Applied Cognitive Psychology, Appl. Cognit. Psychol. **30**: 113–125 (2016)
Published online 12 October 2015 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.3178

Confidence and Eyewitness Identifications: The Cross-Race Effect, Decision Time and Accuracy

CHAD S. DODSON[1]* and DAVID G. DOBOLYI[2]

[1]*Department of Psychology, University of Virginia, Charlottesville, USA*
[2]*McIntire School of Commerce, University of Virginia, Charlottesville, USA*

Summary: Participants encountered same-race and cross-race faces at encoding, completed a series of line-up identification tests and provided confidence ratings by using one of nine different confidence scales. Confidence was less well calibrated with identification accuracy when participants selected a cross-race than a same-race face because of overconfidence. By contrast, there was no cross-race effect on confidence–accuracy calibration when participants responded 'not present'. Whereas confidence was a very strong predictor of accuracy for fast identifications of a line-up face, this was much less the case for slower decisions. Highly confident identifications showed a dramatic drop in accuracy from faster decisions to slower decisions, whereas there was little change in accuracy between faster and slower decisions for moderately confident or weakly confident identifications. Finally, we observed little influence of the format of the nine different confidence scales: numerical and verbal scales produced comparable calibration scores, as did scales with few or many points. Copyright © 2015 John Wiley & Sons, Ltd.

Individuals show less accurate memory for cross-race than same-race faces. This cross-race effect (CRE) is a well-documented phenomenon that occurs across a variety of different kinds of memory tests (Meissner & Brigham, 2001; Sporer, 2001; see Young, Hugenberg, Bernstein & Sacco, 2012, for review). Individuals are worse at recognizing previously encountered cross-race than same-race faces (e.g., Marcon, Susa & Meissner, 2009; Meissner, Brigham & Butz, 2005; Rhodes, Sitzman & Rowland, 2013). They are less able to remember source information about the context in which they encountered cross-race versus same-race faces (e.g., Horry & Wright, 2008; Horry, Wright & Tredoux, 2010). And, of importance for the present paper, individuals are generally less accurate at identifying a cross-race face than a same-race face from a line-up of faces (e.g., Evans, Marcon & Meissner, 2009; Jackiw, Arbuthnott, Pfeifer, Marcon & Meissner, 2008; Smith, Stinson & Prosser, 2004; Wright, Boyd & Tredoux, 2001).

One goal of the present paper is to examine the cross-race effect on (a) confidence ratings that are given to line-up identifications and (b) on the relationship between decision time, confidence and identification accuracy. As a foundation for the study, the following sections will briefly review research on (1) the confidence–accuracy relationship, (2) the interplay between decision time, confidence and accuracy, and (3) ways of measuring confidence.

Confidence–accuracy relationship

The relationship between confidence and accuracy is frequently measured in two different ways, broadly referred to as relative and absolute measures of monitoring (e.g., Koriat & Goldsmith, 1996). In the eyewitness literature, relative measures of the confidence–accuracy relationship have been measured with gamma scores, which refer to how well confidence predicts overall accuracy, such as when individuals give higher confidence ratings to accurate than to inaccurate responses. One limitation, however, with relative measures is that the particular levels of the confidence rating for correct and incorrect responses can have no bearing on the strength of the association between confidence and accuracy. For example, the identical gamma score is obtained when individuals use the lowest confidence rating (e.g., 0) for all incorrect responses and they use either the second lowest confidence rating (e.g., 10) or the highest confidence rating (e.g., 100) for all correct responses. For gamma, the absolute confidence rating does not matter; all that matters is the degree to which lower confidence ratings are assigned to inaccurate than to accurate responses. It is for this reason that these measures of the confidence–accuracy relationship are referred to as relative measures of monitoring because it is the relative difference in confidence for correct and incorrect responses that matters (e.g., Koriat & Goldsmith, 1996). But, as Juslin, Olsson & Winman (1996) argued, because the particular value of the confidence rating is not necessarily important, these relative scores of the confidence–accuracy relationship can have little practical value. Because police do not know *a priori* whether an identification is correct or incorrect, they must focus on the absolute level of confidence of the eyewitness.

Absolute measures of the confidence–accuracy relationship, such as calibration and over–underconfidence scores, focus on the particular level of the confidence rating that is assigned to correct and incorrect responses (e.g., Juslin *et al.*, 1996; Koriat & Goldsmith, 1996). Perfect calibration is shown when linear increases in confidence are paralleled by linear increases in accuracy, such as when individuals are poorly, moderately and highly accurate for responses that are assigned low, moderate and high levels of confidence, respectively. Knowing that a condition is associated with excellent calibration is useful knowledge for investigators because it indicates that high confidence is associated with high accuracy and likewise for low confidence and low accuracy. By contrast, it is impossible to make this inference about high confidence and high accuracy from even an excellent gamma score because it is possible that individuals assigned a high confidence rating to both correct and

* Correspondence to: Chad Dodson, Department of Psychology, University of Virginia, Charlottesville, VA 22904-4400, USA.
E-mail: cdodson@virginia.edu

incorrect responses [e.g., a confidence rating of 100 to all correct and a rating of 90 (or even 99) to all incorrect responses would produce a perfect gamma score].

There are factors in line-up identification studies that reliably affect the confidence–accuracy relationship. Much research shows that when participants choose a face from a line-up—so-called choosers—they generally show a reasonable relative relationship between confidence and accuracy —e.g., Sporer *et al.* (1995) observed an $r_{pb} = .37$ and Sauerland and Sporer (2009) observed a similar correlation score for choosers. Likewise, absolute measures of the confidence–accuracy relationship, such as calibration scores, also show a reasonable relationship for choosers (e.g., Weber & Brewer, 2003; Weber, Brewer, Wells, Semmler and Keast, 2004; Brewer & Wells, 2006; Sauerland & Sporer, 2009). For example, Sauerland and Sporer (2009) observed that accuracy increased from nearly 20% to nearly 80% with increases in confidence from the least confident (i.e., those expressing 0–20% confidence) to the most confident (i.e., those expressing 80–100% confidence). By contrast, the confidence–accuracy relationship for nonchoosers (i.e., those who respond 'not present') is much worse than that for choosers. There frequently is no relationship at all between the confidence in and accuracy of a nonchooser response (i.e., confidence does not predict accuracy; e.g., Sauerland & Sporer, 2009; see Sporer *et al.*, 1995, for a meta-analysis and similar findings). Similarly, nonchoosers tend to show very poor calibration scores because they exhibit similar levels of accuracy across all levels of confidence (e.g., Brewer & Wells, 2006). Overall, many studies have observed that confidence is both reasonably correlated and calibrated with chooser accuracy, but this is not the case for nonchooser accuracy.

Is there a cross-race effect on the relationship between confidence and accuracy? Many studies show that individuals give higher confidence ratings for correct recognition responses to same-race than cross-race faces and that they give lower confidence ratings to incorrect recognition responses to same-race than cross-race faces (e.g., Horry & Wright, 2008; Rhodes *et al.*, 2013). This pattern for confidence to correct and incorrect responses means that confidence–accuracy resolution (i.e., a relative measure) is better for same-race than for cross-race faces, but it is impossible to infer anything from these data about confidence–accuracy calibration (i.e., an absolute measure). Further support for a CRE on relative measures of the confidence–accuracy relationship comes from both recognition and line-up identification studies that have observed stronger gamma scores for same-race than for cross-race faces (e.g., Corenblum & Meissner, 2006; Wright, Boyd & Tredoux, 2001, 2003). These data show that confidence better predicts the accuracy of responses to same-race than cross-race faces. Notably, Wright *et al.* (2001, 2003) are the only researchers to use a line-up identification paradigm and observed that the relative relationship between confidence and accuracy (e.g., gamma scores) is no different from zero for cross-race identifications.

Is there a CRE on either confidence–accuracy calibration or other absolute measures of monitoring (e.g., over–underconfidence)? Are individuals prone to excessive overconfidence when making cross-race line-up identifications? The answers to these questions are unknown as no one has examined them. But, given the value of this knowledge to investigators, the answers to these questions are of great importance.

Decision time, confidence and accuracy

Much research shows that there is a reliable relationship between decision time and the accuracy of positive identifications from lineups (i.e., choosers; e.g., Brewer & Wells, 2006; Dunning & Perretta, 2002; Sauerland & Sporer, 2007, 2009; Smith, Lindsay, Pryke & Dysart, 2001; Sporer, 1993, 1994; Weber *et al*., 2004). Faster decisions are generally more accurate than slower decisions. For example, Sauerland and Sporer (2009) observed that faster line-up decisions (i.e., 6 seconds or less) showed 72% accuracy, but accuracy dropped in half (36% accuracy) for slower decisions (i.e., >6 seconds). One drawback for real world application, however, is that the optimal decision time for an identification varies from study to study and is related to retention interval and other factors (e.g., Brewer, Caon, Todd & Weber, 2006; Sauerland & Sporer, 2007; Weber *et al.*, 2004). So, there does not appear to exist a single time interval that one can point to as always indicating a fast or slow decision. In addition, this relationship between decision time and accuracy appears only to apply to choosers, as many studies have observed that decision time is not meaningfully associated with the accuracy of 'not present' responses to lineups (e.g., Sauerland & Sporer, 2009).

Combining confidence and decision time has proven particularly powerful for predicting accuracy of positive identifications from lineups (e.g., Brewer & Wells, 2006; Sauerland & Sporer, 2007, 2009; Weber *et al.*, 2004). For example, Sauerland and Sporer (2009) observed that fast (6 seconds or less) and confident (90–100%) individuals showed an impressive 97% accuracy rate when they selected someone from a line-up. But, nonchoosers show a very different pattern from choosers: the combination of confidence and decision time provided little value at distinguishing between correct and incorrect nonchooser responses (e.g., Brewer & Wells, 2006).

What of the cross-race effect and confidence and decision time for line-up identifications? Smith *et al.* (2001) were the only researchers to examine this question, and their study is problematic because they did not observe a CRE on line-up identification performance; they found no significant differences in accuracy between own-race choosers and cross-race choosers. Thus, it is still an open question about the relationship between confidence and decision time and the accuracy of cross-race decisions.

Measuring confidence

Does the format of the confidence scale influence the relationship between confidence and accuracy? When measuring the confidence–accuracy relationship, does it matter if individuals use numbers or words or if the scale contains few or many points? Existing research by Weber, Brewer and Margitich (2008) within the context of an eyewitness identification task and by Wallsten and Budescu (1983) within the context of a probability judgment task shows nearly identical

patterns of calibration of confidence to accuracy when participants use either numeric or verbal confidence scales (see also Wallsten, Budescu & Zwick, 1993; see Wallsten & Budescu, 1995, for review). For example, Wallsten *et al.* (1993) required participants to judge the likely accuracy of general knowledge statements with either numeric or verbal confidence scales and observed nearly identical calibration scores for both confidence formats. Despite these findings, many eyewitness researchers assume that calibration scores can be computed properly only when (a) a numerical scale is used and (b) in the context of identifications from a typical line-up that the confidence scale must range from 0% to 100% (e.g., Wells & Penrod, 2011). We examine these latter two assumptions in our study.

We used nine different confidence scales to examine whether the format of the confidence scale matters when measuring the relationship between confidence and accuracy. First, we used a variety of different verbal and numeric scales that allow us to replicate the findings of Weber *et al.* (2008) and Wallsten and colleagues (e.g., Wallsten *et al.*, 1993) about a similar confidence–accuracy relationship with verbal and numeric scales. Second, we varied the number of points on the scale (6 points vs. 11 points vs. 101 points) to examine the hypothesis that more points will increase the sensitivity of the scale. Third, all scales included identical verbal anchors at the endpoints (i.e., not at all confident and completely confident). But, we included numeric scales that ranged either from 0% to 100% or from 50% to 100%. Typically, calibration studies use numeric scales that correspond to chance performance, such as using a 50% to 100% confidence scale for a two alternative forced choice task because chance accuracy is 50%. We deliberately used different numeric ranges in order to examine the assumption that assessing calibration with a typical line-up requires that the confidence scale ranges from 0% to 100%. Put another way, we examine how flexible individuals are at mapping a sense of confidence onto a confidence scale that substantially varies in range (i.e., 0–100% vs. 50–100%), as well as other dimensions, such as verbal versus numeric format. Finally, for the verbal confidence scales, we manipulated whether every point on the verbal scale was labeled or only the endpoints on the scale were labeled. In other words, one would expect that when given a 6-point or an 11-point scale—ranging from 'not at all confident' to 'completely confident'—that individuals should be better calibrated at mapping their subjective sense of confidence on to a verbal scale that has a label at every point, as opposed to a scale that consists of unlabeled points except for the anchor labels at the endpoints. Overall, these nine different confidence scales allow us to examine the degree to which confidence–accuracy relationships (e.g., calibration) are tied to the particular format of the confidence scale (see the Procedure section for a list and example of each of the scales).

Overview

This study answers a number of questions about the cross-race effect on line-up identification accuracy and the interplay between the CRE, accuracy, confidence and decision time. We used a line-up recognition paradigm (e.g., Meissner, Tredoux, Parker & MacLin, 2005; Dobolyi & Dodson, 2013) in which participants first encoded a series of same-race and cross-race faces and then encountered a series of lineups consisting of a mixture of lineups that either contained or did not contain a different photo of a previously seen person (i.e., target-present lineups and target-absent lineups, respectively). Participants provided a confidence rating for all responses, which allowed us to examine the relationship between confidence and accuracy.

METHOD

Participants

Participants were 1656 individuals who completed the task over the Internet via Amazon's Mechanical Turk (www.mturk.com). Individuals were restricted to US users, and a check of IP addresses showed that 96.92% of participants were located in the USA, with all states—except Wyoming—represented by the sample. There were 1482 Caucasian-Americans (mean age = 23.74 years, $SD = 3.13$, range = 18–30, 49.83% female)[1] and 174 African-Americans (mean age = 23.54 years, $SD = 3.29$, range = 18–30, 64.37% female).

Design

We used a 9 (confidence-scale format) × 2 (line-up race: black, white) × 2 (participant race: black, white) × 2 (target-present vs. target-absent lineups) mixed factorial design, with confidence-scale format and participant race as between-subjects factors and line-up race and target presence/absence as within-subjects factors (see later for how we implement this design by using 12 lineups). *A priori* power analyses using G*POWER (Faul, Erdfelder, Lang & Buchner, 2007) showed that our sample size and an alpha level of .05 provide us with over 90% power to detect small-sized effects—using Cohen's (1988) criteria—that are between factors (e.g., confidence-scale format) in the ANOVA and we will have over 99% power to detect small-sized effects that are within factors.

Materials

We used the materials, procedure and browser-based framework that were used in Dobolyi and Dodson (2013). There were 12 target faces (six black and six white), with a casual version (i.e., smiling facial expression and wearing street clothes) and a formal version (i.e., neutral facial expression and wearing a maroon t-shirt) of each face. The photos of the faces were from a noncommercial database (the Meissner Face Database). The casual photo of the person was shown at encoding, whereas the line-up contained the formal version of the person. Each target face was associated with two six-person lineups: (a) a target-present line-up consisted of the target face along with five foil faces and (b) a target-absent line-up consisted of six foils (i.e., aforementioned five foils and an additional foil that replaced the target face). A mock witness paradigm verified that all 12 lineups were fair: that is, using the scoring method of Tredoux (1999), we

[1] Technical error prevented the recording of the age of one white participant.

obtained average *E*-scores across all lineups of 4.66 (95% CI = 4.26, 5.07) and an average target selection rate of 18% (95% CI = 10%, 26%). See Dobolyi and Dodson (2013) for details about the materials.

Procedure

Participants were instructed to pay careful attention to a series of faces as their memory for them would be tested later. They further were informed that the faces would appear one at a time and that some of them may repeat.

During the encoding phase, a casual version of each face was shown for 3 seconds with a 1-second interstimulus interval. Each of the 12 target faces was presented twice in a random order, with the following three constraints: (a) that all 12 target faces appeared once before any repeated; (b) no identical target face appeared consecutively; and (c) that no more than three faces of a particular race appeared back-to-back. To counteract primacy and recency effects, the first two faces and last two faces of the encoding phase were filler faces that never appeared during the test phase. Overall, then, the encoding phase contained 28 trials: the 12 target faces seen twice, plus the four filler faces. When the encoding phase was finished, there was a 5-minute distraction task.

Participants were informed that the test phase consisted of a series of six-person lineups, with some lineups containing a previously seen person and in other lineups all six faces were never seen before. Specifically, they were told, 'You will now go through a series of lineups in which your goal is to determine whether or not one of the people you saw earlier is present in each lineup. Lineups will consist of six faces shown together. In each lineup, only *one* face may correspond to someone you saw earlier, but be aware that the photo will not be identical. It is possible for all six people in a lineup to be ones you have not seen earlier. Either way, focus on just the faces: all lineup faces will be shown with identical clothing and a neutral facial expression. If you recognize a person in a lineup, click on that face to highlight it. If you do not recognize any of the people, click on the "Not Present" option.' Each line-up consisted of a two-row by three-column grid of faces with a 'Not Present' option centered underneath it.

Upon making a response, a confidence scale appeared underneath the line-up, and participants provided a confidence rating about the likely accuracy of their response. Participants encountered only one of the nine different confidence scales (described later). Each scale was presented as a horizontal line with the anchors 'Not at All Confident' and 'Completely Confident' at the left and right endpoints, respectively. Each scale consisted of either six or 11 discrete points that the participant could select for their confidence rating, in addition to a numeric slider scale that contained 101 points. The different scales were as follows:

1. Numeric, 6 points, 50–100% range: 50, 60, 70, 80, 90, 100
2. Numeric, 6 points, 0–100% range: 0, 20, 40, 60, 80, 100
3. Numeric, 11 points, 50–100% range: 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100
4. Numeric, 11 points, 0–100% range: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
5. Numeric, continuous (via a slider), 0–100% range
6. Verbal, 6 points: Not at All Confident, Quite Unconfident, Somewhat Unconfident, Somewhat Confident, Quite Confident, Completely Confident
7. Verbal, 6 points, but only endpoints are labeled as Not at All Confident and Completely Confident with no labels for the intervening points
8. Verbal, 11 points: Not at All Confident, Extremely Unconfident, Quite Unconfident, Rather Unconfident, Somewhat Unconfident, As Confident as Unconfident, Somewhat Confident, Rather Confident, Quite Confident, Extremely Confident, Completely Confident
9. Verbal, 11 points, but only endpoints are labeled as Not at All Confident and Completely Confident with no labels for the intervening points

Finally, each participant encountered 12 lineups during the test phase, one for each of the target faces seen at encoding. Three white lineups and three black lineups were randomly assigned as target-present lineups (i.e., contained a previously seen face), and the rest were target-absent lineups (i.e., did not contain a previously seen face). Presentation of the lineups was random with the constraint that there were (a) no more than three consecutive target-present or target-absent lineups and (b) no more than three consecutive black or white lineups.

RESULTS

Effect sizes are measured with Cohen's (1988) *d* for *t*-tests and partial eta-squared (i.e., η_p^2) for ANOVAs.

Accuracy and response bias

We assessed accuracy with d' scores that were based on the correct identification rate to target-present lineups and the false alarm rate from target-absent lineups.[2] The advantage of d' over other measures of accuracy, such as diagnosticity ratios or the correct identification rate on target-present lineups, is that d' is less susceptible than other measures to the confounding influence of changes in decision criteria (e.g., the tendency to select any face in a line-up or conversely the tendency to respond 'not present'). A 2 (participant race: black, white) × 2 (line-up race: black, white) × 9 (confidence scale type) ANOVA revealed a significant interaction between race of participant and race of line-up faces, $F(1, 1638) = 32.07$, $MSE = 1.07$, $p < .001$, $\eta_p^2 = .02$, 95% CI [0.01, 0.03]. There is a significant cross-race effect for both groups of participants: black participants showed significantly higher d' scores for black lineups ($M = 2.04$, $SD = 1.36$) than for white lineups ($M = 1.51$, $SD = 1.39$), $t(173) = 4.75$, $p < .0001$, Cohen's $d = 0.38$, 95% CI [0.22, 0.54], whereas white participants showed significantly higher d' scores for white lineups ($M = 1.61$, $SD = 1.27$) than for black lineups ($M = 1.48$,

[2] To compute d', we divided the false alarm rate by six to compensate for the greater number of foils in the TA line-up as compared to the single target in the TP line-up (e.g., Meissner *et al.*, 2005). Moreover, because values of d' are undefined when hit rates or false alarm rates are equal to zero, all hit rates and false alarm rates were transformed by adding .1 to the numerator and .2 to the denominator (see Dobolyi & Dodson, 2013, for details).

$SD = 1.24$), $t(1481) = 3.46$, $p = .0006$, Cohen's $d = 0.10$, 95% CI [0.03, 0.18]. This interaction qualifies the significant main effects of participant race, $F(1, 1638) = 7.07$, $MSE = 2.15$, $p < .01$, $\eta_p^2 = .01$, 95% CI [0.00, 0.01] and line-up race, $F(1, 1638) = 11.69$, $MSE = 1.07$, $p < .001$, $\eta_p^2 = .01$, 95% CI [0.00, 0.02]. There were no other significant effects in this analysis; all Fs < 1.30. Overall, then, we replicated the same-race versus cross-race effect on identification accuracy and critically showed that this CRE occurred for both black participants and white participants. In addition, we observed no statistically significant differences in accuracy across the different confidence rating scale formats.

As for response bias, we examined the signal detection score, C, in which higher values correspond to a bias to respond 'not present' and lower values correspond to a bias to choose a face. A 2 (participant race: black, white) × 2 (line-up race: black, white) × 9 (confidence scale type) ANOVA revealed an effect of line-up race, $F(1, 1638) = 4.06$, $MSE = 0.24$, $p = .04$, $\eta_p^2 = .00$, 95% CI [0.00, 0.01] and no other significant effects, all Fs < 1.65. Participants were significantly more biased to respond 'not present' to white lineups ($M = 0.86$, $SD = 0.57$) than to black lineups ($M = 0.78$, $SD = 0.55$), but there was no cross-race effect on response bias. Overall, the presence of a cross-race effect on identification accuracy and the absence of one on response bias are consistent with Jackiw *et al.* (2008).

Confidence and accuracy

Absolute measure of confidence and accuracy: Calibration

We assessed the calibration of confidence to accuracy with calibration error scores, which measure how well confidence ratings align with accuracy. For instance, excellent calibration between accuracy and confidence is shown when individuals exhibit perfect, moderate and chance performance for responses that are assigned, respectively, high (e.g., 100), medium (e.g., 60) and low (e.g., 0) confidence ratings. In this way, the absolute value of the confidence ratings corresponds to the absolute level of performance for items that are assigned these ratings. Calibration error scores were derived for each individual by taking the absolute difference between predicted accuracy, as indicated by the confidence rating, and actual accuracy at each level of confidence (see Koriat & Goldsmith, 1996). This difference is then further weighted by the frequency of responses at each level of confidence.[3]

[3] We transformed the data from all scales to a 6-point scale that ranged from 0% to 100% so as to remove a potential confound of computing scores based on different-sized scales. But, when participants chose an individual from a line-up (i.e., choosers), we obtained the identical pattern of effects for all measures when the analyses were based on either the raw (untransformed) or transformed data. Similarly, for nonchoosers (i.e., when participants responded 'not present' to a line-up), the raw data and the transformed data produced the identical pattern of effects for Somers' D scores, although there were slight differences for calibration and over/underconfidence scores. Specifically, for calibration scores, both the raw and transformed data produced effects of confidence scale type, participant race and line-up race, except these latter two effects were marginally significant when the data from all the scales had been transformed to 6 points. For over/underconfidence scores, both the raw and transformed data showed an effect of participant race, but the transformed data showed an additional effect of confidence format (see the text for discussion of this effect). Overall, there is a consistent pattern of results from both the raw and transformed data; our particular transformation did not produce the particular pattern of results.

Figure 1 shows a calibration plot for choosers (Figure 1a) and nonchoosers (Figure 1b) to same-race and cross-race lineups, collapsed across the different confidence-scale conditions. Chooser accuracy refers to the likelihood of making a correct identification given that a face was chosen from a line-up, and following others (e.g., Brewer & Wells, 2006), it is computed by the following formula: (correct identification$^{\text{target present}}$)/(correct identification$^{\text{target present}}$ + false identification$^{\text{target absent}}$). By contrast, nonchooser accuracy refers to the likelihood of correctly responding 'not present' when an individual responded 'not present' to a line-up and is computed with the following formula: (correct rejection$^{\text{target absent}}$)/(correct rejection$^{\text{target absent}}$ + miss$^{\text{target present}}$). In both figures, the black diagonal line shows perfect calibration in which the lowest point of the confidence scale is associated with chance performance, the highest point of the scale is associated with perfect performance and the points in between are associated with corresponding degrees of accuracy. Greater deviation from the black diagonal line corresponds to less accurate calibration.

Figure 1 makes two points clear: (1) although there are significant deviations from the diagonal, confidence and accuracy are reasonably calibrated for choosers. For nonchoosers, however, the flat lines show that confidence has very little correspondence with accuracy; and (2) although the lines for same-race and cross-race performance are similar, cross-race identification for choosers is clearly less calibrated (i.e., farther from the diagonal than the same-race calibration curve).

Table 1 presents the calibration scores from each confidence-scale condition for choosers and nonchoosers to same-race and different-race lineups. A 2 (participant race: black, white) × 2 (line-up race: black, white) × 9 (confidence scale type) ANOVA of the calibration error scores when individuals choose a face showed a significant interaction between line-up race and race of participant, $F(1, 1600) = 21.26$, $MSE = 0.029$, $p < .0001$, $\eta_p^2 = .01$, 95% CI [0.01, 0.02]. Black participants were less calibrated with white lineups ($M = 0.37$, $SD = 0.22$) than black lineups ($M = 0.30$, $SD = 0.24$), $t(168) = 3.50$, $p < .001$, $d = 0.31$, 95% CI [0.13, 0.49], whereas white participants were less calibrated with black lineups ($M = 0.38$, $SD = 0.20$) than white lineups ($M = 0.36$, $SD = 0.20$), $t(1448) = 3.26$, $p < .01$, $d = 0.10$, 95% CI [0.03, 0.17]. This interaction between line-up race and participant race qualifies the significant main effects of participant race, $F(1, 1600) = 6.95$, $MSE = 0.051$, $p < .01$, $\eta_p^2 = .00$, 95% CI [0.00, 0.01], and line-up race, $F(1, 1600) = 6.42$, $MSE = 0.029$, $p < .01$, $\eta_p^2 = .00$, 95% CI [0.00, 0.01]. There were no other significant effects from this ANOVA, all Fs < 1.57. Overall, then, when individuals chose a face from a line-up, they were better calibrated when making judgments about same-race than different-race lineups. Moreover, as seen in Table 1, there were no statistically significant differences between the nine different confidence rating conditions in the calibration of confidence to chooser accuracy.

With respect to the calibration of confidence to accuracy when individuals do not choose a face, a 2 (participant race: black, white) × 2 (line-up race: black, white) × 9 (confidence scale type) ANOVA produced an effect of confidence scale type, $F(8, 1407) = 2.89$, $MSE = 0.044$, $p < .01$, $\eta_p^2 = .02$, 95% CI [0.00, 0.02], and no other significant effects, all

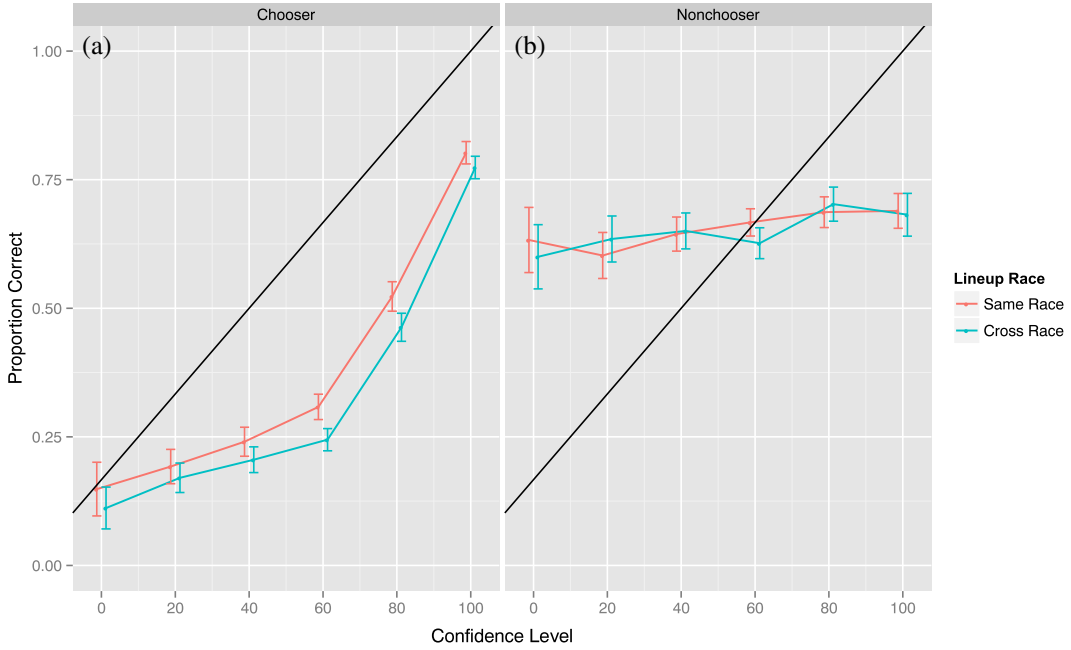


Figure 1. Calibration of confidence-level to accuracy for same-race and cross-race lineups when participants either choose a face from the lineup (i.e., Chooser lineup) or respond "not present" (i.e., NonChooser lineup). Error bars represent 95% confidence intervals.

Fs < 3.23. Student–Newman–Keuls tests showed that when individuals used the 50–100, 6-point confidence scale, they were significantly less calibrated than when individuals used any of the other scales, except for the 50–100, 11-point scale. Similarly, the 50–100, 11-point scale was associated with significantly worse calibration than the 0–100, 11-point scale, the labeled, verbal 6-point scale, or the labeled, verbal 11-point scale.

Absolute measure of confidence and accuracy: Over/underconfidence

Overconfidence and underconfidence are shown in Figure 1 by points on the calibration curve that fall below and above the diagonal, respectively. Nearly all points of the same-race and cross-race chooser-calibration curves fall below the diagonal and thus reflect overconfidence. For example, consider chooser accuracy when participants use the top-most confidence rating (e.g., 100%). Instead of showing 100% accuracy at this top-most rating, accuracy is much worse. In other words, participants believe that they are performing better than they actually are. The other key pattern is that for choosers, the cross-race curve reflects more overconfidence than the same-race curve. The formula for computing over/underconfidence (OU) scores is nearly identical to that for computing calibration. However, OU scores are based on the difference between predicted accuracy, as indicated by the confidence rating, and actual accuracy at each level of confidence—in contrast to calibration that is based on the absolute value of this difference. All other aspects of deriving OU scores are identical to those for calibration. Positive OU scores indicate overconfidence (i.e., the confidence rating is greater than the corresponding level of accuracy), and negative OU scores reflect underconfidence.

A 2 (participant race: black, white) × 2 (line-up race: black, white) × 9 (confidence scale type) ANOVA of OU scores when individuals choose a face yielded a significant interaction between participant race and line-up race, $F(1, 1600) = 20.93$, $MSE = 0.05$, $p < .0001$, $\eta_p^2 = .01$, 95% CI [0.01, 0.02]. As seen in Table 1, white participants were significantly more overconfident when choosing a face from a black line-up ($M = 0.28$, $SD = 0.27$) than from a white line-up ($M = 0.22$, $SD = 0.28$), $t(1448) = 6.28$, $p < .001$, $d = 0.22$, 95% CI [0.15, 0.29], whereas black participants showed the opposite pattern and were significantly more overconfident when choosing a face from a white line-up ($M = 0.27$, $SD = 0.28$) than from a black line-up ($M = 0.21$, $SD = 0.29$), $t(168) = 2.58$, $p = .01$, $d = 0.22$, 95% CI [0.05, 0.40]. There was also a main effect of scale type, $F(8, 1600) = 3.09$, $MSE = 0.10$, $p < .01$, $\eta_p^2 = .02$, 95% CI [0.00, 0.02], and no other significant effects in this analysis, all Fs < 1.20. As presented in Table 1, Student–Newman–Keuls tests confirmed that individuals who used the 50–100, 6-point confidence scale were significantly less overconfident than individuals who used any of the other scales. The remaining scales did not differ from each other.

Finally, when individuals responded 'not present' and did not choose a face, a 2 (participant race: black, white) × 2 (line-up race: black, white) × 9 (confidence scale type) ANOVA of OU scores showed a significant effect of participant race, $F(1, 1407) = 9.07$, $MSE = 0.15$, $p < .01$, $\eta_p^2 = .01$, 95% CI [0.00, 0.02]. White participants ($M = -0.04$, $SD = 0.35$) were more underconfident than black participants ($M = 0.03$, $SD = 0.33$), but both groups of participants were close to the optimal value of 0, which represents the absence of overconfidence or underconfidence. There was also a main effect of confidence scale type, $F(8, 1407) = 2.49$, $MSE = 0.15$, $p = .01$, $\eta_p^2 = .01$, 95% CI [0.00, 0.02], and no other significant effects in this analysis, all Fs < 1.96. Student–Newman–Keuls tests showed that individuals who used the 50–100, 6-point confidence scale were significantly more underconfident than individuals who used any of the

Table 1. Calibration and over–underconfidence scores as a function of participant race, line-up race and confidence scale format

					Chooser				Nonchooser			
		Confidence scale format			Calibration		OU		Calibration		OU	
Participant race	Line-up race	Numeric/Verbal	Range	Points	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
White	Black	Numeric	50–100	6	0.35	0.01	0.19	0.02	0.42	0.02	−0.12	0.03
	Black	Numeric	0–100	6	0.38	0.02	0.29	0.02	0.38	0.02	−0.08	0.03
	Black	Numeric	50–100	11	0.39	0.02	0.26	0.02	0.42	0.02	−0.10	0.03
	Black	Numeric	0–100	11	0.38	0.02	0.29	0.02	0.36	0.02	−0.02	0.03
	Black	Numeric	0–100	101	0.37	0.02	0.27	0.02	0.38	0.02	−0.04	0.03
	Black	Verbal*		6	0.39	0.02	0.28	0.02	0.38	0.02	−0.03	0.03
	Black	Verbal		6	0.38	0.01	0.28	0.02	0.38	0.02	−0.04	0.03
	Black	Verbal*		11	0.39	0.02	0.32	0.02	0.37	0.02	−0.02	0.03
	Black	Verbal		11	0.38	0.01	0.29	0.02	0.35	0.02	−0.01	0.03
	Mean				**0.38**	**0.01**	**0.28**	**0.01**	**0.38**	**0.01**	**−0.05**	**0.01**
	White	Numeric	50–100	6	0.33	0.01	0.14	0.02	0.42	0.02	−0.14	0.03
	White	Numeric	0–100	6	0.37	0.02	0.24	0.02	0.36	0.02	−0.07	0.03
	White	Numeric	50–100	11	0.35	0.02	0.23	0.02	0.39	0.02	−0.03	0.03
	White	Numeric	0–100	11	0.36	0.02	0.26	0.02	0.33	0.02	0.03	0.03
	White	Numeric	0–100	101	0.35	0.02	0.23	0.02	0.33	0.02	−0.03	0.03
	White	Verbal*		6	0.36	0.02	0.22	0.02	0.35	0.02	0.01	0.03
	White	Verbal		6	0.37	0.01	0.22	0.02	0.32	0.02	−0.01	0.02
	White	Verbal*		11	0.38	0.02	0.26	0.02	0.39	0.02	−0.01	0.03
	White	Verbal		11	0.35	0.01	0.22	0.02	0.31	0.01	−0.04	0.02
	Mean				**0.36**	**0.01**	**0.22**	**0.01**	**0.36**	**0.01**	**−0.03**	**0.01**
Black	Black	Numeric	50–100	6	0.24	0.03	0.06	0.06	0.45	0.06	−0.16	0.10
	Black	Numeric	0–100	6	0.29	0.05	0.23	0.06	0.40	0.05	−0.02	0.10
	Black	Numeric	50–100	11	0.25	0.05	0.12	0.07	0.30	0.04	0.05	0.07
	Black	Numeric	0–100	11	0.28	0.06	0.21	0.07	0.31	0.04	0.10	0.07
	Black	Numeric	0–100	101	0.29	0.06	0.24	0.07	0.28	0.04	−0.07	0.05
	Black	Verbal*		6	0.33	0.05	0.25	0.06	0.40	0.07	−0.01	0.13
	Black	Verbal		6	0.26	0.05	0.14	0.06	0.33	0.03	0.05	0.07
	Black	Verbal*		11	0.32	0.07	0.25	0.08	0.32	0.06	0.09	0.08
	Black	Verbal		11	0.41	0.07	0.33	0.08	0.38	0.04	0.10	0.07
	Mean				**0.30**	**0.02**	**0.21**	**0.02**	**0.35**	**0.02**	**0.02**	**0.03**
	White	Numeric	50–100	6	0.31	0.04	0.15	0.06	0.43	0.06	−0.02	0.11
	White	Numeric	0–100	6	0.40	0.04	0.29	0.06	0.23	0.04	−0.04	0.05
	White	Numeric	50–100	11	0.36	0.04	0.22	0.05	0.36	0.06	−0.01	0.08
	White	Numeric	0–100	11	0.32	0.07	0.29	0.08	0.28	0.05	0.16	0.06
	White	Numeric	0–100	101	0.36	0.05	0.24	0.06	0.38	0.05	0.00	0.08
	White	Verbal*		6	0.44	0.05	0.40	0.06	0.36	0.06	0.07	0.10
	White	Verbal		6	0.36	0.04	0.24	0.07	0.28	0.04	0.04	0.07
	White	Verbal*		11	0.37	0.06	0.24	0.08	0.38	0.05	0.03	0.09
	White	Verbal		11	0.41	0.06	0.36	0.06	0.38	0.04	0.18	0.08
	Mean				**0.37**	**0.02**	**0.27**	**0.02**	**0.34**	**0.02**	**0.04**	**0.03**

Note; Numeric/Verbal refers to whether the confidence ratings are labeled with numbers or words. Range refers to whether the numeric confidence scale ranges from either 0 to 100 or 50 to 100. Points refers to whether there are 6 or 11 different ratings on the confidence scale, in addition to the numeric slider scale that had 101 points. Verbal* indicates that these scales had none of their points labeled except for the anchors at the opposite ends of the scale. Chooser and Nonchooser refer to line-up decisions involving either the selection of one of the line-up faces or a response of 'not present', respectively. Calibration and Over–Underconfidence (OU) measure the alignment of confidence to accuracy with higher scores indicating *worse* performance.

other scales, except they were not significantly different from those who used the 0–100, 6-point scale. The remaining scales were comparable to each other.

Relative measure of confidence and accuracy

Although Goodman–Kruskal gamma scores have been the most popular method of measuring the relative confidence/accuracy relationship, growing research demonstrates fundamental problems with this measure (e.g., Masson & Rotello, 2008). We used Somers' *D*, which is computed by a nearly identical formula to that of gamma and does not suffer from the same faults (e.g., Pannu & Kaszniak, 2005). Like gamma scores, Somers' *D* ranges from −1 to 1 and measures the correlation between accuracy and confidence, with a score of 0 indicating no relationship and scores closer to 1 and −1 reflecting perfect positive and negative relationships, respectively.

For choosers, a 2 (participant race: black, white) × 2 (line-up race: black, white) × 9 (confidence scale type) ANOVA of Somers' *D* scores indicated no significant effects, all Fs < 1.21. There was no effect of either the type of confidence scale or whether individuals were responding to same-race or cross-race faces. White participants showed comparable Somers' *D* scores when responding to white lineups ($M = 0.52$, $SD = 0.58$) and black lineups ($M = 0.56$, $SD = 0.53$). Similarly, black participants' Somers' *D* scores were similar for responses to white lineups ($M = 0.57$, $SD = 0.49$) and black lineups ($M = 0.63$, $SD = 0.49$). But, it is important to highlight that these chooser Somers' *D* scores were well above 0, which means that confidence is predictive of accuracy. As for

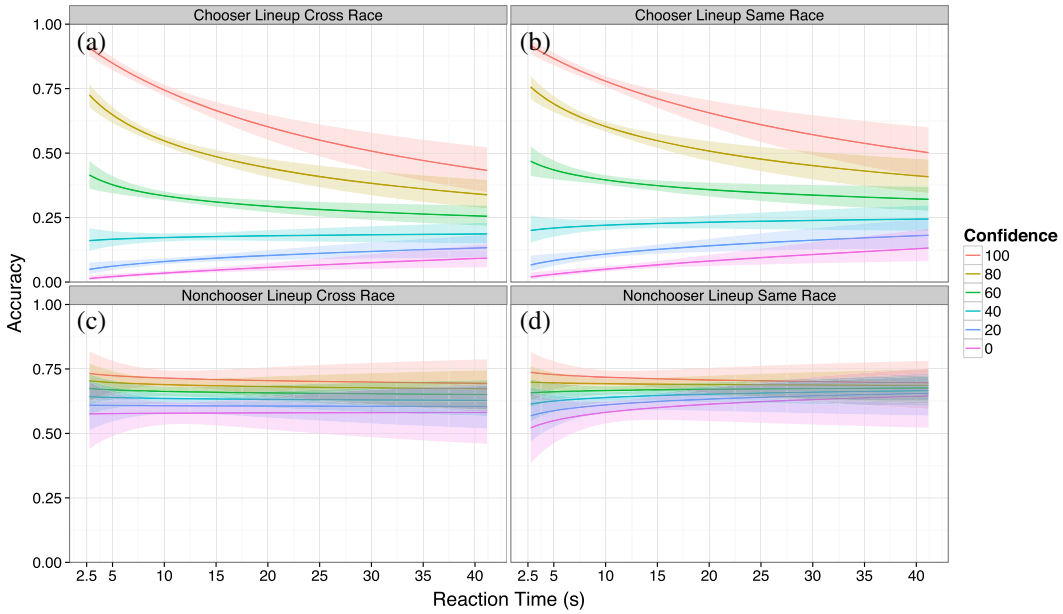


Figure 2. Identification accuracy for Chooser lineups and rejection accuracy for Nonchooser lineups as a function of confidence-level, decision-time to make a response, and same- vs. cross-race faces in the lineup. Error shading represents 95% confidence intervals.

nonchoosers, the same ANOVA of Somers' *D* scores also showed no significant effects, all $Fs < 1.56$. Somers' *D* scores for white participants were 0.09 ($SD = 0.61$) and 0.07 ($SD = 0.65$) for responses to white and black lineups, respectively. And, for black participants, Somers' D scores were 0.08 ($SD = 0.57$) and −0.04 ($SD = 0.57$) for responses to white and black lineups, respectively. Consistent with the results of others (e.g., Sauerland & Sporer, 2009; see Sporer, Penrod, Read & Cutler, 1995), Somers' *D* scores for nonchoosers were much less accurate than the scores for choosers and were close to a value of 0. When participants respond 'not present', confidence has nearly no predictive value for determining accuracy.

Accuracy across changes in decision time and levels of confidence

It is impossible to use the signal detection measure of accuracy, d', to measure changes in accuracy across different decision times because there are too much missing data. This measure is based on the combination of correct identifications and false identifications, and an individual participant rarely makes both of these responses with the identical decision time and with the identical level of confidence—both of which are required if one wanted to trace changes in d' across different decision times and different levels of confidence. So, instead, following what others have done, we measured (a) the correct identification accuracy: the frequency of identifying the target when participants encountered a target-present line-up; and (b) the correct rejection accuracy: the frequency of responding 'not present' when the line-up did not include a previously encountered person, i.e., in a target-absent line-up.

Figure 2a–d is derived from the mixed effects analyses (described later) and shows how identification accuracy (Figure 2a and b) and rejection accuracy (Figure 2c and d) change with (a) increasing decision time, (b) different levels of confidence and (c) same-race and cross-race faces. Note that the error shading for each line in these figures represents a 95% confidence interval. Consistent with previous studies (e.g., Sauerland & Sporer, 2009), we observed that rejection accuracy—or the rate at which individuals correctly respond 'not present'—is relatively constant across all levels of confidence and all decision times. In other words, neither confidence nor decision time is much associated with the accuracy of saying that the target is 'not present' in the line-up. In addition, there appears to be no substantive difference in rejection accuracy for same-race and cross-race faces. But, this is not the case for identification accuracy.

The accuracy of an identification is highest for responses that are both fast and confident (e.g., see the top two confidence ratings). As shown in Figure 2a and b, replicating Sauerland and Sporer (2009), fast (e.g., 2.5 seconds) and maximally confident line-up identifications are associated with correct identification rates of roughly 90%. Although accuracy is better overall for same-race than cross-race identifications, there is little effect of line-up race on the overall shape of the curve. In other words, the CRE appears to exert a comparable effect on identification accuracy at all confidence levels and all decision times.

We used logistic regression fitted via generalized linear mixed effects models in the *lme4* package (Bates, Maechler, Bolker & Walker, 2014) within *R* (R Core Team, 2014) to analyze accuracy (i.e., binary correct and incorrect responses) via a five-way interaction of the following fixed-effect factors: Decision Time (continuous, log-transformed),[4] Confidence (continuous, 0–100), Confidence

[4] We used the Box–Cox transformation method to select the best correction for reaction time to achieve optimal normality of model residuals (Box & Cox, 1964). For these data, log transformations were preferable to inverse reaction times (i.e., 1/RT). Of a total of 19 872 observations (i.e., 1656 participants * 12), 20 were removed because they were not recorded because of technical error. We also used a cutoff of 3 median absolute deviations to trim log-transformed outliers. This ultimately led to the removal of 294 additional observations, or 1.48% of the raw data (or 1.58% including the 20 missing observations, i.e., 294 + 20 = 314, leaving 19 558 for the final sample).

Scale Type (nine different scales), Line-up Decision (Identification vs. Rejection) and Same vs. Cross-Race Line-up. Random effects consisted of an intercept within participant to group participants' responses together.[5] The full model in Wilkinson–Rogers notation can be written as follows: Accuracy ~ DecisionTime * Confidence * ConfidenceScaleType * LineupDecision * SameVsCrossRace + (1 | Participant).

This analysis of accuracy scores produced many significant effects and interactions, which were evaluated using likelihood ratio tests (e.g., Barr, Levy, Scheepers & Tily, 2013). Although we will list all of the significant effects, we only concentrate on the highest order interactions because they moderate all lower order effects. There were significant main effects of Decision Time, $\chi^2(1) = 7.80$, $p < .01$, Confidence, $\chi^2(1) = 1055.09$, $p < .0001$, Line-up Decision, $\chi^2(1) = 1267.58$, $p < .0001$, and Line-up Race, $\chi^2(1) = 17.23$, $p < .001$, which were qualified by interactions between (a) Decision Time and Confidence, $\chi^2(1) = 34.95$, $p < .0001$, (b) Decision Time and Line-up Decision, $\chi^2(1) = 8.62$, $p < .01$, (c) Confidence and Line-up Decision, $\chi^2(1) = 670.52$, $p < .0001$, (d) Line-up Decision and Line-up Race, $\chi^2(1) = 12.15$, $p < .001$, (e) Confidence and Confidence Scale, $\chi^2(8) = 22.14$, $p < .01$, (f) Confidence Scale and Line-up Decision, $\chi^2(8) = 21.86$, $p < .01$, and (g) Decision Time, Confidence and Line-up Decision, $\chi^2(1) = 24.19$, $p < .0001$.

The interaction between Decision Time, Confidence and Line-up Decision is clearly visible in Figure 2a–d by comparing the top two figures (Figure 2a and b), which show the accuracy of a line-up identification, with the bottom two figures, which show the accuracy of responding 'not present'—rejection accuracy (Figure 2c and d). Whereas higher confidence and faster decision times are clearly associated with higher accuracy of an identification (top figures), they are not associated with accuracy of a 'not present' response (bottom figures). Moreover, there is another important part of the identification data: the top figures are clear that it is not the case that all identification responses—at all levels of confidence—are meaningfully related to changes in decision time. It is primarily the two highest levels of confidence that show a significant decrease in accuracy with increases in decision time. For example, the accuracy of the most confident identifications (i.e., the top line) drops nearly in half from ~90% accuracy at the fastest decision times to ~50% accuracy at the slowest decision times. By contrast, for the lowest three levels of confidence, there is very little change in the accuracy of an identification across all decision times.

The Line-up Decision and Line-up Race interaction occurred because there was a cross-race effect on identification accuracy (i.e., Same-Race accuracy $M = 0.39$, $SE = 0.01$ vs. Cross-Race accuracy $M = 0.33$, $SE = 0.01$, $\chi^2(1) = 30.60$, $p < .0001$) but not on rejection accuracy (i.e., Same-Race accuracy $M = 0.67$, $SE = 0.01$ vs. Cross-Race accuracy $M = 0.66$, $SE = 0.01$, $\chi^2(1) = .22$).

Lastly, there were two significant interactions involving the type of confidence scale. Appendix A shows the interaction between Confidence Scale Type and Line-up Decision, which was produced by higher identification accuracy when participants received the 50–100, 6-point confidence scale than any of the other confidence-scale types. By contrast, rejection accuracy was comparable across all confidence scale types. We have no explanation for this difference in identification accuracy with this particular confidence scale. There was also an interaction between Confidence Scale Type and Confidence Level. As seen in Appendix A, this interaction is driven by the larger differences in overall accuracy between the confidence levels for the (a) 0–100, 11-point scale, (b) verbal, 6-point labeled scale, and (c) the verbal, 11-point labeled scale, as compared to the other scales. Overall, we suspect that these interactions are mainly driven by the high power of our design as a result of our large sample size.

GENERAL DISCUSSION

In this study, participants encountered same-race and cross-race faces at encoding and then completed a series of line-up recognition tests that either contained a previously encountered person (i.e., a different photo of the same person seen at encoding) or did not contain a previously seen person. Consistent with past studies, we observed (a) better accuracy when participants responded to same-race than to cross-race faces (e.g., Meissner & Brigham, 2001), (b) better calibration of confidence to accuracy when participants chose a face from a line-up (i.e., choosers) than when they responded 'not present' (i.e., nonchoosers; e.g., Weber & Brewer, 2003) and (c) that the combination of confidence and decision time was a very powerful predictor of chooser accuracy, with relatively fast and confident decisions associated with roughly 90% accuracy (e.g., consistent with Sauerland & Sporer, 2009). By contrast, confidence and decision time were not substantial predictors of nonchooser accuracy (e.g., Sauerland & Sporer, 2009).

There are four novel findings from this study. First, we observed a cross-race effect on the calibration of confidence to accuracy for choosers but not for nonchoosers. When participants selected a face from a line-up, confidence was more closely aligned with accuracy (e.g., high confidence = high accuracy and low confidence = low accuracy) when participants selected a same-race than a cross-race face. Moreover, the reason for this CRE on calibration is that participants were significantly more overconfident when choosing a cross-race than a same-race face. These same-race versus cross-race effects on calibration and on overconfidence occurred in both our African-American and our Caucasian-American participants. Given the importance of eyewitness confidence on jury decision-making and the fact that mistaken eyewitness testimony is one of the major causes of false convictions, these results indicate that witnesses are vulnerable to being significantly more

[5] We attempted to fit multiple random effect models for multi-model selection via Akaike Information Criterion (Burnham & Anderson, 2002) by including additional grouping factors and slope effects consistent with maximal models (see Barr, 2013; Barr *et al.*, 2013; however, none of these models successfully converged despite using multiple optimizers and rescaling and centering of continuous variables, likely due to substantial model complexity (e.g., the five-way interaction of the fixed effects consisted of 2 * 2 * 9 (36) factor levels, in addition to the interaction of the two continuous variables).

overconfident—perhaps causing more false convictions—when they make cross-race than same-race identifications.

One potential underlying mechanism for this cross-race effect on calibration and overconfidence involves the same processes that produce a CRE on memory accuracy. Although calibration is not necessarily tied to the accuracy of a response (e.g., Busey, Tunnicliff, Loftus & Loftus, 2000; Dodson, Bawa & Krueger, 2007; Palmer, Brewer, Weber & Nagesh, 2013), it is frequently the case that conditions that produce increased accuracy also produce increased calibration (e.g., Deffenbacher, 2008). In our study, participants showed a strong correlation between their d' score (i.e., measure of accuracy) and their chooser-calibration score: the correlation between the same-race d' and calibration scores was $r(1449) = -.63$, 95% CI [−0.66, −0.60], $z = 28.21$, $p < .0001$ for our Caucasian-American participants, and it was $r(169) = -.74$, 95% CI [−0.81, −0.67], $z = 12.37$, $p < .0001$ for our African-American participants; for the cross-race identifications, the correlation between the cross-race d' and calibration scores was $r(1449) = -.67$, 95% CI [−0.70, −0.64], $z = 30.67$, $p < .0001$ for our Caucasian-American participants, and it was $r(169) = -.69$, 95% CI [−0.76, −0.61], $z = 10.99$, $p < .0001$ for our African-American participants. Note that these correlation scores are negative because better calibration scores are associated with lower values whereas the opposite is true for d' scores. Overall, for both same-race and cross-race identifications, participants with better accuracy showed better calibration.

Moreover, the magnitude of the CRE on memory accuracy (i.e., worse d' score for cross-race than same-race faces) was related to the magnitude of the CRE on chooser-calibration scores. For Caucasian-American participants, the correlation between their (a) change in d' score for same-race and cross-race faces and (b) change in chooser-calibration score for same-race and cross-race faces was $r(1449) = -.57$, 95% CI [−0.60, −0.54], $z = 24.71$, $p < .0001$; for African-American participants, this correlation score was $r(169) = -.62$, 95% CI [−0.71, −0.52], $z = 7.79$, $p < .0001$. So, the underlying mechanism that causes a CRE on calibration is likely the same mechanism that is producing a CRE on identification accuracy because changes in both scores as a function of the CRE are highly correlated. We assume that cross-race faces decrease the kind and amount of information that is retrieved, which reduces accuracy and in turn impairs the correspondence between confidence and accuracy. But, more fundamentally, the cause of the cross-race effect is not completely known. Some researchers argue that the CRE is a case of a more general phenomenon of in-group/out-group differences, whereas other researchers argue for differences in perceptual expertise to account for the CRE—and both of these accounts may separately contribute to the CRE (see Rhodes, 2013, for a review).

The second novel finding from this study involves the relationship between decision time, confidence and the accuracy of an identification from chooser lineups. Although we replicated the well-established finding that identification accuracy is related to how long it takes participants to make a decision (e.g., Dunning & Stern, 1994; Sporer, 1992; Weber & Brewer, 2006), our findings place a strong boundary condition on this relationship. The relationship between decision time and accuracy is not the same for highly confident, moderately confident and weakly confident identifications. The top panels of Figure 2 clearly show that it is only when participants are highly confident (i.e., the top two confidence ratings) that there is a dramatic decrease in identification accuracy from faster decisions to slower decisions (e.g., for the highest level of confidence within the same-race condition, there is a drop in performance from 91.63% to 50.19% for decision times of 2.77 to 41.17 seconds). By contrast, when participants are moderately or weakly confident in their identification, there is very little change in accuracy from faster to slower decisions. From a practical perspective, this means that a high confidence identification will be much less effective at predicting accuracy for slower responses than for faster responses, but there is little change in accuracy for low and moderate confidence responses across all decision times.

Theoretically, the pattern of data that needs to be explained is why longer decision times are associated with either dramatic or no changes in identification accuracy, depending on the confidence in the identification (e.g., as shown in Figure 2, increasing decision time is associated with a large drop in the accuracy of high confidence responses, but it is associated with no change in accuracy for low confidence responses). To explain this differing relationship between accuracy, confidence and decision time that is clear in the top panels of Figure 2, we propose a Time–Confidence–Criteria account. We argue that the criteria about the kind and amount of memorial information that justify a particular confidence rating change with decision time. Longer decision times are associated with using increasingly liberal criteria for making a response at all confidence ratings. For example, high confidence ratings for slower responses are based on less vivid and less plentiful memorial information, as compared to high confidence ratings for faster responses. Or, conversely, participants use a stricter criterion for the kind and amount of memorial information that is required to make a high confidence response at faster than at slower decisions. Because identification accuracy is likely poor when it is based on less vivid and less plentiful memorial information, this account can easily explain the severe loss of accuracy for high confidence identifications with increasing decision time.

But, why is it that accuracy only shows a dramatic decline with increasing decision time for high confidence responses and not for low confidence responses (and only a small drop in accuracy for moderate confidence responses) if, according to our account, the same process of using a looser criterion with increasing decision time also occurs for making moderate and low confidence responses? The magnitude of the decrease in accuracy is determined by the distance from chance performance. Consider low confidence responses. These are responses that are presumably based on little, if any, diagnostic memorial information, which is why performance is no different from chance, even for the fastest responses. Loosening the criteria with increasing decision time for what constitutes as a low confidence response has little functional consequence because performance is already at chance and performance cannot get worse than chance.

The third major finding from this study is that the cross-race effect exerts a similar effect on the accuracy of an identification, regardless of either the particular level of confidence or the particular decision time. In other words, we observed no evidence that suggests that the CRE on identification accuracy is more powerful for either (1) low versus high confident responses (or vice versa) or (2) slower versus faster decision times. As mentioned earlier, our large sample size provided us with a large amount of power to detect even small-sized effects and so we argue that the CRE is a phenomenon that affects all kinds of memorial information.

Finally, the fourth significant finding involves the effect of using different confidence scales. We replicated Weber, Brewer and Margitich (2008) and conceptually similar findings by Wallsten and colleagues (e.g., Wallsten *et al.*, 1993): (a) both our numeric and verbal confidence scales produced comparable calibration scores and over–underconfidence scores and (b) both numeric and verbal scales varied in the same way when individuals either chose a face from a line-up or responded 'not present'. The novel aspect of our study, however, is that no one has systematically examined so many scales in order to answer the following questions about measuring the relationship between confidence and accuracy: (1) does it matter if the confidence scale consists of few points (i.e., 6 points) or many points (i.e., 11 points or 101 points with the slider scale)?; (2) when using numeric confidence scales, does it matter—at least for our identification task—if the low end of the scale is fixed at 0% or 50% (i.e., 0–100% vs. 50–100%)?; and (3) more generally, how resistant is the confidence–accuracy relationship to changes in the format of the confidence scale?

In terms of measuring the relationship between confidence and accuracy, the size (i.e., number of points on the scale) and format of the scale do not seem to matter much. Our data indicate that participants show a strong degree of flexibility in mapping a feeling of confidence onto a provided confidence scale so that they are generally comparably calibrated regardless of the format of the confidence scale. We contrasted scales with 6 points, 11 points and 101 points (i.e., slider) and observed no substantial differences in either the effectiveness of the scale at measuring the calibration of confidence to accuracy or the sensitivity of the scale to cross-race effects. Moreover, consider the effects of using a numerical confidence scale that ranges from either 50–100% or 0–100%: if participants had ignored our labels for the endpoints of the scale and, instead, had interpreted the values of the scale literally so that, for instance, a confidence rating of 50% means a 50% chance of being correct, then participants' CA relationship would have been much more impaired when using the 50–100% scale than the 0–100% scale. But, this was generally not the case. When individuals chose a face from a line-up, we observed no difference between the 0–100% scale and the 50–100% scale in calibration scores, although individuals tended to be more overconfident with the 0–100% scale than the 50–100% scale. Overall, we suspect that when individuals understand how the confidence scale maps onto accuracy (e.g., that the endpoints correspond to chance and perfect performance, respectively), then the format of the scale generally does not matter.

In conclusion, we observed that individuals tend to be overconfident when selecting a cross-race face from a line-up that worsens the relationship between their confidence and accuracy of an identification for cross-race than same-race faces. In addition, high confidence in an identification is a tremendously powerful predictor of accuracy when the decision is made quickly. But, this predictor is extremely sensitive to decision time and quickly loses its predictive power with increases in decision time. By contrast, although moderate and weakly confident identifications are less powerful predictors of accuracy, they also generally maintain their respective predictive power across changes in decision time.

ACKNOWLEDGEMENT

This research was supported by National Science Foundation Grant SES 0925145.

REFERENCES

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. doi: 10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-7). http://CRAN.R-project.org/package=lme4

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, *26*, 211–252.

Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. *Law and Human Behavior*, *30*, 31–50.

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. doi: 10.1037/1076-898X.12.1.11

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Corenblum, B., & Meissner, C. A. (2006). Recognition of faces of ingroup and outgroup children and adults. *Journal of Experimental Child Psychology*, *93*, 187–206.

Deffenbacher, K. A. (2008). Estimating the impact of estimator variables on eyewitness identification: A fruitful marriage of practical problem solving and psychological theorizing. *Applied Cognitive Psychology*, *22*, 815–826. doi: 10.1002/acp.1485

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied*, *19*, 345–357. doi: 10.1037/a0034596

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging*, *22*(1), 122–133.

Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identifications. *Journal of Applied Psychology*, *87*, 951–962.

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67*, 818–835.

Evans, J. R., Marcon, J. L., & Meissner, C. A. (2009). Cross-racial lineup identification: Assessing the potential benefits of context reinstatement. *Psychology, Crime & Law*, *15*, 19–28.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi: 10.3758/BF03193146

Horry, R., & Wright, D. B. (2008). I know your face but not where I saw you: Context memory is impaired for other-race faces. *Psychonomic Bulletin & Review*, *15*, 610–614. doi: 10.3758/PBR.15.3.610

Horry, R., Wright, D. B., & Tredoux, C. G. (2010). Recognition and context memory for faces from own and other ethnic groups: A remember–know investigation. *Memory & Cognition*, *38*, 134–141. doi: 10.3758/MC.38.2.134

Jackiw, L. B., Arbuthnott, K. D., Pfeifer, J. E., Marcon, J. L., & Meissner, C. A. (2008). Examining the cross-race effect in lineup identification using Caucasian and First Nation samples. *Canadian Journal of Behavioral Science*, *40*, 52–57.

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*, 1304–1316.

Koriat, A. & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517.

Marcon, J. L., Susa, K. J., & Meissner, C. A. (2009). An examination of the cross-race effect in a repetition-lag paradigm. *Psychonomic Bulletin and Review*, *16*, 99–103.

Masson, M. E. J., & Rotello, C. M. (2008). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 509–527. doi: 10.1037/a0014876.

Meissner, C. A. & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy & Law*, *7*, 3–35. doi: 10.1037//1076-8971.7.1.3

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: A dual-process approach. *Applied Cognitive Psychology*, *19*, 545–567.

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, *33*(5), 783–792. doi: 10.3758/BF03193074

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

Pannu, J. K., & Kaszniak, A. W. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychology Review*, *15*, 105–130. doi:10.1007/s11065-005-7091-6.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rhodes, G. (2013). Face recognition. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 46–68). New York, NY: Oxford University Press.

Rhodes, M. G., Sitzman, D. M., & Rowland, C. A. (2013). Monitoring and control of learning own-race and other-race faces. *Applied Cognitive Psychology*, *27*, 553–563. doi: 10.1002/acp.2948

Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime & Law*, *13*, 611–625.

Sauerland, M. & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46–62.

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

Smith, S. M., Stinson, V., & Prosser, M. A. (2004). Do they all look alike? An exploration of decision-making strategies in cross-race facial identifications. *Canadian Journal of Behavioral Science*, *36*, 146–154.

Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision-times and person descriptions of choosers and nonchoosers. *European Journal of Social Psychology*, *22*, 157–180.

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78*, 22–33.

Sporer, S. L. (1994). Decision-times and eyewitness identification accuracy in simultaneous and sequential lineups. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 300–327). New York: Cambridge University Press.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law*, *7*, 36–97.

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

Tredoux, C. (1999). Statistical considerations when determining measures of lineup size and lineup bias. *Applied Cognitive Psychology*, *13*(S1), S9-S26.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, *29*, 151–173.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing—General-principles and empirical-evidence. *Knowledge Engineering Review*, *10*, 43–62.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*, 176–190.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence–accuracy calibration in face recognition. *Journal of Applied Psychology*, *88*, 490–499.

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

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

Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: The unruly 10–12 second rule. *Journal of Experimental Psychology: Applied*, *10*, 139–147.

Wells, G. L., & Penrod, S. D. (2011). Eyewitness identification research: Strengths and weaknesses of alternative methods. In B. Rosenfeld, & S. D. Penrod (Eds.), *Research methods in forensic psychology*. Hoboken, NJ: John Wiley and Sons.

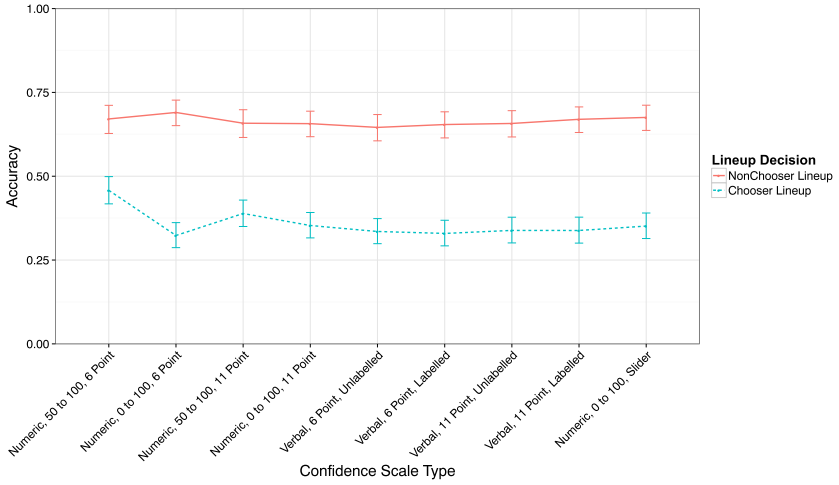
Wright, D. B., Boyd, C. E., & Tredoux, C. G. (2001). A field study of own-race bias in South Africa and England. *Psychology, Public Policy, and Law*, *1*, 119–133.

Wright, D. B., Boyd, C. E., & Tredoux, C. G. (2003). Inter-racial contact and the own-race bias for face recognition in South Africa and England. *Applied Cognitive Psychology*, *17*, 365–373.

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2012). Perception and motivation in face recognition: A critical review of theories of the cross-race effect. *Personality and Social Psychology Review*, *16*, 116–142. doi: 10.1177/1088868311418987

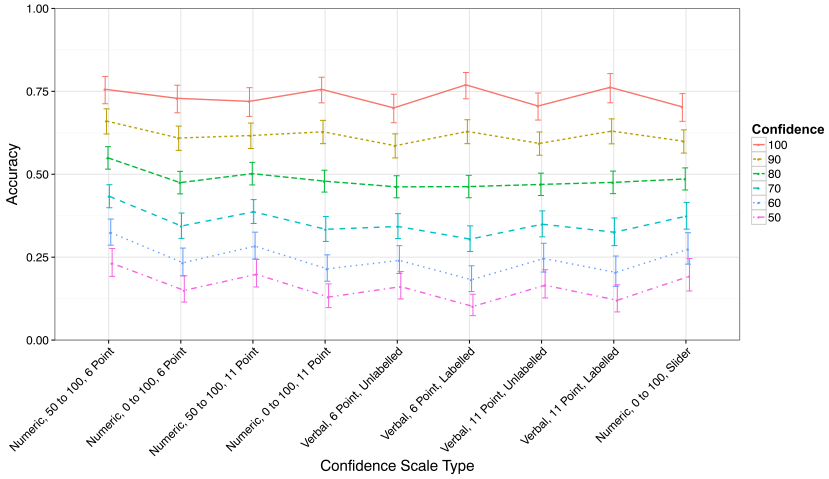
APPENDIX A

Confidence Scale Type x Lineup Decision Interaction



Note. Chooser accuracy refers to the rate of identifying the target and nonchooser accuracy refers to the rate of correctly responding "not present." Error bars represent 95% CIs.

Confidence Scale Type x Confidence Level Interaction



Note. Error bars represent 95% CIs.