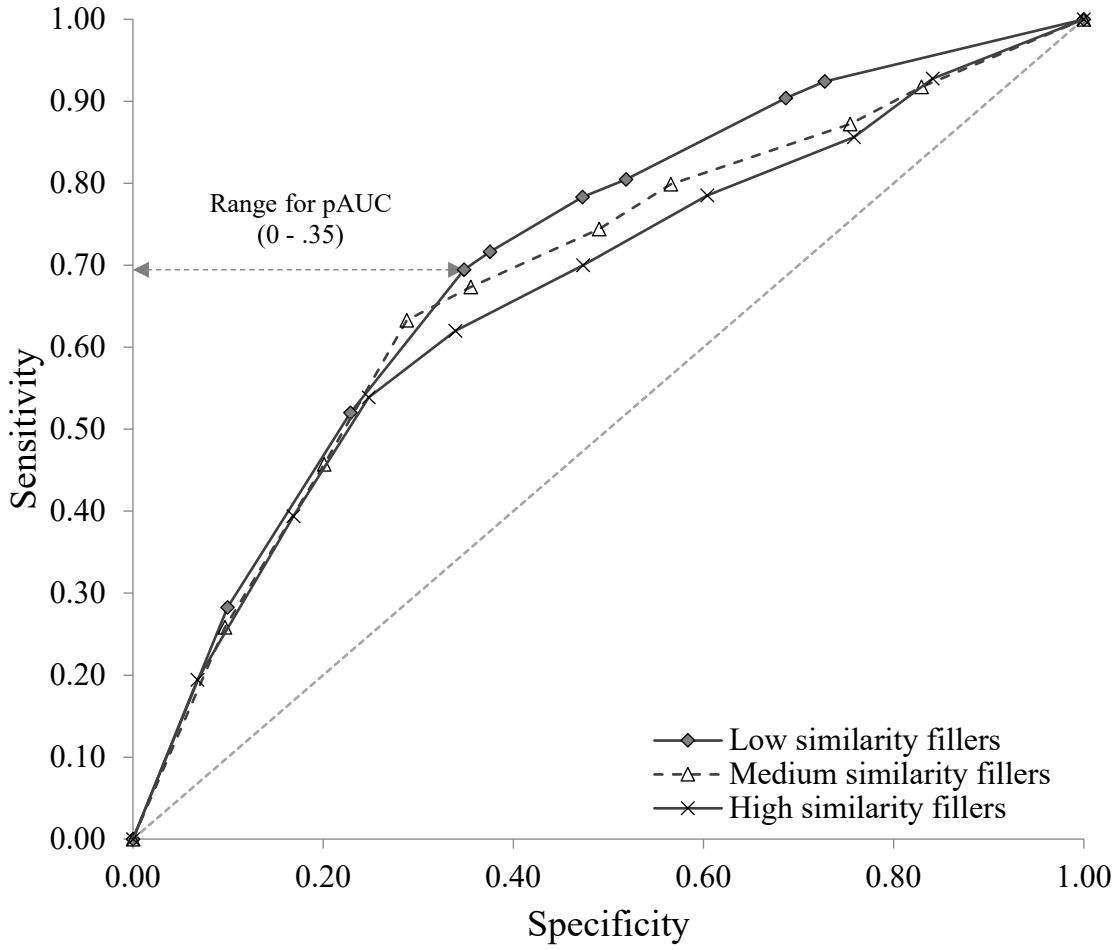

Figure S2. ROC curves for the 3-person low, medium and high similarity filler lineups. AUC = .710, 95% CI [.685, .735], .683, 95% CI [.657, .710], and .671, 95% CI [.645, .697] in the low, medium and high similarity conditions (low vs. high p = .033, D = 2.14; low vs. medium p = .151, D = 1.43, medium vs. high p = .512, D = 0.66); pAUC = .140, 95% CI [.125, .155], .138, 95% CI [.123, .153], and .137, 95% CI [.123, .151] in the low, medium and high similarity conditions (ps > .817, Ds < 0.24)

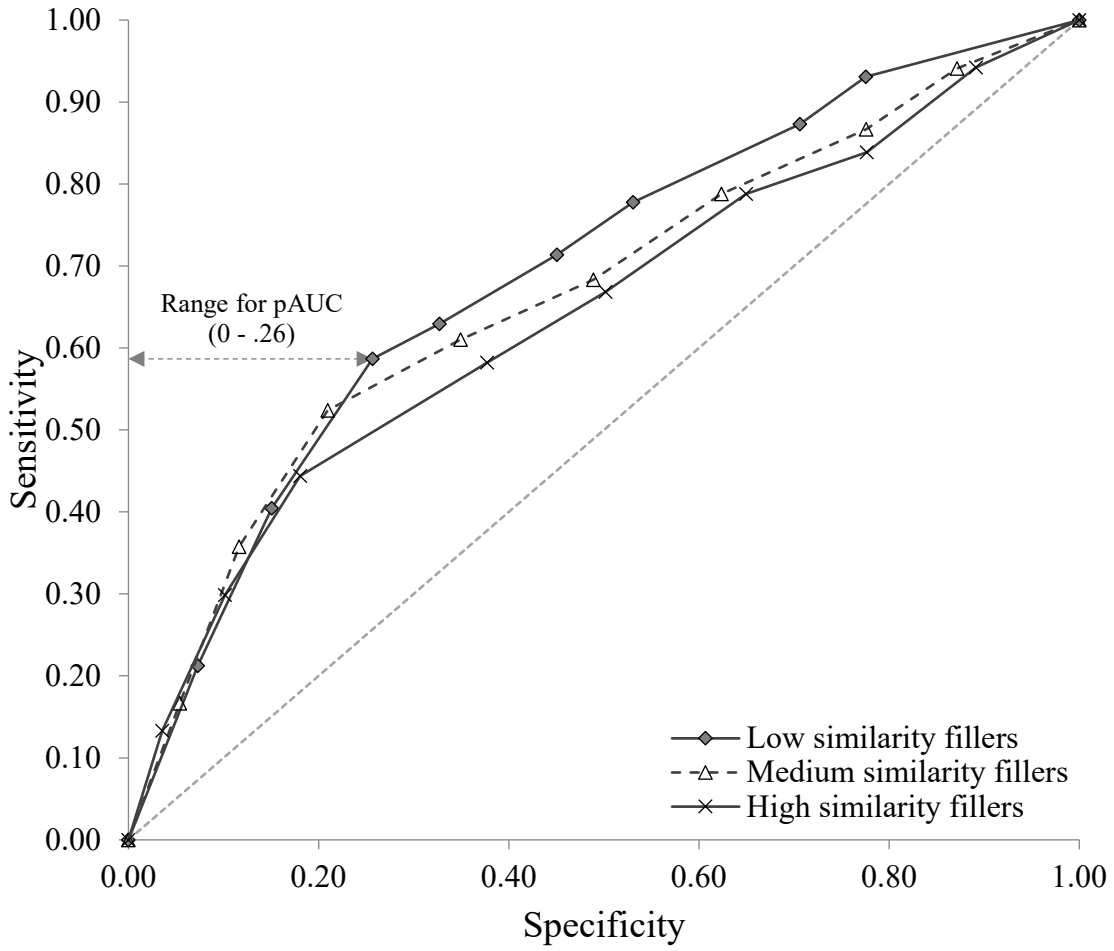

Figure S3. ROC curves for the 6-person low, medium and high similarity filler lineups. AUC = .694, 95% CI [.668, .719], .668, 95% CI [.641, .693], and .643, 95% CI [.617, .669] in the low, medium and high similarity conditions (low vs. high p = .007, D = 2.69; low vs. medium p = .158, D = 1.41, medium vs. high p = .199, D = 1.29); pAUC = .086, 95% CI [.075, .098], .089, 95% CI [.079, .100], and .083, 95% CI [.074, .093] in the low, medium and high similarity conditions (ps > .440, Ds < 0.78)

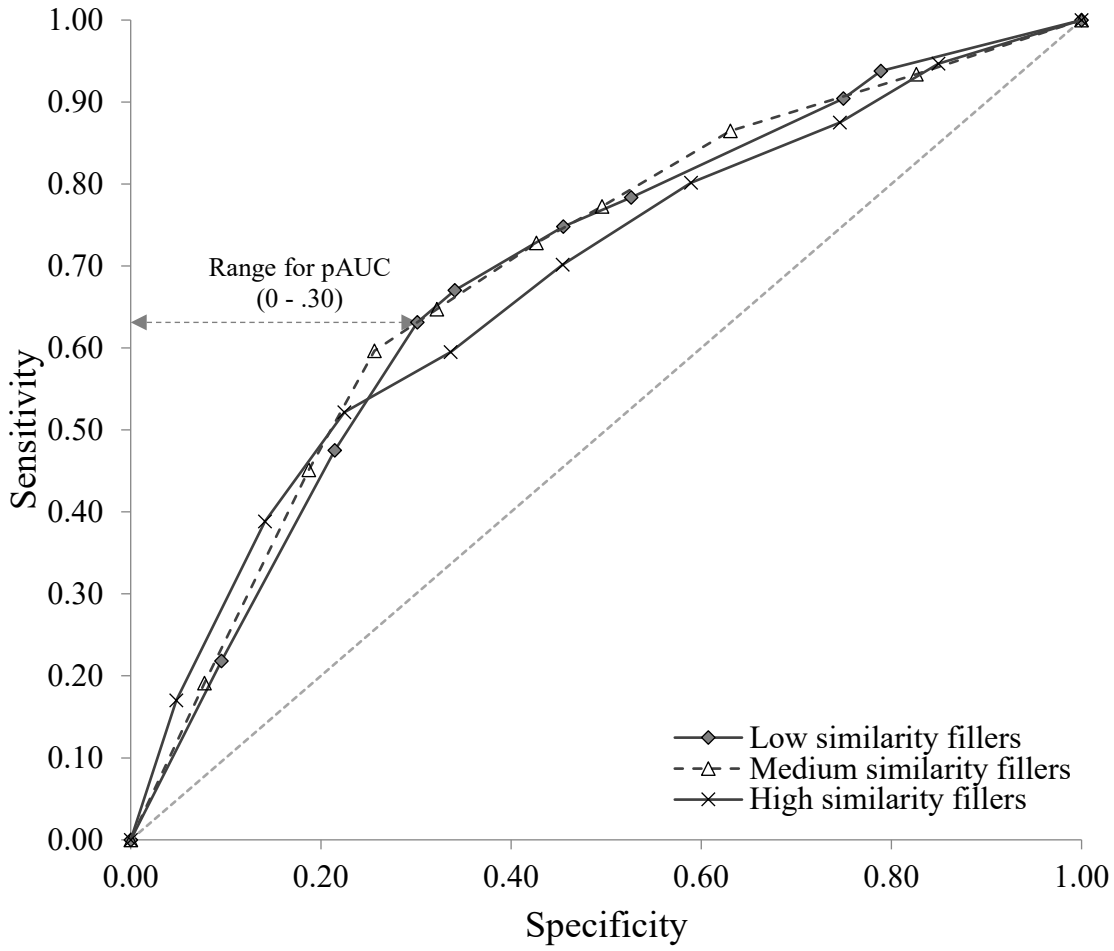

Figure S4. ROC curves for the first trial low, medium and high similarity filler lineups. AUC = .690, 95% CI [.660, .720], .698, 95% CI [.669, .727], and .681, 95% CI [.652, .710] in the low, medium and high similarity conditions (*p*s > .413, Ds < 0.83); pAUC = .099, 95% CI [.084, .115], .106, 95% CI [.090, .122], and .109, 95% CI [.097, .123] in the low, medium and high similarity conditions (*p*s > .317, Ds < 1.01)

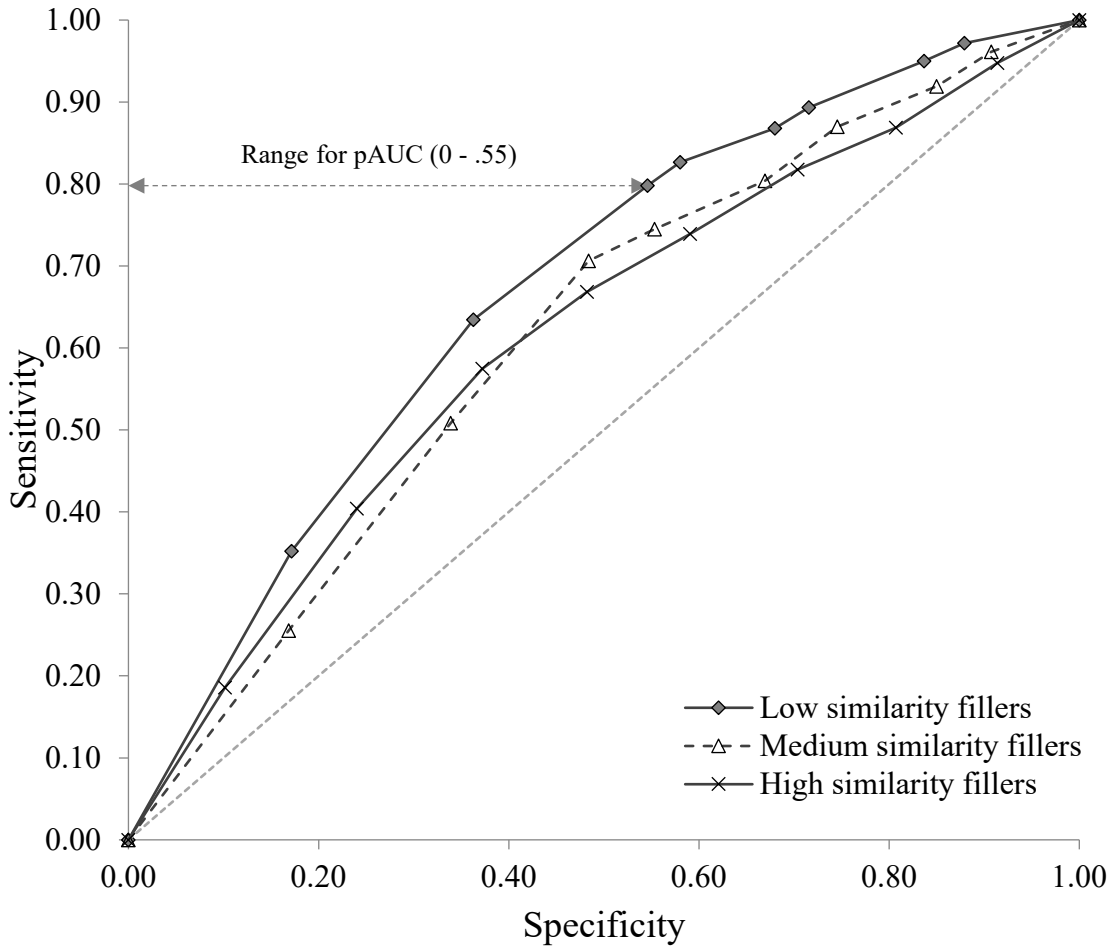

Figure S5. ROC curves for the Stimulus 1 low, medium and high similarity filler lineups. AUC = .671, 95% CI [.641, .700], .617, 95% CI [.586, .648], and .616, 95% CI [.584, .647] in the low, medium and high similarity conditions (low vs. high p = .011, D = 2.53; low vs. medium p = .015, D = 2.42, medium vs. high p = .957, D = 0.05); pAUC = .259, 95% CI [.236, .282], .222, 95% CI [.199, .247], and .230, 95% CI [.208, .252] in the low, medium and high similarity conditions (low vs. high p = .083, D = 1.73; low vs. medium p = .030, D = 2.18, medium vs. high p = .651, D = 0.45)

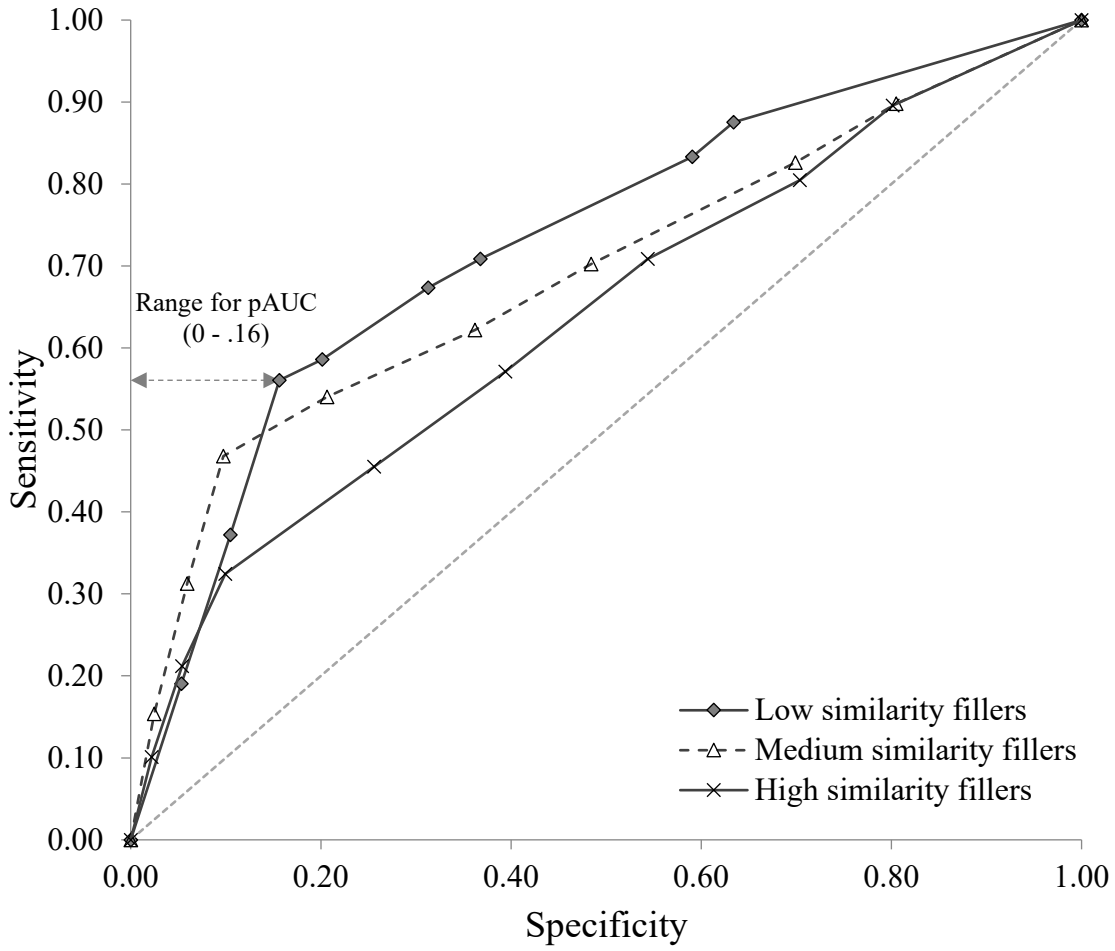

Figure S6. ROC curves for the Stimulus 2 low, medium and high similarity filler lineups. AUC = .730, 95% CI [.701, .758], .692, 95% CI [.661, .721], and .638, 95% CI [.608, .669] in the low, medium and high similarity conditions (low vs. high $p < .001$, D = 4.28; low vs. medium $p = .072$, D = 1.80, medium vs. high $p = .015$, D = 2.42); pAUC = .046, 95% CI [.036, .056], .055, 95% CI [.047, .064], and .040, 95% CI [.033, .046] in the low, medium and high similarity conditions (low vs. high $p = .331$, D = 0.97; low vs. medium $p = .145$, D = 1.46, medium vs. high $p = .004$, D = 2.91).

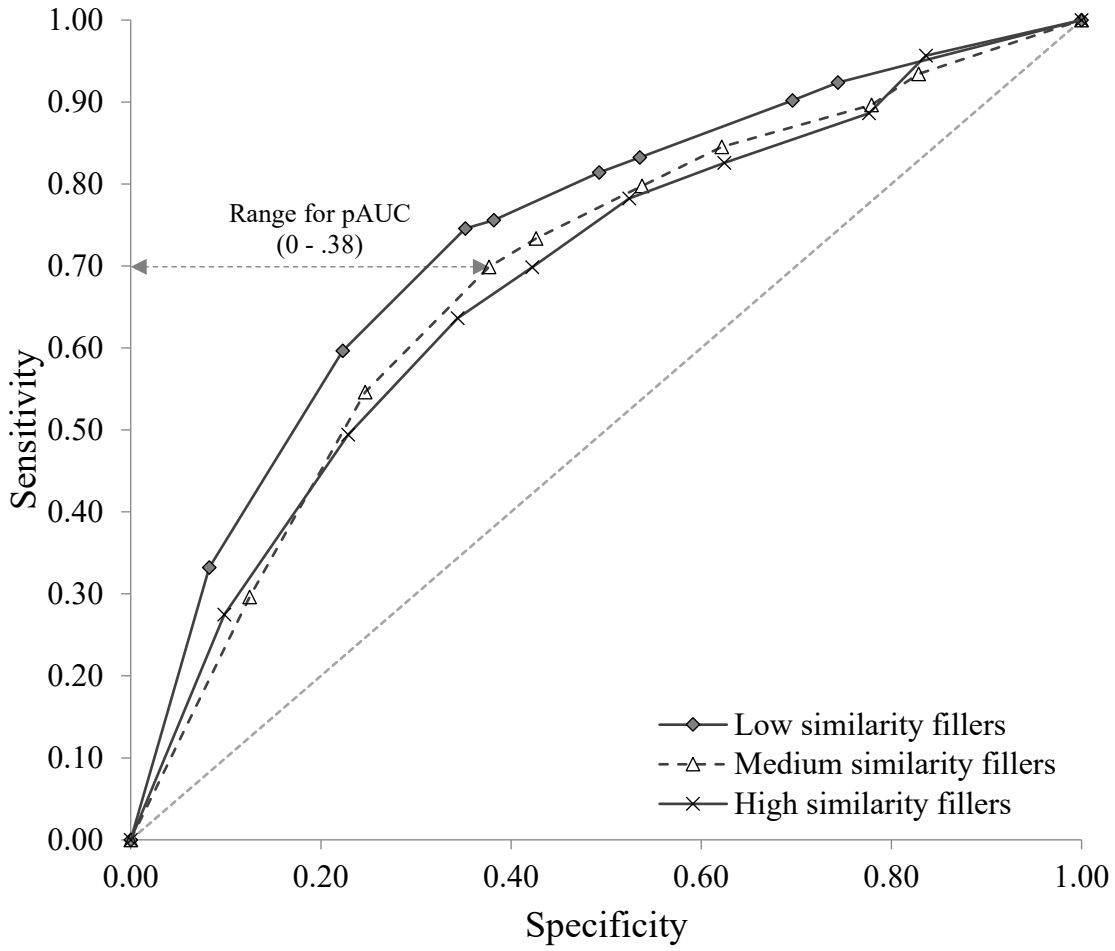

Figure S7. ROC curves for the Stimulus 3 low, medium and high similarity filler lineups. AUC = .739, 95% CI [.712, .767], .689, 95% CI [.659, .718], and .683, 95% CI [.653, .712] in the low, medium and high similarity conditions (low vs. high p =.006, D = 2.75; low vs. medium p = .014, D = 2.46, medium vs. high p = .771, D = 0.29); pAUC = .186, 95% CI [.170, .205], .153, 95% CI [.135, .171], and .152, 95% CI [.134, .169] in the low, medium and high similarity conditions (low vs. high p =.007, D = 2.70; low vs. medium p = .011, D = 2.53, medium vs. high p = .950, D = 0.06).

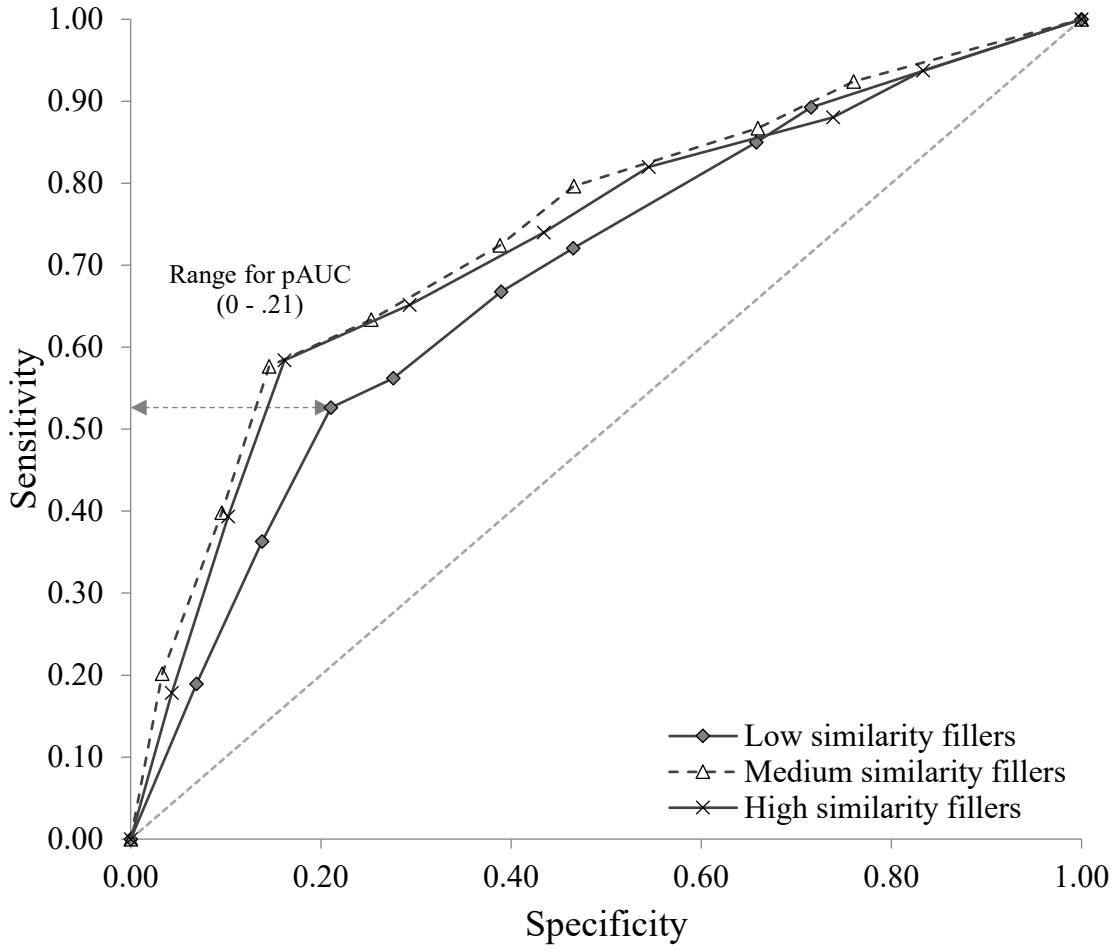

Figure S8. ROC curves for the Stimulus 4 low, medium and high similarity filler lineups. AUC = .687, 95% CI [.656, .716], .744, 95% CI [.716, .771], and .728, 95% CI [.699, .756] in the low, medium and high similarity conditions (low vs. high p =.050, D = 1.97; low vs. medium p = .005, D = 2.81, medium vs. high p = .416, D = 0.81); pAUC = .058, 95% CI [.047, .069], .085, 95% CI [.074, .096], and .079, 95% CI [.067, .090] in the low, medium and high similarity conditions (low vs. high p =.009, D = 2.60; low vs. medium p = .001, D = 3.40, medium vs. high p = .450, D = 0.76).

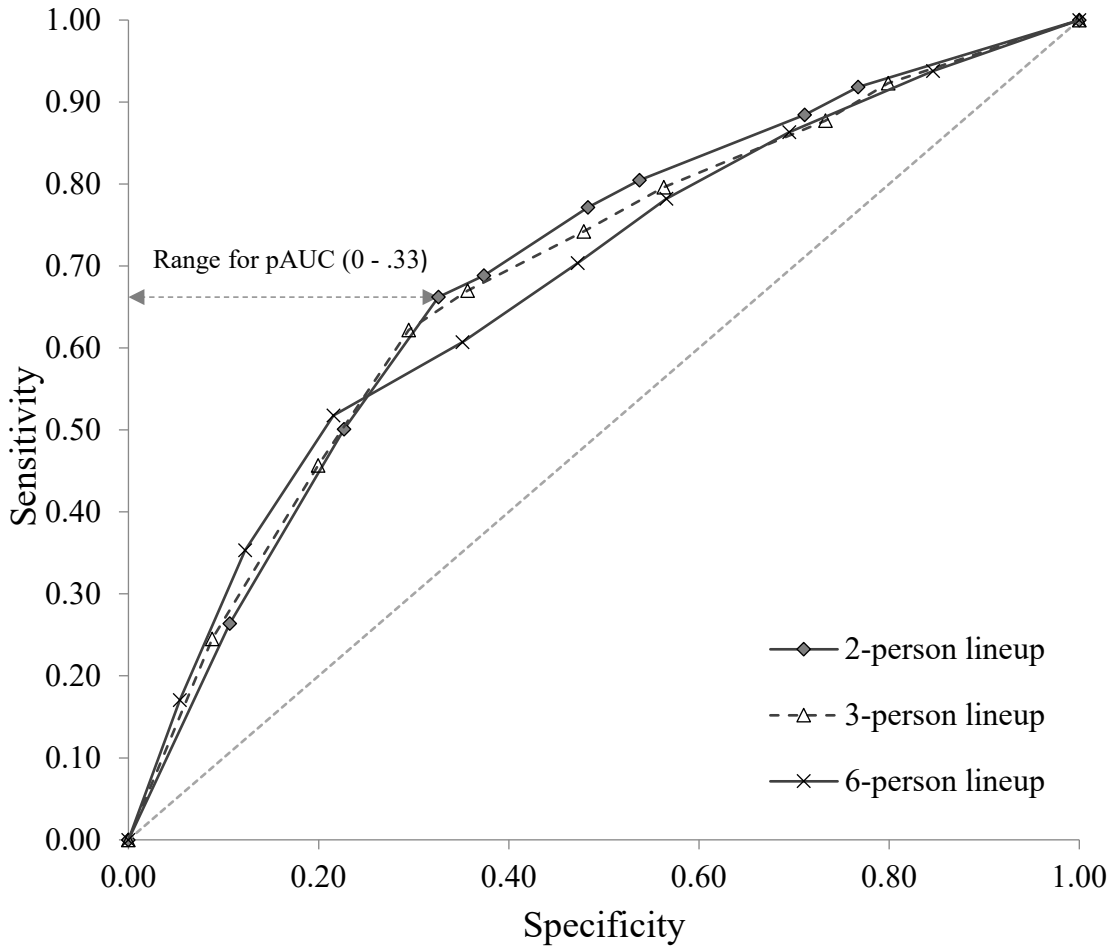

Figure S9. ROC curves for the 2-person, 3-person and 6-person lineups. AUC = .696, 95% CI [.681, .710], .687, 95% CI [.672, .703], and .680, 95% CI [.665, .694] in the 2-, 3- and 6-person lineup conditions (*p*s > .125, Ds < 1.54); pAUC = .120, 95% CI [.113, .129], .124, 95% CI [.115, .132], and .127, 95% CI [.119, .134] in the 2-, 3- and 6-person lineup conditions (*p*s > .271, Ds < 1.11).

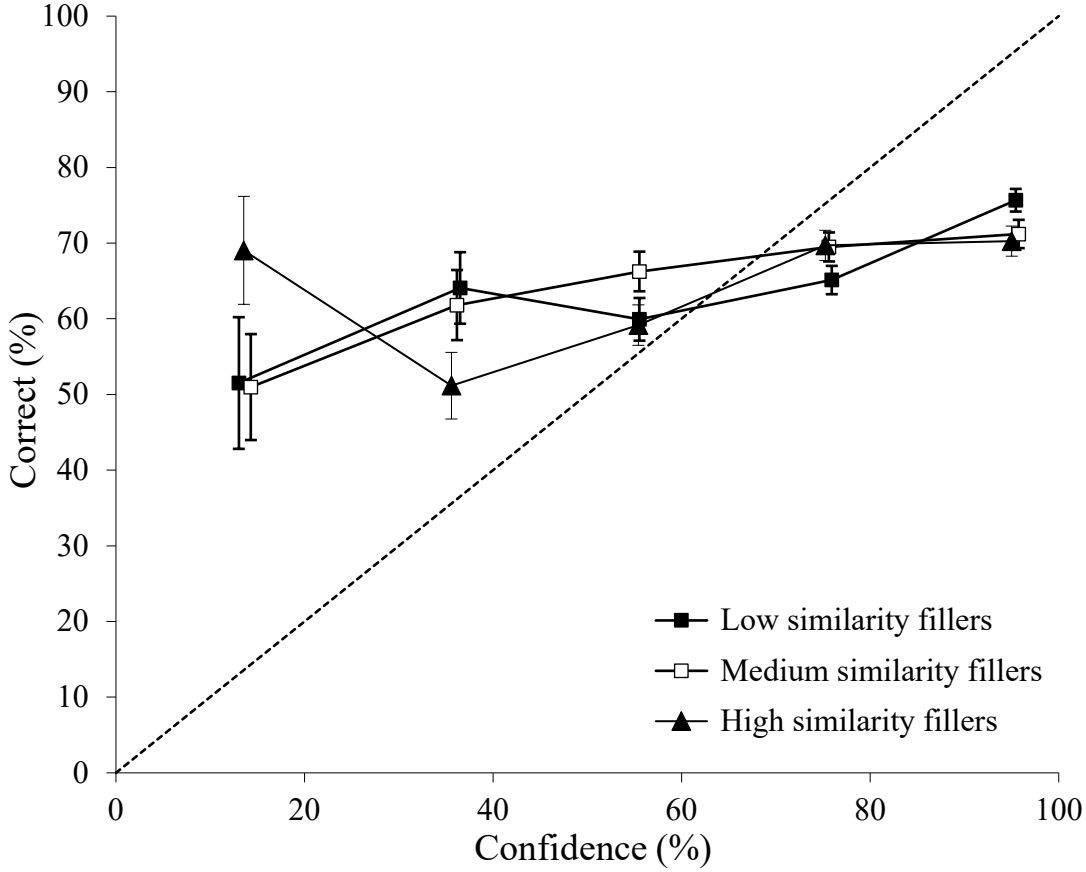

Figure S10. Overall data: Calibration curves for non-choosers in the low, medium and high similarity conditions.

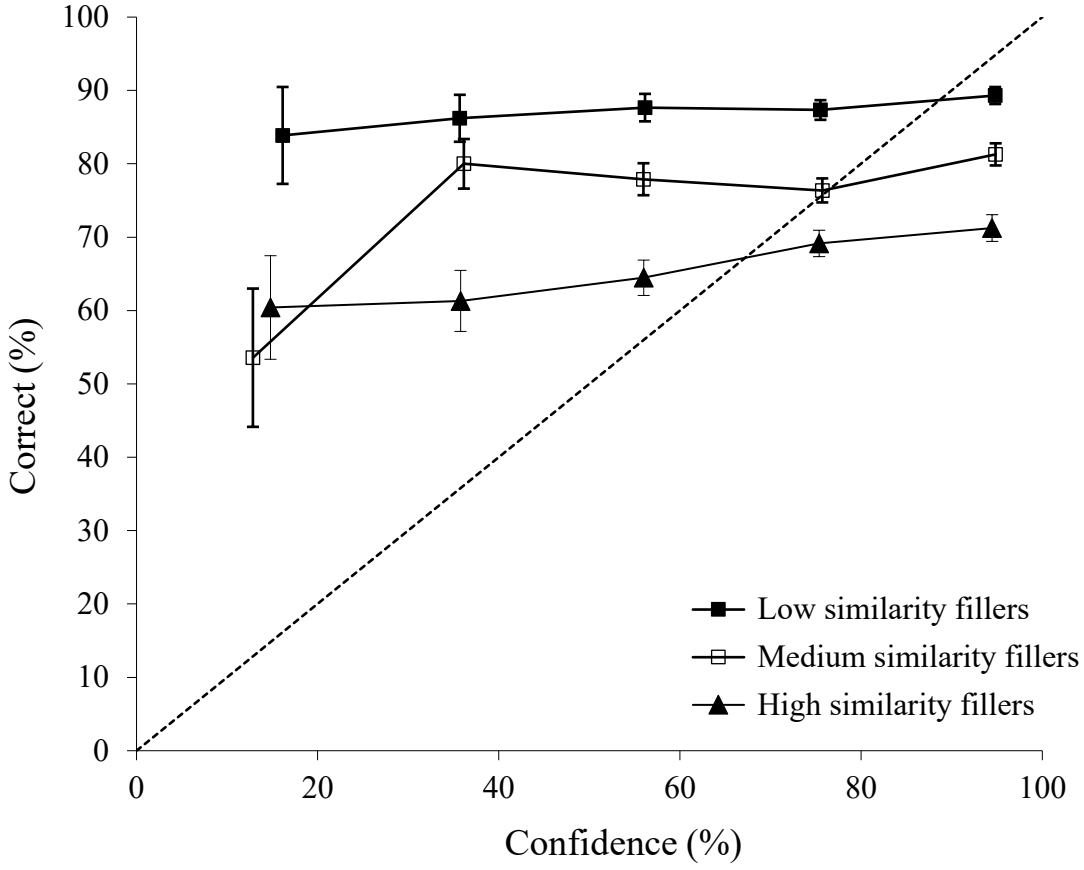

Figure S11. Overall data: Calibration curves for target-present choosers in the low, medium and high similarity conditions.

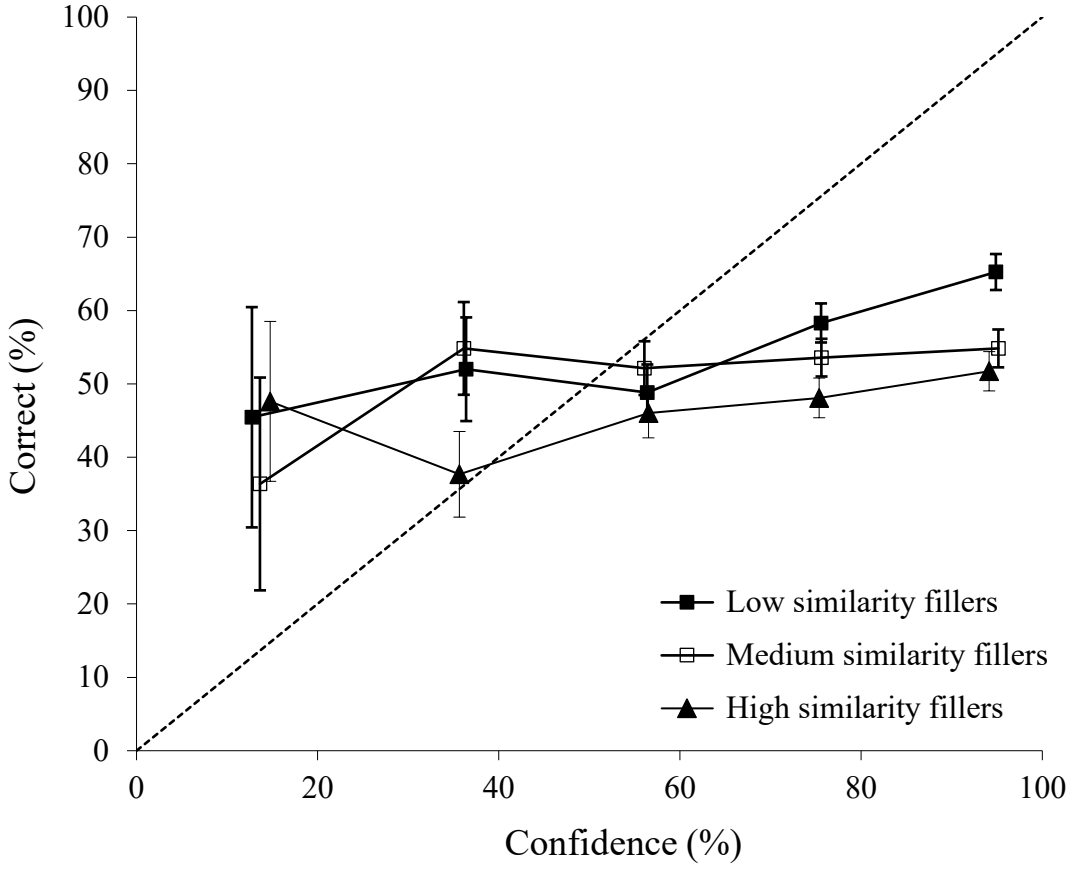

Figure S12. Overall data: Calibration curves for choosers from 2-person lineups in the low, medium and high similarity conditions.

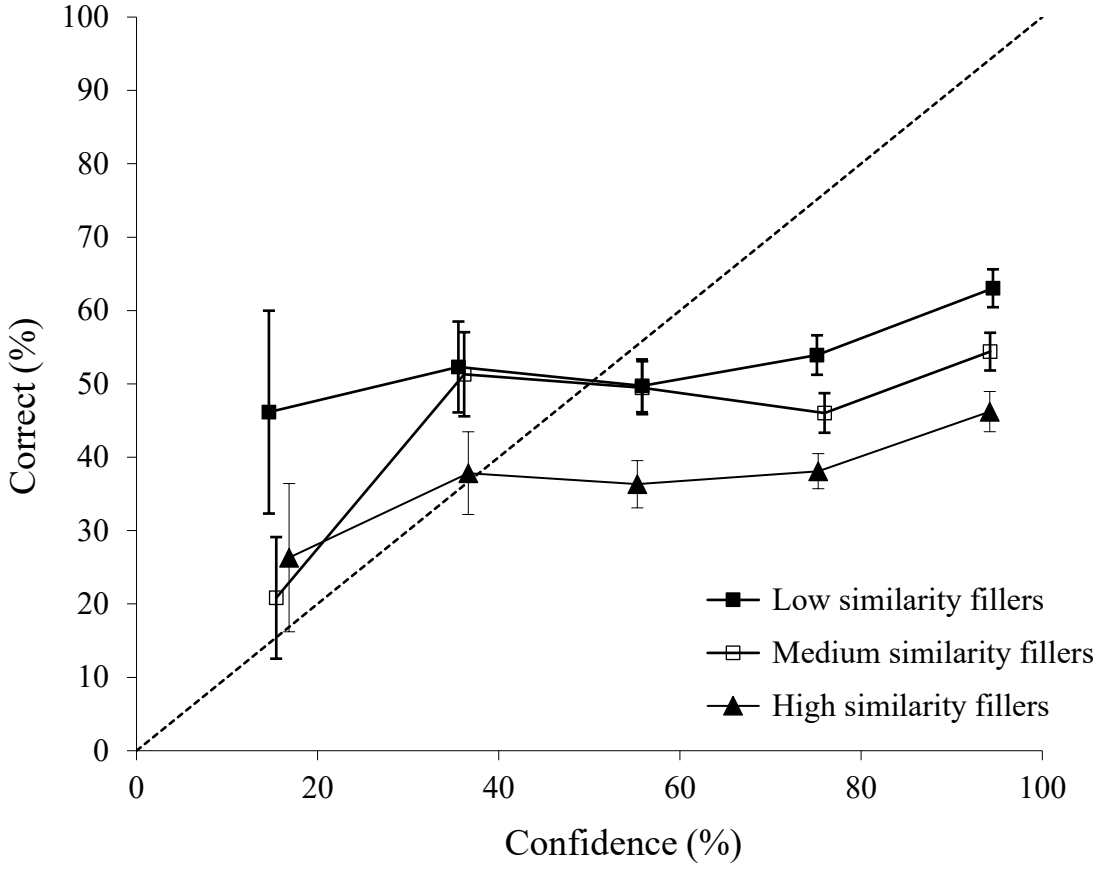

Figure S13. Overall data: Calibration curves for choosers from 3-person lineups in the low, medium and high similarity conditions.

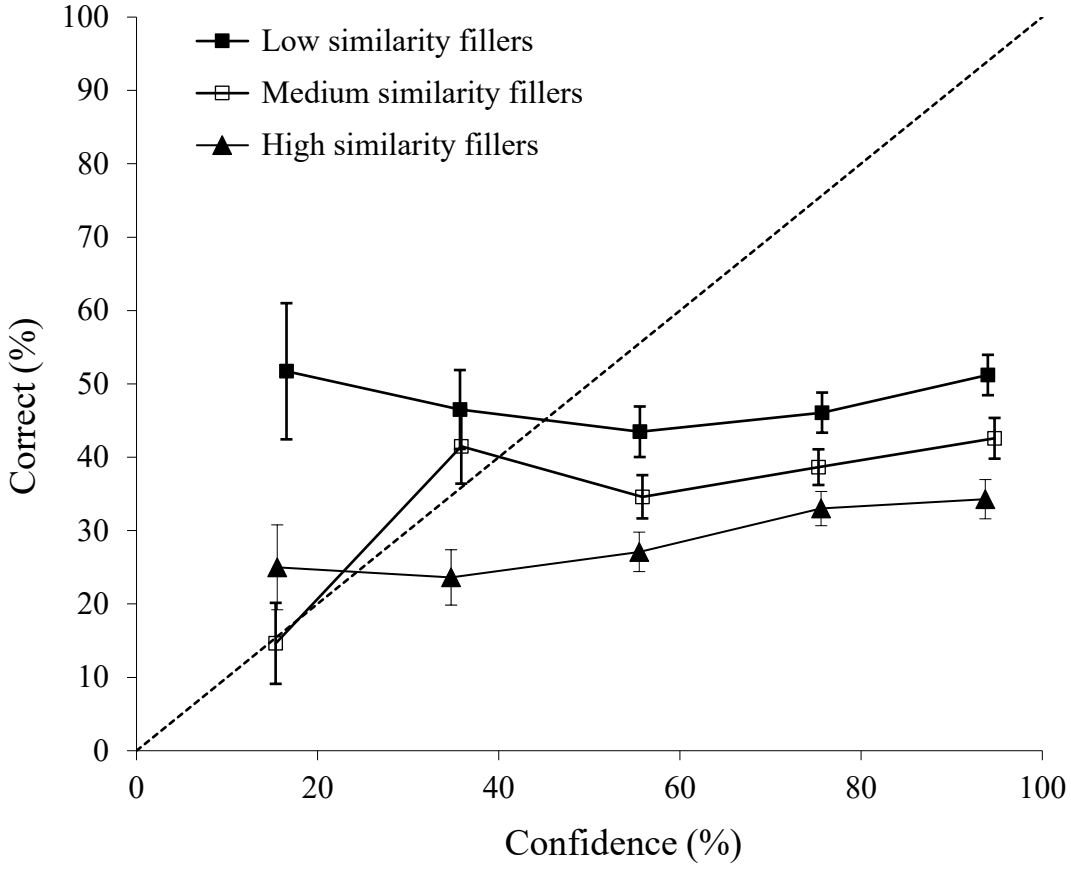

Figure S14. Overall data: Calibration curves for choosers from 6-person lineups in the low, medium and high similarity conditions.

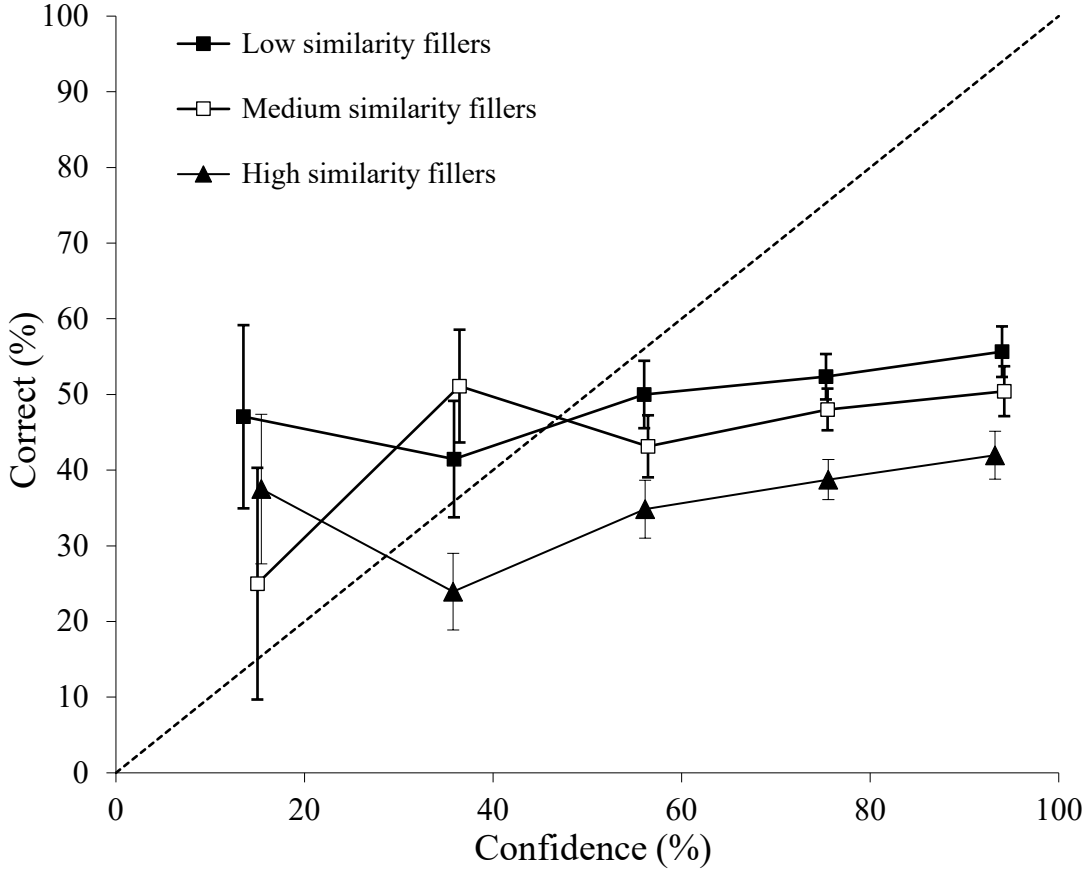

Figure S15. First trial data: Calibration curves for choosers in the low, medium and high similarity conditions.

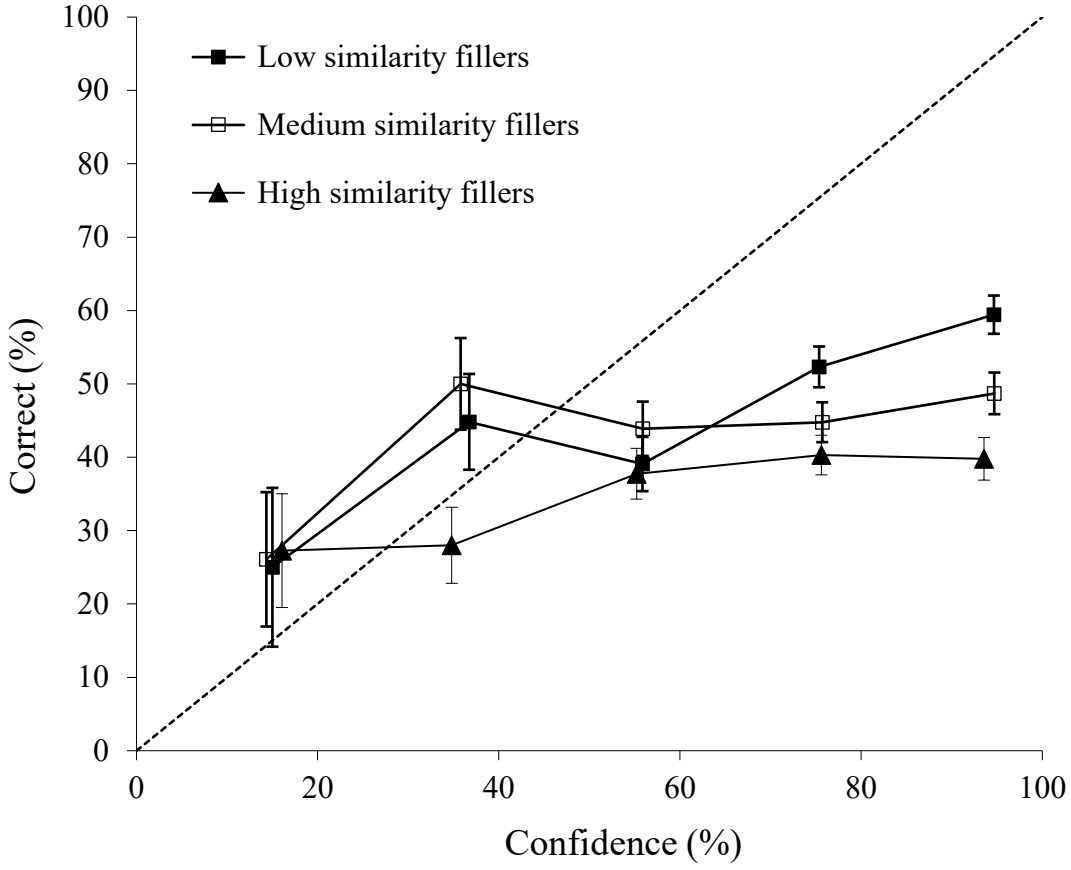

Figure S16. Stimulus 1 data: Calibration curves for choosers in the low, medium and high similarity conditions.

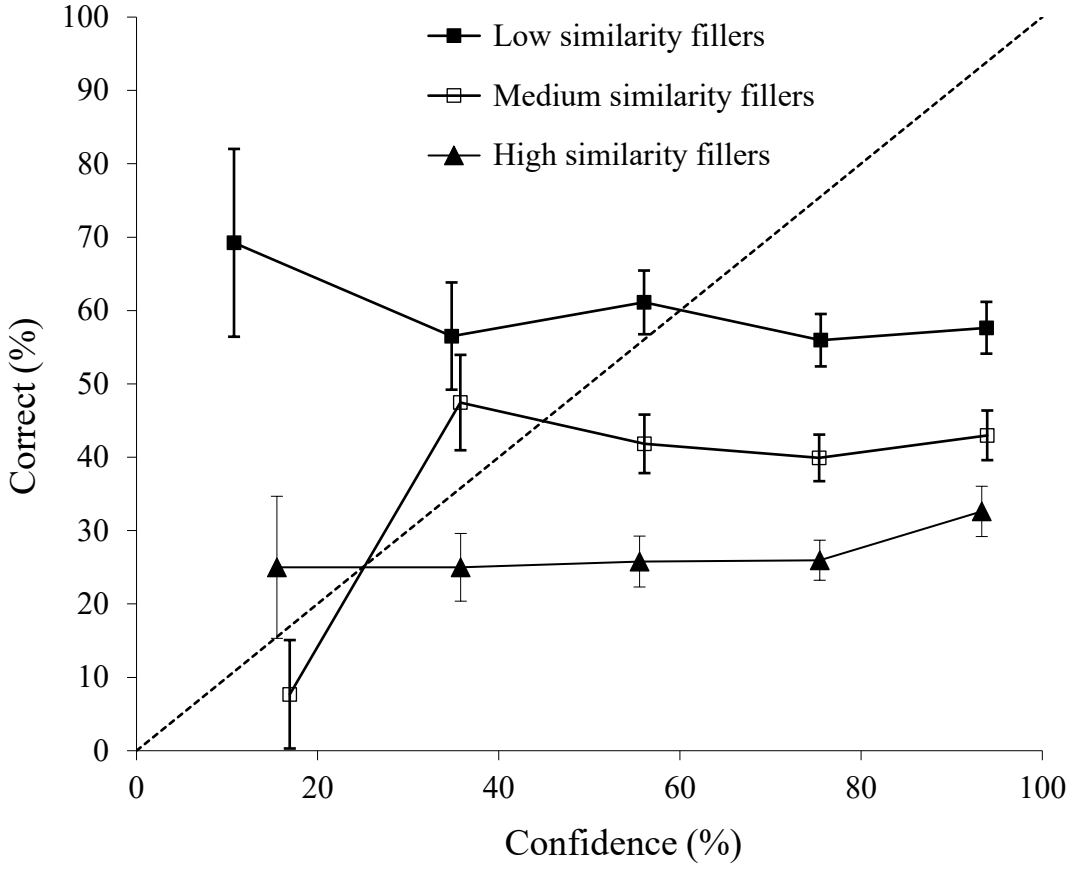

Figure S17. Stimulus 2 data: Calibration curves for choosers in the low, medium and high similarity conditions.

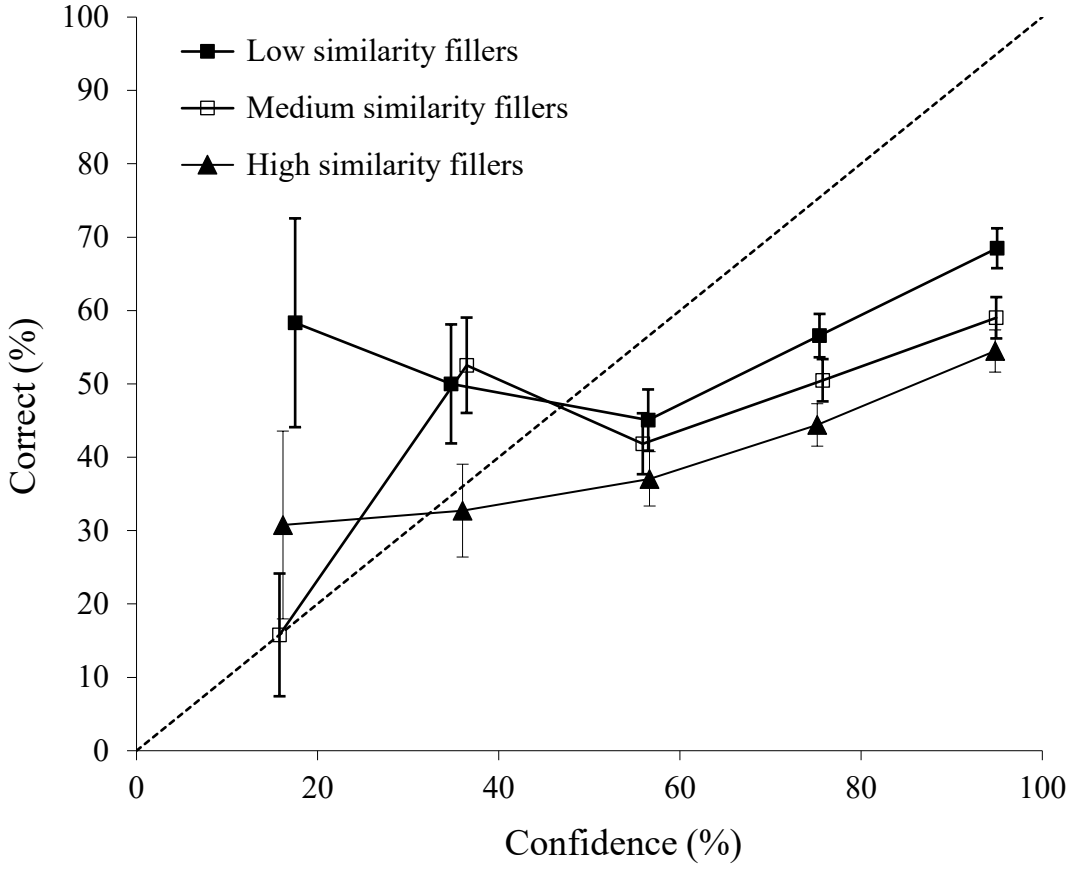

Figure S18. Stimulus 3 data: Calibration curves for choosers in the low, medium and high similarity conditions.

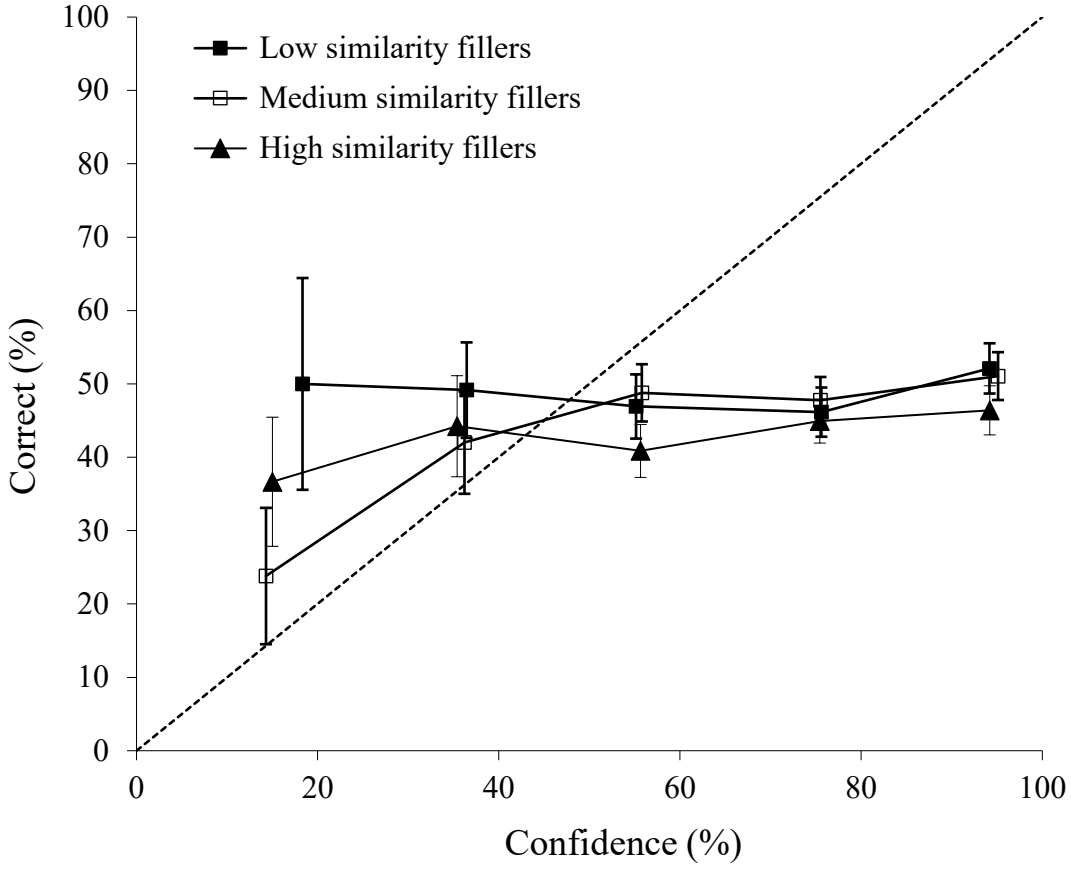

Figure S19. Stimulus 4 data: Calibration curves for choosers in the low, medium and high similarity conditions.

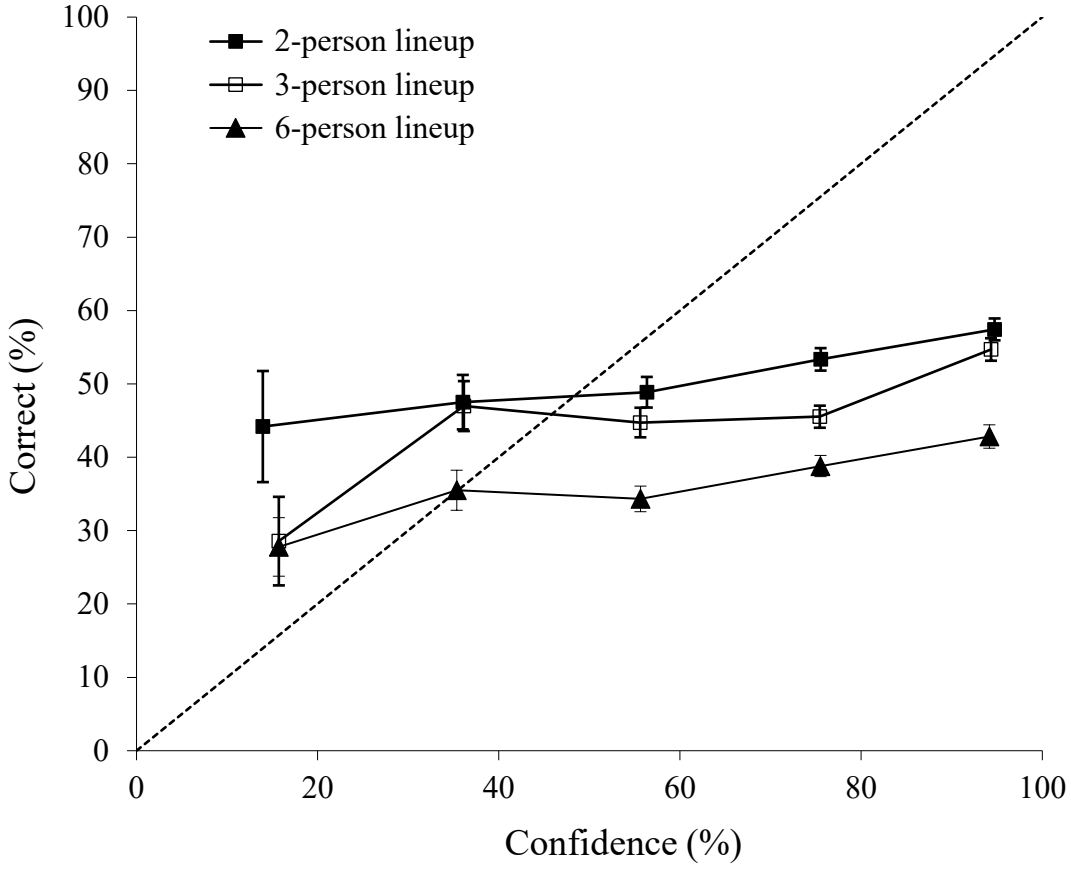

Figure S20. Overall data: Calibration curves for choosers from the 2-person, 3-person and 6-person lineups.

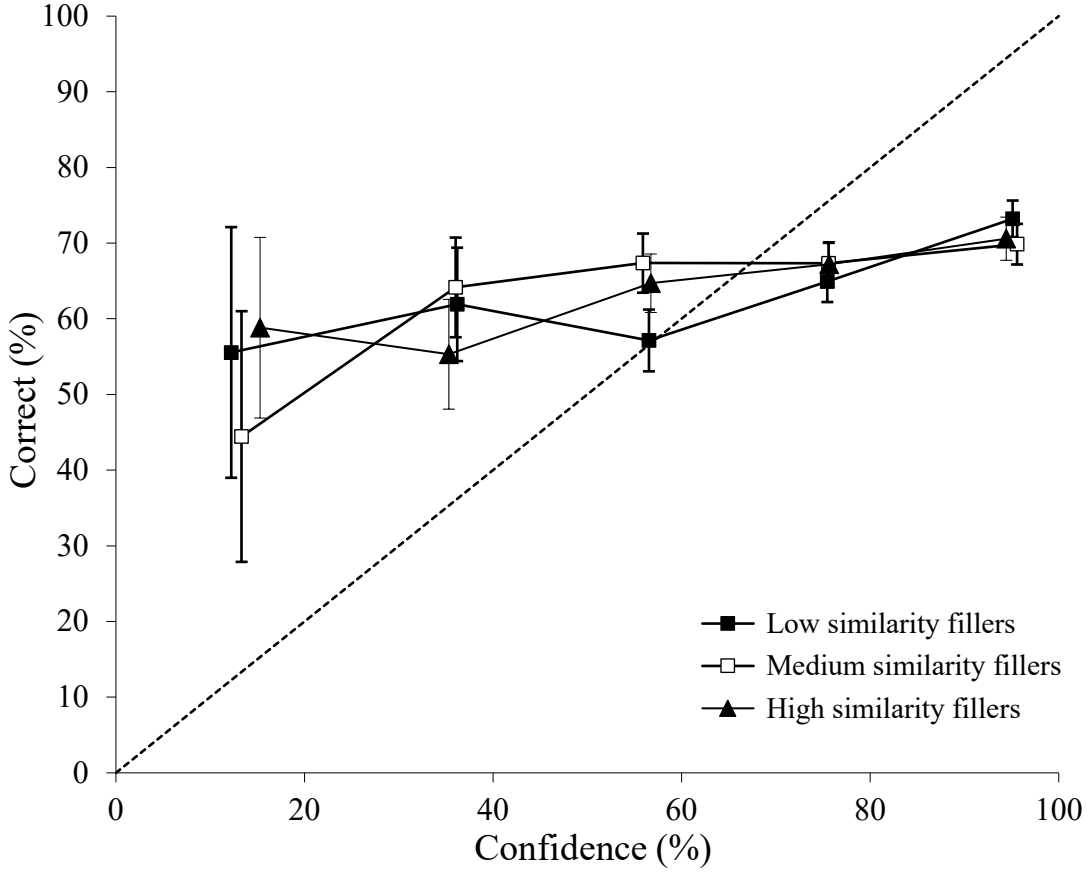

Figure S21. Overall data: CAC curves for suspect choosers from 2-person lineups in the low, medium and high similarity conditions.

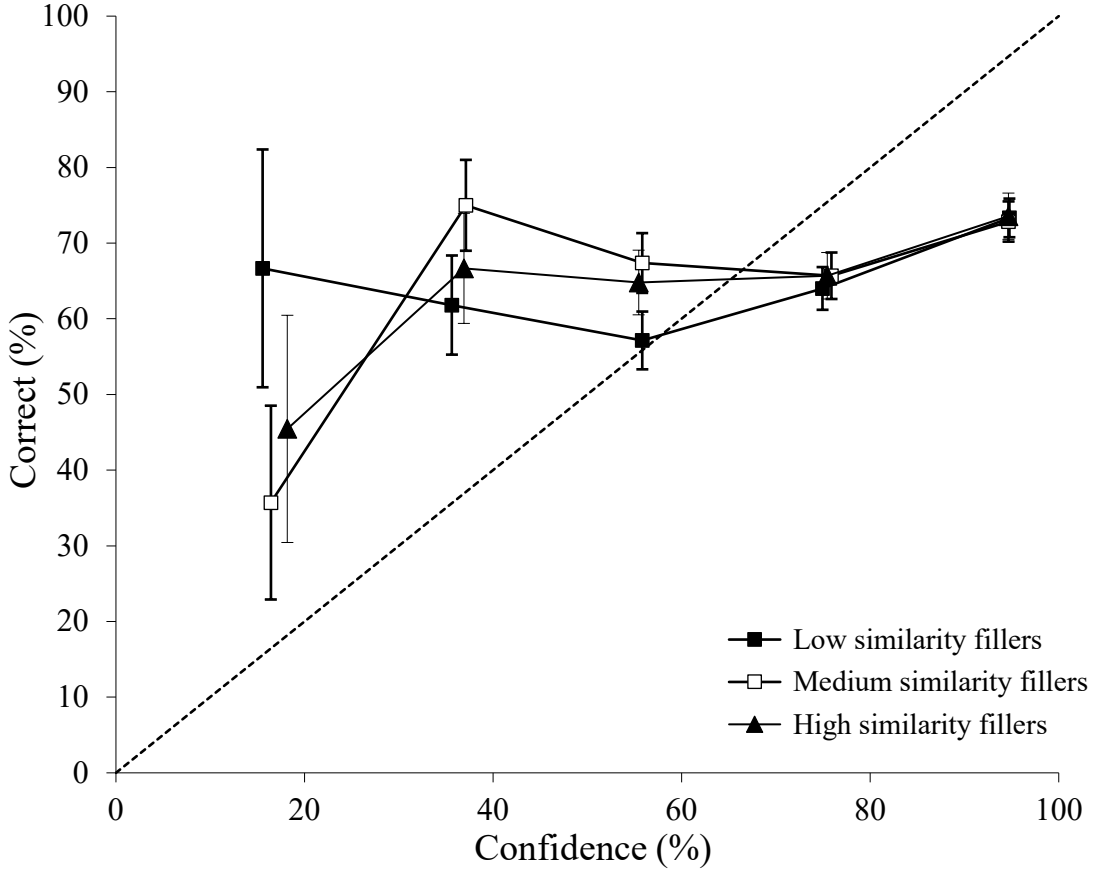

Figure S22. Overall data: CAC curves for suspect choosers from 3-person lineups in the low, medium and high similarity conditions.

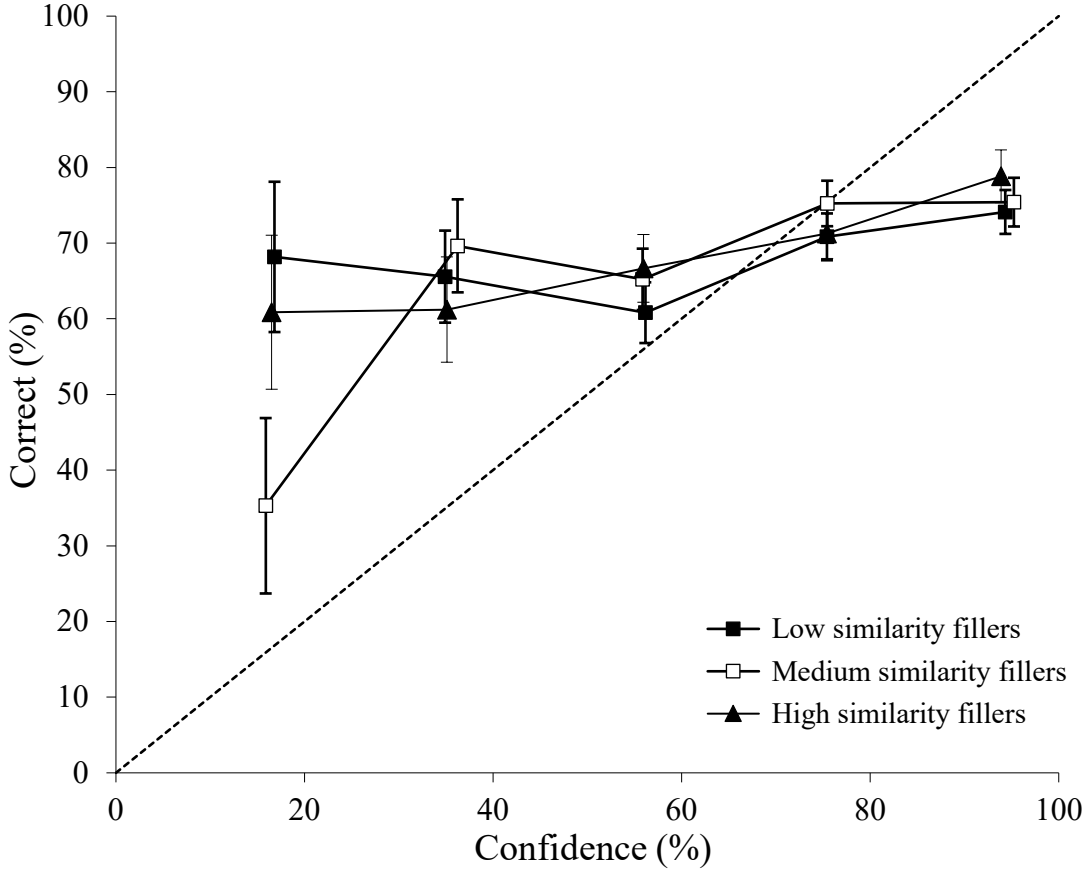

Figure S23. Overall data: CAC curves for suspect choosers from 6-person lineups in the low, medium and high similarity conditions.

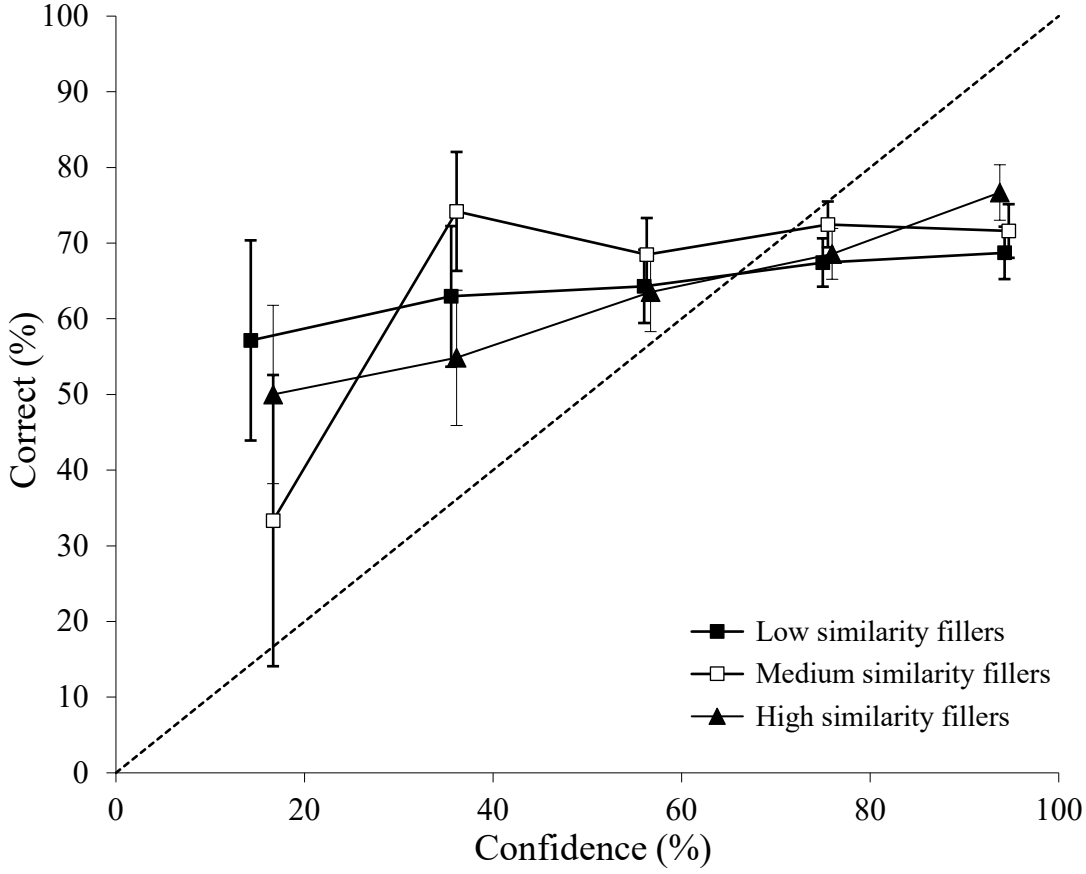

Figure S24. First trial data: CAC curves for suspect choosers in the low, medium and high similarity conditions.

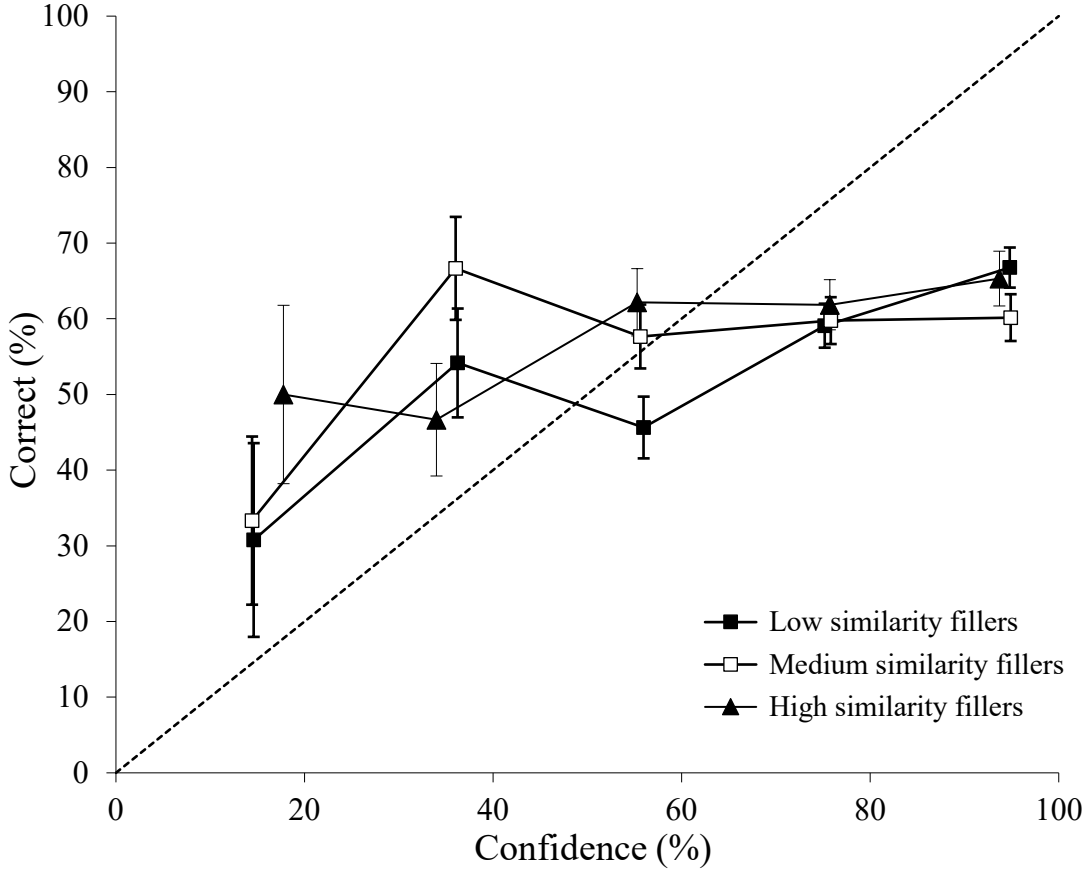

Figure S25. Stimulus 1 data: CAC curves for suspect choosers in the low, medium and high similarity conditions.

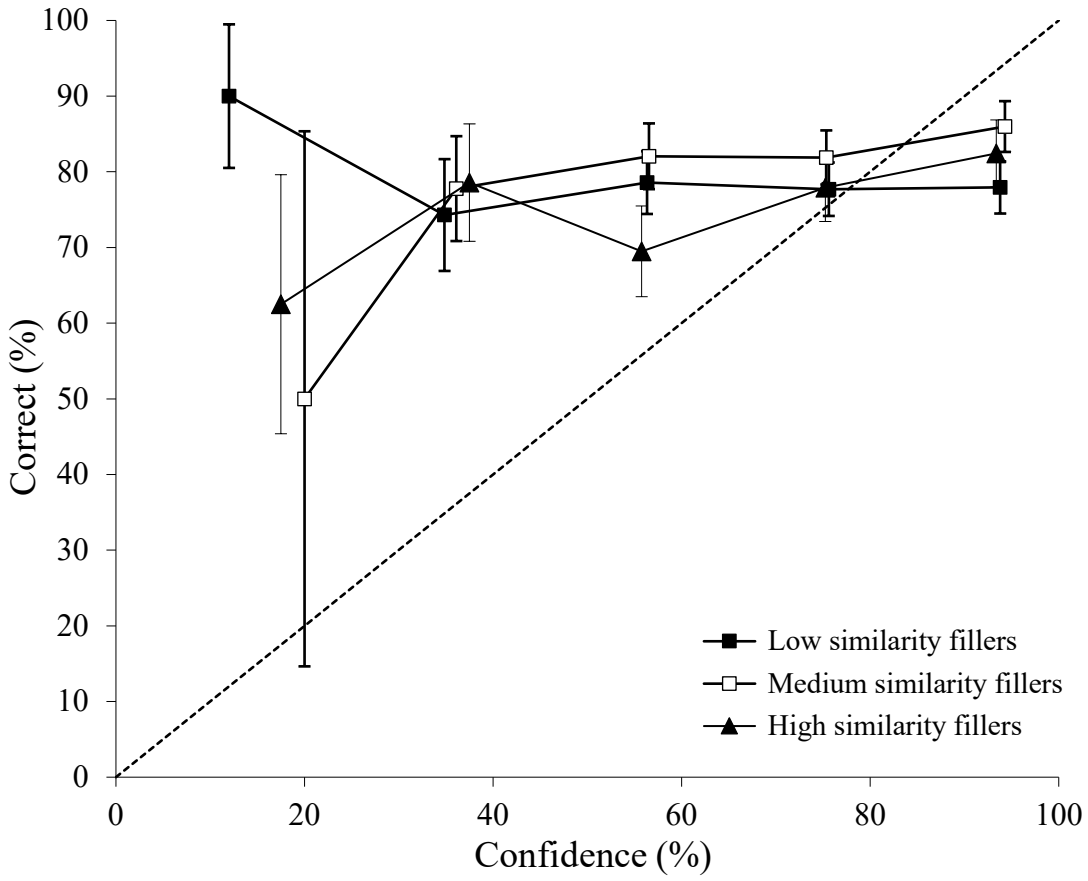

Figure S26. Stimulus 2 data: CAC curves for suspect choosers in the low, medium and high similarity conditions.

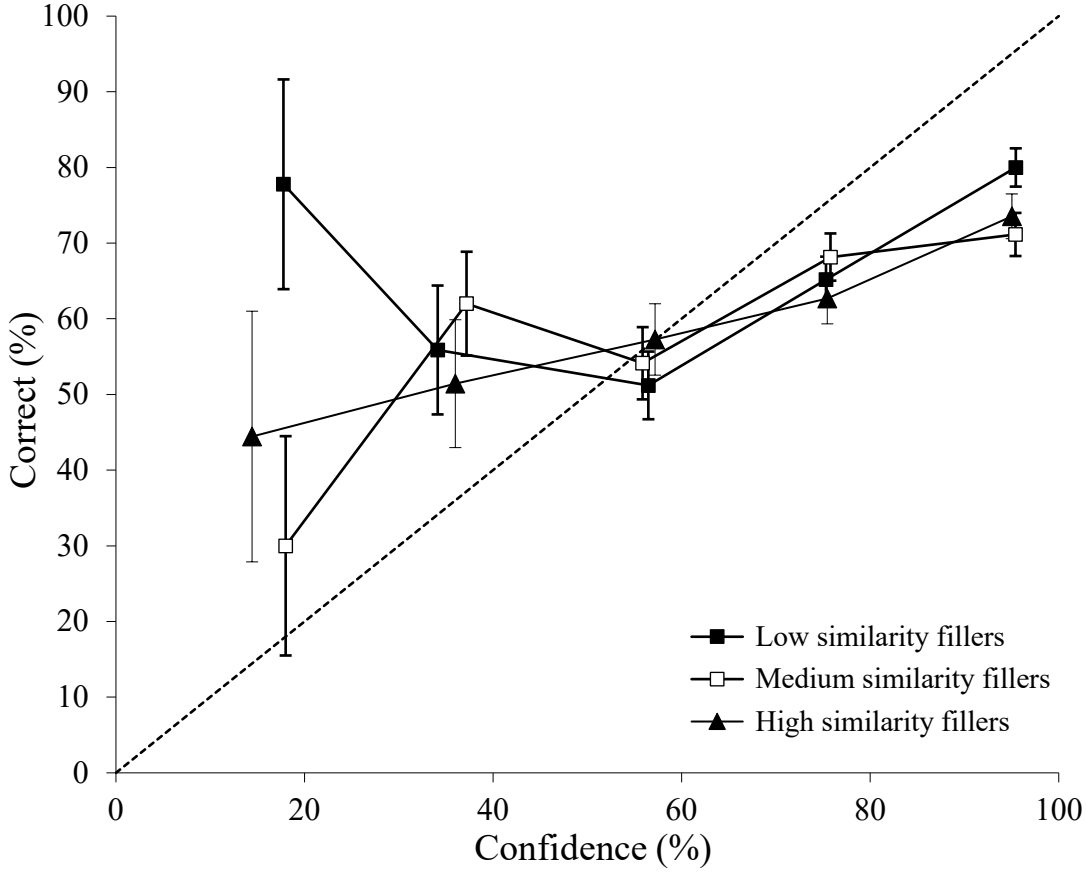

Figure S27. Stimulus 3 data: CAC curves for suspect choosers in the low, medium and high similarity conditions.

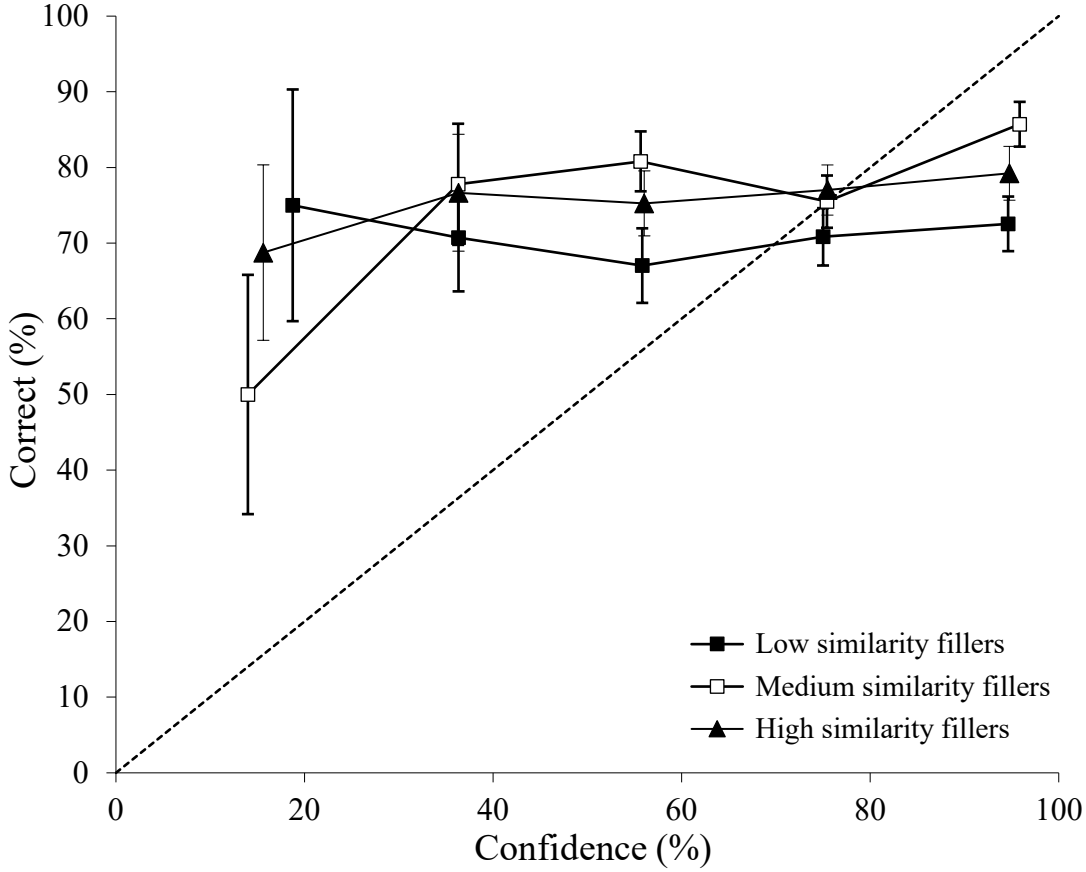

Figure S28. Stimulus 4 data: CAC curves for suspect choosers in the low, medium and high similarity conditions.

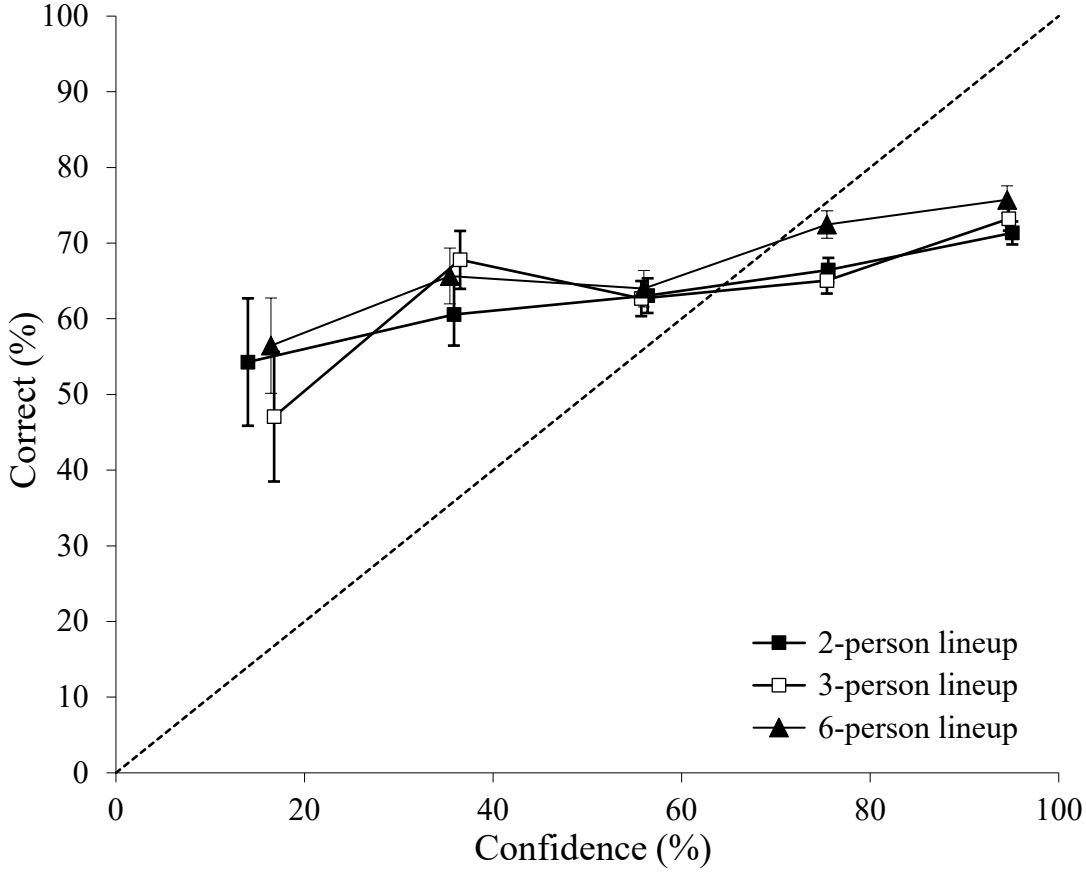

Figure S29. Overall data: CAC curves for choosers from the 2-person, 3-person and 6-person lineups.