




Evidence for a confidence–accuracy relationship in memory for same- and cross-race faces

Thao B. Nguyen, Kathy Pezdek & John T. Wixted


To cite this article: Thao B. Nguyen, Kathy Pezdek & John T. Wixted (2017) Evidence for a confidence–accuracy relationship in memory for same- and cross-race faces, *The Quarterly Journal of Experimental Psychology*, 70:12, 2518-2534, DOI: [10.1080/17470218.2016.1246578](https://doi.org/10.1080/17470218.2016.1246578)


To link to this article: <https://doi.org/10.1080/17470218.2016.1246578>

 [View supplementary material](#) 



 Published online: 05 Nov 2016.

 [Submit your article to this journal](#) 

 Article views: 312

 [View related articles](#) 

 [View Crossmark data](#) 

 Citing articles: 4 [View citing articles](#) 

Evidence for a confidence–accuracy relationship in memory for same- and cross-race faces

Thao B. Nguyen^a, Kathy Pezdek^a and John T. Wixted^b

^aDepartment of Psychology, Claremont Graduate University, Claremont, CA, USA; ^bDepartment of Psychology, University of California, San Diego, CA, USA

ABSTRACT

Discrimination accuracy is usually higher for same- than for cross-race faces, a phenomenon known as the cross-race effect (CRE). According to prior research, the CRE occurs because memories for same- and cross-race faces rely on qualitatively different processes. However, according to a continuous dual-process model of recognition memory, memories that rely on qualitatively different processes do not differ in recognition accuracy when confidence is equated. Thus, although there are differences in overall same- and cross-race discrimination accuracy, confidence-specific accuracy (i.e., recognition accuracy at a particular level of confidence) may not differ. We analysed datasets from four recognition memory studies on same- and cross-race faces to test this hypothesis. Confidence ratings reliably predicted recognition accuracy when performance was above chance levels (Experiments 1, 2, and 3) but not when performance was at chance levels (Experiment 4). Furthermore, at each level of confidence, confidence-specific accuracy for same- and cross-race faces did not significantly differ when overall performance was above chance levels (Experiments 1, 2, and 3) but significantly differed when overall performance was at chance levels (Experiment 4). Thus, under certain conditions, high-confidence same-race and cross-race identifications may be equally reliable.

ARTICLE HISTORY

Received 29 April 2016
Accepted 26 September 2016
First Published Online 5
November 2016

KEYWORDS

Confidence–accuracy
relationship; Cross-race
effect; Face recognition
memory; Metacognition

Discrimination accuracy is higher when recognizing individuals of the same race than a different race, a phenomenon known as the cross-race effect (CRE; Malpass & Kravitz, 1969). Meissner, Brigham, and Butz (2005) and others (Goldinger, He, & Papesh, 2009; Hills & Lewis, 2006; Hugenberg, Young, Bernstein, & Sacco, 2010) argue that the CRE is a result of qualitative differences in encoding same- and cross-race faces. Consistent with the classic dual-process framework of familiarity and recollection (Mandler, 1980), Meissner et al. reported that recollection, which they measured using the remember/know paradigm, is more likely to occur for same- than for cross-race faces. This is because observers are more likely to undergo effortful elaboration and encode qualitatively diagnostic information of same- than cross-race faces, and are thus more likely to retrieve specific episodic information to help differentiate among similar

faces during recognition for same- than for cross-race faces.

Although qualitative differences in memory may result in higher recognition accuracy for same- than for cross-race faces, according to the continuous dual-process model of recognition memory (Wixted & Mickes, 2010), memories based on qualitatively different processes may not always differ in old/new recognition accuracy. According to Wixted and Mickes's (2010) continuous dual-process model of recognition memory, which incorporates signal detection theory (Banks, 1970) and classic dual-process models (Mandler, 1980), memory strength is comprised of two dimensions of recollection and familiarity. Traditionally, recollection-based memories (i.e., made with “remember” judgments) are considered to represent stronger memories than familiarity-based memories (i.e., made with “know” judgments;

Donaldson, 1996; Dunn, 2004). However, Wixted and Mickes (2010) argued that the remember/know paradigm does not assess only memory strength, but also the quality or content of the memory. To dissociate strength from quality of a memory signal, Wixted and Mickes tested participants on their recognition memory for words. They asked participants to make a confidence rating in their old/new recognition judgment and then asked them to make a remember/know/guess judgment for each word they judged to be "old". The remember/know/guess judgments were used to assess the quality or content of the memory, whereas the confidence ratings were used to assess the strength of the memory. According to signal detection theory, each level of confidence represents a different response criterion. A high confidence rating represents a more conservative criterion, or a higher threshold for memory strength, than a low confidence rating. Thus, an observer would need a stronger memory signal to make a high-confidence "old" judgment than a low-confidence "old" judgment, and this would be true whether the memory is recollection or familiarity based.

Wixted and Mickes (2010) reported that, overall, remember judgments yielded higher old/new discrimination accuracy than know and guess judgments. However, and more importantly, when confidence was equated, old/new discrimination accuracy did not differ between high-confidence remember (i.e., recollection-based) and high-confidence know (i.e., familiarity-based) judgments. In contrast, high-confidence remember and high-confidence know judgments differed in source accuracy (i.e., the colour or location in which the word was presented during the study phase). Thus, recollection-based and familiarity-based memories differ qualitatively even when they are equated for overall memory strength. Similar findings were reported in a subsequent and detailed study reported by Ingram, Mickes, and Wixted (2012). These results suggest that when comparing judgments made with the same confidence level (i.e., those similar in memory strength), memories that rely on qualitatively different processes, such as memory for same- versus cross-race faces (Goldinger et al., 2009; Hills & Lewis, 2006; Hugenberg et al., 2010; Meissner et al., 2005), may not significantly differ.

Extending the dual-process model of recognition memory to memory for same- and cross-race faces, if high-confidence memories that are based on

qualitatively different processes do not differ in recognition accuracy, then the magnitude of the CRE should be attenuated at the highest level of confidence. The key point is that comparing memory for same- and cross-race faces similar in memory strength, which we refer to as confidence-specific accuracy, is different from examining overall same- and cross-race discrimination accuracy. Overall discrimination accuracy assesses the average difference in memory, and there is no doubt that it differs for same- versus cross-race recognition, perhaps because same-race faces are more likely to involve recollection (Meissner et al., 2005) or qualitatively superior information (Goldinger et al., 2009; Hills & Lewis, 2006; Hugenberg et al., 2010). In contrast, confidence-specific accuracy assesses differences in recognition accuracy for memories (that can differ qualitatively) that are similar in memory strength, and it is an open question as to whether a CRE is evident once confidence is equated. Our study investigates this issue as it applies to eyewitness memory. Based on past research on discrimination accuracy and the CRE, cross-race identifications are viewed as less reliable than same-race identifications. However, if we compare confidence-specific accuracy, the more informative measure for triers of fact, cross-race identifications may not always be less reliable, especially if they are made with high confidence.

Numerous variables reported to affect overall discrimination accuracy at encoding have been reported not to affect confidence-specific accuracy. Using an eyewitness identification paradigm, Palmer, Brewer, Weber, and Nagesh (2013) tested the effects of exposure duration and divided attention on memory to assess whether the confidence-accuracy (CA) relationship is weakened under poor encoding conditions. They reported that confidence was a reliable predictor of accuracy in all experimental conditions for "choosers" (i.e., those who made an identification) but not "non-choosers" (i.e., those who do not make an identification). More importantly, the proportion correct at the highest level of confidence did not differ as a function of retention interval, exposure duration, or divided attention. Thus, although overall discrimination accuracy has been reported to be poor under specific conditions (e.g., short exposure duration, divided attention at encoding, etc.; Deffenbacher, Bornstein, McGorty, & Penrod, 2008), the results of Palmer et al. (2013) suggest that these variables do not affect confidence-specific accuracy. Although Palmer et al. used an eyewitness identification

paradigm, their results are consistent with Wixted and Mickes's (2010) continuous dual-process model of recognition memory.

Recently, Dodson and Dobolyi (2015) tested cross-race recognition accuracy using a simultaneous line-up procedure. Not surprisingly, they found a significant cross-race effect in that d' was higher for same- than for cross-race identifications (and this was true for both Black and White participants). Separately, they also examined accuracy at each level of confidence, which was measured using a 100-point confidence scale. Overall, calibration, a measure of how well confidence ratings align with accuracy, was significantly better for same- than for cross-race identifications, and at each level of confidence the proportion correct was slightly higher for same- than for cross-race identifications. However, looking at Dodson and Dobolyi's Figure 1a, the proportion correct was consistently higher for same- than for cross-race faces, suggesting that the CA relationship may not differ as a function of face race. Although there was a small but significant CRE on measures of calibration and the proportion correct data, the vastly stronger effect was the relationship between confidence and accuracy. For example, for decisions made with zero confidence, accuracy was very low for both same- and cross-race identifications (15% correct and 11% correct, respectively, estimated from their Figure 1a). For decisions made with 100% confidence, accuracy was much higher for both same- and cross-race identifications (80% correct and 77% correct, respectively).

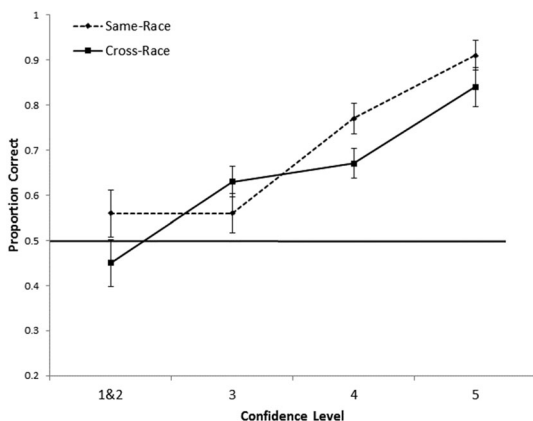


Figure 1. Confidence-specific accuracy assessed with confidence-accuracy characteristic (CAC) curves in Experiment 1. Chance performance is denoted by the horizontal line. Error bars represent standard error.

Thus, although a reliable CRE was still evident even after controlling for confidence, it was of negligible magnitude compared to low-versus-high-confidence identifications.

We asked whether the same- versus cross-race manipulation affects confidence-specific accuracy (in addition to overall discrimination accuracy) using an old/new recognition procedure. In an old/new recognition memory test, only one face is shown at a time. This test is thus more similar to a show-up than to a line-up (a line-up was the procedure used in Dodson and Dobolyi, 2015). A show-up is a standard police eyewitness identification procedure in which an eyewitness (often a victim) is presented with a suspect who has been apprehended shortly after a crime was committed. The police want to know whether the witness recognizes the suspect, yes or no, as the person who committed the crime. Thus, a show-up is a real-world old/new recognition procedure. Although line-ups are reported to result in higher discrimination accuracy than show-ups (Wixted & Mickes, 2015), these two identification procedures may or may not differ in terms of how race affects the confidence-accuracy relationship.

We investigated same- and cross-race confidence-specific accuracy by performing secondary analyses on datasets obtained from four old/new recognition memory studies on same- and cross-race faces. These are datasets obtained from Nguyen and Pezdek (2015), Blandón-Gitlin, Pezdek, Saldivar, and Steelman (2014), O'Brien and Wasson (2015), and Pezdek, O'Brien, and Wasson (2012). All four of these studies tested memory for a list of recently presented faces. Faces from all four studies were obtained from a database of male faces used by Meissner et al. (2005). Meissner et al. provided two photographs of each face: (a) each man smiling and dressed in a casual shirt, and (b) each man with a neutral facial expression and dressed in a maroon coloured shirt. In all four studies, the first set was presented during the study phase, and the second set was presented during the test phase. Consistent with other studies discussed thus far (Dodson & Dobolyi, 2015; Palmer et al., 2013; Wixted & Mickes, 2010), in our four studies, confidence ratings were collected immediately after each old/new recognition judgment during the initial (and only) test phase. This provides optimal conditions for a reliable CA relationship. As Brewer (2006) suggested, confidence may be predictive of accuracy only if the confidence rating is made shortly after the initial recognition judgment. If there is a substantial delay

between the recognition and the confidence judgment, the information about evidence strength that is used to make an old/new judgment may not be accessible to the observer, and he or she is likely to rely on less accurate information to make the meta-cognitive judgment.

Consistent with previous research on the CA relationship (Palmer et al., 2013), we defined accuracy at each level of confidence as $\text{no. hits}_c / (\text{no. hits}_c + \text{no. false alarms}_c)$, where c indicates that the hits and false alarms were made with a specific level of confidence. In each of the four studies, this proportion correct (confidence-specific accuracy) was computed separately for same- and cross-race faces for each level of confidence for each participant. We focus on analyses of “old” responses, which is similar to analysing data from only “choosers” in an eyewitness identification paradigm. In other words, this proportion is the probability that a positive identification or an “old” judgment is correct, a useful measure of reliability for triers of fact. It is reasonable to focus on “old” responses rather than “new” responses because it is only the cases in which an eyewitness identifies the suspect that proceed to trial. We did assess the CA relationship for “new” responses to compare it to the CA relationship for “old” responses and briefly report these results below. Consistent with past research (Palmer et al., 2013; Sporer, Penrod, Read, & Cutler, 1995), the CA relationship for “new” responses (i.e., non-choosers) was weaker and less reliable than the CA relationship for “old” responses (i.e., choosers); therefore, we do not focus on the results for “new” responses.

Calibration statistics were not computed for these studies because confidence was not collected on a 0–100-point scale. Still, one can examine the CA relationship by plotting the confidence–accuracy characteristic (CAC) curves (Mickes, 2015). CAC curves are created by plotting the average proportion correct [$\text{no. hits}_c / (\text{no. hits}_c + \text{no. false alarms}_c)$] for each level of confidence. CAC curves and confidence-specific accuracy allow researchers to determine how reliable or trustworthy same- and cross-race face judgments are. Receiver operating characteristic (ROC) curves and other measures of discrimination accuracy (e.g., d') do not convey that information (see Mickes, 2015, for a thorough discussion of ROC versus CAC analyses).

To illustrate further the difference between discrimination accuracy (ROC or d') and confidence-specific accuracy (CAC), we present a hypothetical dataset in Table 1. Suppose participants studied

Table 1. Comparison of d' and confidence-specific accuracy in a hypothetical data set.

Condition	No. hits	Hit rate	No. false alarms	False-alarm rate	d'
Condition A	48	.96	32	.64	1.39
Condition B	24	.48	16	.32	0.42

Table 1a. Comparison of d' and confidence-specific accuracy in a hypothetical data set.

Condition	Confidence	No. hits	No. false alarms	Proportion correct
Condition A	5	18	2	.9
	4	16	4	.8
	3	6	4	.6
	2	6	14	.3
	1	2	8	.2
	Total	48	32	
Condition B	5	9	1	.9
	4	8	2	.8
	3	3	2	.6
	2	3	7	.3
	1	1	4	.2
	Total	24	16	

faces under two experimental conditions. Condition A (e.g., same-race faces or long exposure duration) yields higher discrimination accuracy, measured by d' ; on the other hand, Condition B (e.g., cross-race faces or short exposure duration) yields lower discrimination accuracy. Although these two conditions differ radically in discrimination accuracy, the CAC curves for Conditions A and B are identical. This is because d' and ROC curves measure discrimination accuracy whereas confidence-specific accuracy (proportion correct) and CAC curves measure reliability or trustworthiness. Thus, although d' (discrimination accuracy) is reportedly lower for cross-race faces, a cross-race judgment may not always be less reliable than a same-race judgment.

Experiment 1: Nguyen and Pezdek (2015)

Method

In Experiment 1 of their study, Nguyen and Pezdek (2015) tested 48 Caucasian undergraduate and graduate students for recognition memory of 48 White (same-race) and 48 Black (cross-race) faces, half old and half new. The study was a 2 [race of target face: White (same-race), Black (cross-race)] \times 2 (exposure time: 1.5 s, 5 s) \times 2 (interstimulus interval, ISI: 3 s, 9 s) within-subjects factorial design. The experiment was conducted on a computer using PsychoPy software

(Peirce, 2007), and participants were tested individually. The study phase consisted of two blocks, counter-balanced across participants. ISI was the blocked variable; exposure time was randomized within each block. Half of the 48 study faces (24 White and 24 Black) in each block were presented for 1.5 s each; half were presented for 5 s each. In one block, participants were presented 24 faces (12 White and 12 Black) each followed by an ISI of 3 s. In the other block, 24 different faces were followed by an ISI of 9 s. Each study face was presented individually. Across participants, each face was equally often presented with a 1.5-s or 5-s exposure time and following a 3-s or 9-s ISI. After the study phase, participants had four minutes to sort a list of words alphabetically. This served as a distractor task to avoid recency effects. The test phase followed. Each of the 96 test faces was presented individually, randomly arranged across participants. Participants pressed the “1” key to indicate an “old” face and the “0” key to indicate a “new” face. Participants rated their confidence on a scale of 1 (*completely guessing*) to 5 (*absolutely sure I’m correct*) by pressing one of the 1–5 number keys and were instructed to use the full range of the scale. Each face equally often served as a target or a foil face. For the purposes of the current analyses, race of target face results were collapsed across exposure time and ISI conditions because we were only interested in whether confidence-specific accuracy differed between same- and cross-race faces.

Results

There was a significant difference in discrimination accuracy between same- and cross-race faces as measured by the signal detection measure d' , 1.10 and 0.75, respectively, $F(1, 47) = 28.66$, $p < .001$, $d = 0.77$. Note that this is a large effect size. Given this significant CRE when examining old/new discrimination accuracy assessed with d' , the similarity in trustworthiness (i.e., confidence-specific accuracy or the proportion of correct “old” responses) of same- and cross-race faces, as measured by confidence-specific accuracy or CAC analysis, cannot be attributed to a manipulation failure of the race of face variable.

Next, we assessed confidence-specific accuracy for “old” responses across participants. A 2 (race of target face: same-race, cross-race) \times 4 (confidence: 1 & 2, 3, 4, 5) repeated measures analysis of variance (ANOVA) was conducted on the proportion correct data. There were few observations at the two lowest levels of

confidence so we collapsed across Levels 1 and 2 for all analyses to reduce noise. The mean proportion correct for each condition is plotted in Figure 1. We conducted chi-square tests of independence to ensure that race of target face and confidence were orthogonal factors (i.e., observers were not more likely to rate cross-race faces with lower levels of confidence than same-race faces).¹ Results from the ANOVA indicated that there was not a significant main effect of race of target face, $F(1, 47) = 2.72$, $p = .11$, $\eta_p^2 = .06$, power = .37, but there was a significant main effect of confidence level, $F(3, 45) = 31.48$, $p < .001$, $\eta_p^2 = .68$, power = 1.00. Confidence accounted for much more variance in proportion correct than race of target face, $\eta_p^2 = .68$ and $\eta_p^2 = .06$, respectively. Post hoc pairwise comparisons were conducted to examine the main effect of confidence. For these comparisons, and all other pairwise comparisons in this paper, we used a Bonferroni correction of α/n , where $\alpha = .05$, and n is the number of pairwise comparisons. All comparisons were one-tailed. With a Bonferroni correction of $\alpha = .008$, the proportion correct at the highest level of confidence was significantly higher than the proportion correct at each of the other three lower levels of confidence (all $ps < .001$). Furthermore, the proportion correct values among each of the three lower confidence levels were not statistically significantly different from one another (all $ps < .01$).

There was also a significant interaction between race of target face and confidence level, $F(3, 45) = 3.19$, $p = .03$, $\eta_p^2 = .18$, power = .70. To explore further whether the confidence–accuracy relationship differed as a function of race of target face, dependent t tests were conducted on the proportion correct for same- and cross-race faces at each of the four levels of confidence. A Bonferroni correction of $\alpha = .0125$ was used. To make the mean estimates of proportion correct more stable, for each t test, we removed participants who did not make at least two ratings at that particular confidence level; therefore, the sample sizes across all t tests were slightly different (total $N = 48$). The dependent t tests indicated that the proportion correct for same- and cross-race judgments did not significantly differ at Confidence Level 5, $t(45) = 1.32$, $p = .10$, power = .25, at Confidence Level 4, $t(46) = 2.26$, $p = .01$, power = .60, at Confidence Level 3, $t(46) = -1.74$, $p = .09$, power = .40, or at Confidence Levels 1 and 2, $t(39) = 1.55$, $p = .06$, power = .33. Although the significant interaction between race of target face and confidence level indicated that differences in memory accuracy between same- and cross-

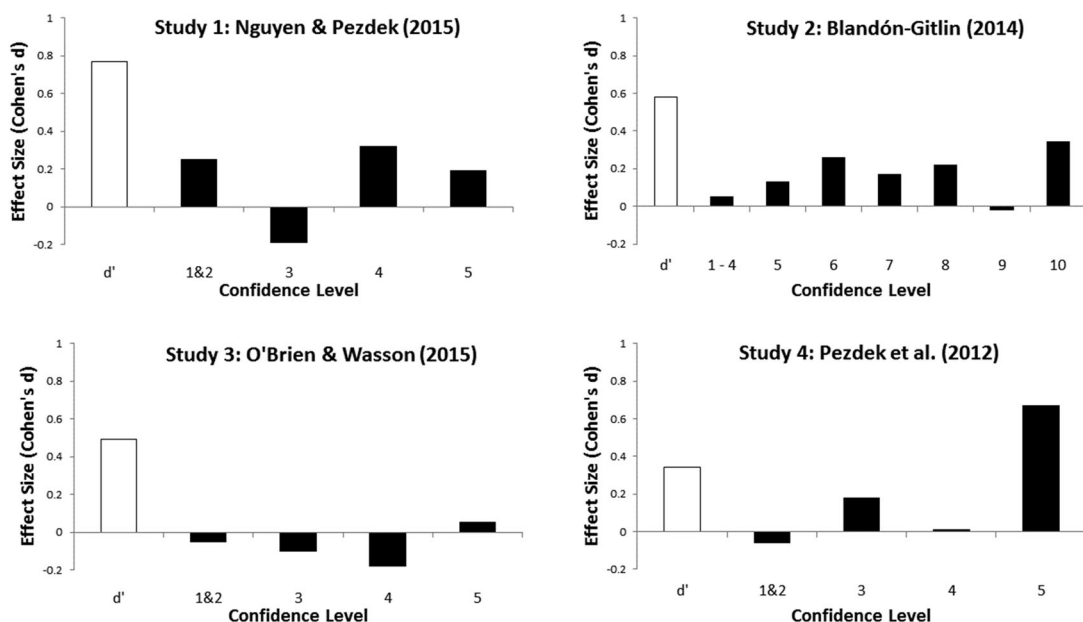


Figure 2. A comparison of the effect size of the difference in discrimination accuracy (as measured by d') and confidence-specific accuracy (as measured by proportion correct for each level of confidence) between same-race and cross-race faces in Experiment 1, Experiment 2, Experiment 3, and Experiment 4.

race faces varied with confidence level, the pairwise comparisons indicated that confidence-specific accuracy did not significantly differ between same- and cross-race faces at any confidence level.

The small sample size in this study (and the following studies) may have resulted in low power to detect statistically significant differences between same- and cross-race face confidence-specific accuracy, and a conservative Bonferroni correction may result in Type II errors. It is thus useful to compare effect sizes for each comparison. Figure 2 (top left panel) shows a comparison of the effect sizes as measured by Cohen's d with Hedges's g adjustment for small sample sizes. Although the size of the difference in discrimination accuracy (d') between same- and cross-race faces is large, the effect sizes for confidence-specific accuracy indicate small differences between same- and cross-race face accuracy at each level of confidence. Based on this comparison of effect sizes, it is clear that the magnitude of the CRE is larger for discrimination than confidence-specific accuracy. These results suggest that the CRE may be attenuated when confidence is equated.

We also ran mixed-effects logistic regression analyses using the `glmer` function in R to address potential issues with conducting ANOVAs on aggregate proportions. Results from two mixed-effects models

are presented in Table 2. The first model consisted of random intercepts of participants (as judgment trials were nested within participants) and the main effects of face race and confidence. In the second model, we added the interaction between face race and confidence. A model comparison indicated that adding the interaction term did not improve model fit; however, the main effect of face race was no longer statistically significant. Thus, when confidence was equated, accuracy for same- and cross-race faces did not significantly differ. In contrast, confidence was a significant predictor of accuracy even with face race in the model; higher levels of confidence were associated with higher accuracy than lower levels of confidence. These results indicate that confidence is more informative of accuracy than is face race. The nonsignificant interaction between face race and confidence indicates that the CA relationship did not differ between same- and cross-race faces.

We also assessed whether participants who differ in magnitude of the CRE also differ in their CA relationships. In a third mixed-effects logistic regression model, we included random slopes of face race and found that this did not improve the fit of the model. Given that the magnitude of the CRE did not vary in our samples, we do not discuss these results further.

Table 2. Results from mixed-effects logistic regression analyses.

Experiment	Model 1		Model 2			Models 1 and 2 comparison
	Face race	Confidence	Face race	Confidence	Face Race × Confidence	
1	$z = 3.52 (.39)^{***}$	$z = 11.35 (.57)^{***}$	$z = 0.44 (.14)$	$z = 8.24 (.54)^{***}$	$z = 0.77 (.07)$	$\chi^2(1) = 0.60$
2	$z = 3.96 (.46)^{***}$	$z = 15.05 (.41)^{***}$	$z = 1.49 (.51)$	$z = 12.27 (.41)^{***}$	$z = -0.14 (-.01)$	$\chi^2(1) = 0.02$
3	$z = 1.39 (.26)$	$z = 3.88 (.32)^{***}$	$z = -0.08 (-.05)$	$z = 2.56 (.28)^*$	$z = 0.50 (.08)$	$\chi^2(1) = 0.25$
4	$z = 2.09 (.40)^*$	$z = 6.65 (.63)^{***}$	$z = -1.03 (-.78)$	$z = 3.83 (.49)^{***}$	$z = 1.61 (.31)$	$\chi^2(1) = 2.63$

Note: Coefficient estimates (log odds) are in parentheses. Random effects consisted of random intercepts of participant on accuracy. Face race and confidence were entered as fixed effects. Model 1 in Wilkison–Rogers notation is as follows: accuracy \sim face race + confidence + (1|participant); Model 2 is as follows: accuracy \sim (face race \times confidence) + (1|participant).

* $p < .05$. *** $p < .001$.

Overall, as results from the mixed-effects logistic regression mirror those from the ANOVA, we discuss only the ANOVA results moving forward.

The CA relationship for “new” responses was also examined. Using a Bonferroni correction of $\alpha = .0125$, results from dependent t tests indicated that same- and cross-race confidence-specific accuracy did not differ at any of the four confidence levels in Experiment 1 ($ps > .21$). The CA relationship for “new” responses in Experiment 1 is presented in [Figure 3](#) (top left panel).

Discussion

In summary, Experiment 1 results indicate that confidence is predictive of accuracy, more so than race of face, when calculated as the proportion of correct “old” responses. Furthermore, confidence-specific accuracy for same- and cross-race faces did not significantly differ at each of the four levels of confidence after a Bonferroni correction. Still, there was a small effect of race (slightly lower accuracy for cross-race judgments) that would be significant if a less stringent correction for multiple testing was used. Thus, our results are much like those reported by Dodson and Dobolyi (2015) using line-ups. Specifically, these results suggest that high-confidence cross-race identifications are as trustworthy as, or only slightly less trustworthy than, high-confidence same-race identifications. To see whether these results replicate on another dataset, in Experiment 2 we performed secondary analyses of data published by Blandón-Gitlin et al. (2014).

Experiment 2: Blandón-Gitlin et al. (2014)

Method

In Experiment 1 of their study, Blandón-Gitlin et al. (2014) tested 43 Caucasian male college students for

recognition memory of 50 White (same-race) and 50 Black (cross-race) faces, half old and half new. The study was a 2 (condition: placebo, oxytocin) \times 2 [race of target face: same-race (White faces) and cross-race (Black faces)] mixed design. Condition was between subjects; race of target face was within subjects. Each of the 50 (half White, half Black) study faces was presented individually for 2.5 s in the study phase. In the test phase, participants were presented 100 faces and made an old/new judgment for each along with a confidence rating on a scale of 1 (*not at all confident*) to 10 (*very confident*). Participants were instructed to use the full range of the scale. Although there was a reported CRE for the placebo but not the oxytocin condition, for the purposes of the current analyses, race of target face results were collapsed across the placebo and oxytocin conditions to have a sufficient sample size.²

Results

There was a significant difference in discrimination accuracy between same- and cross-race faces as measured by the signal detection measure d' , 1.89 and 1.56, respectively, $F(1, 41) = 14.22$, $p = .001$, $d = 0.58$. As in Experiment 1, the manipulation of race of target face in Experiment 2 was successful in eliciting at CRE based on measures of discrimination accuracy.

Next, we assessed confidence-specific accuracy across participants. A 2 (race of target face: same-race, cross-race) \times 7 (confidence: 1–4, 5, 6, 7, 8, 9, 10) repeated measures ANOVA was conducted on the proportion correct data. Similar to Experiment 1, there were too few observations at the lower levels of confidence on the 10-point scale, so we collapsed across Levels 1 through 4. The mean proportion correct for each condition is plotted in [Figure 4](#) (top panel). We also conducted analyses on the placebo and oxytocin groups separately. Mean proportion

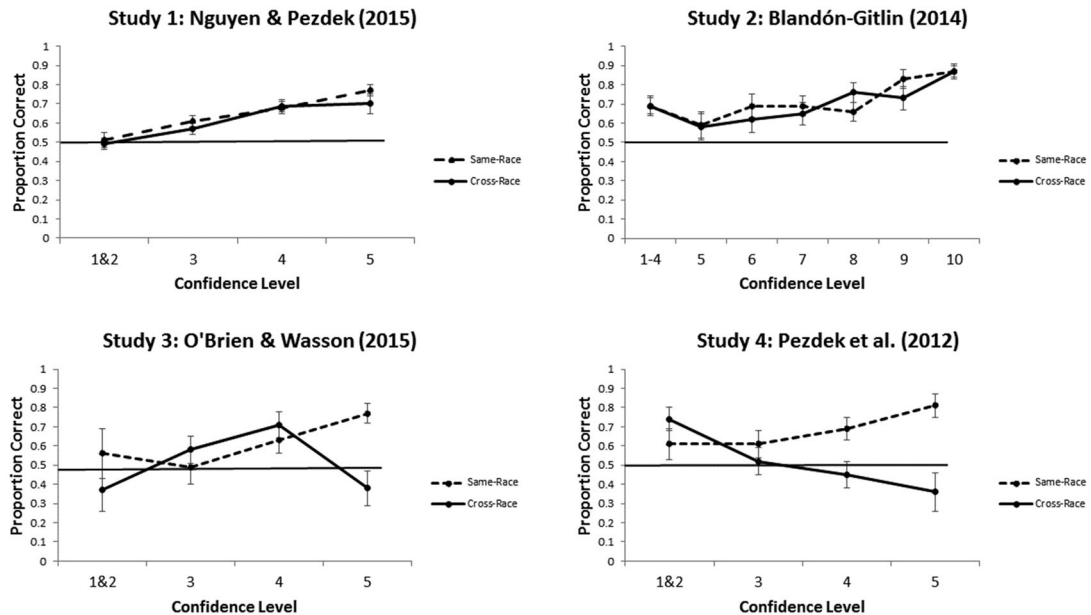


Figure 3. Confidence-specific accuracy for “new” responses assessed with confidence–accuracy characteristic (CAC) curves in Experiment 1, Experiment 2, Experiment 3, and Experiment 4. Chance performance is denoted by the horizontal line. Error bars represent standard error.

correct for the placebo and oxytocin groups are plotted in Figure 4 (bottom panels). The conclusions we drew from the data did not differ when we analysed the groups separately versus together; therefore, we report results from both groups together (see Footnote 2). Results from the ANOVA indicated no significant main effect of race of target face, $F(1, 42) = 2.01$, $p = .16$, $\eta_p^2 = .05$, power = .28, but a significant main effect of confidence level, $F(6, 37) = 53.09$, $p < .001$, $\eta_p^2 = .90$, power = 1.00. Confidence accounted for a much larger amount of variance in proportion correct than race of target face, $\eta_p^2 = .90$ and $\eta_p^2 = .05$, respectively. With a Bonferroni correction of $\alpha = .008$, the proportion correct was significantly higher at the highest level of confidence than at each of the other six lower levels of confidence (all $ps < .001$). These results replicate Experiment 1; overall, higher levels of confidence were associated with higher levels of accuracy for both same- and cross-race faces.

In contrast to Experiment 1, in Experiment 2, the interaction between race of target face and confidence level was not significant, $F(6, 37) = 0.57$, $p = .75$, $\eta_p^2 = .08$, power = .20. Nevertheless, we conducted dependent t tests on the proportion correct data at each of the seven levels of confidence. Similar to Experiment 1, we removed participants

who did not make at least two ratings at any specific confidence level; therefore, the sample sizes slightly differed across t tests (total $N = 43$). With a Bonferroni correction of $\alpha = .007$, the proportion correct did not significantly differ between same- and cross-race faces at Confidence Level 10, $t(42) = 2.22$, $p = .03$, power = .58, at Confidence Level 9, $t(37) = -0.15$, $p = .56$, power = .05, at Confidence Level 8, $t(39) = 1.41$, $p = .17$, power = .28, at Confidence Level 7, $t(36) = 1.05$, $p = .30$, power = .18, at Confidence Level 6, $t(33) = 1.49$, $p = .15$, power = .30, at Confidence Level 5, $t(28) = 0.70$, $p = .49$, power = .10, or at Confidence Level 1–4, $t(26) = 0.24$, $p = .81$, power = .06. Consistent with Experiment 1, confidence-specific accuracy did not differ between same- and cross-race faces at any confidence level; however, there was slightly lower accuracy for cross-race judgments that would be significant if a less stringent correction for multiple testing was used. Furthermore, the small sample sizes resulted in low power to detect significant effects. Thus, it is more important to compare effect sizes for these differences.

Figure 2 (top right panel) illustrates the effect size comparison for d' and confidence-specific accuracy for each level of confidence. Based on this comparison of effect sizes, it is clear that the magnitude of the CRE is larger for discrimination than

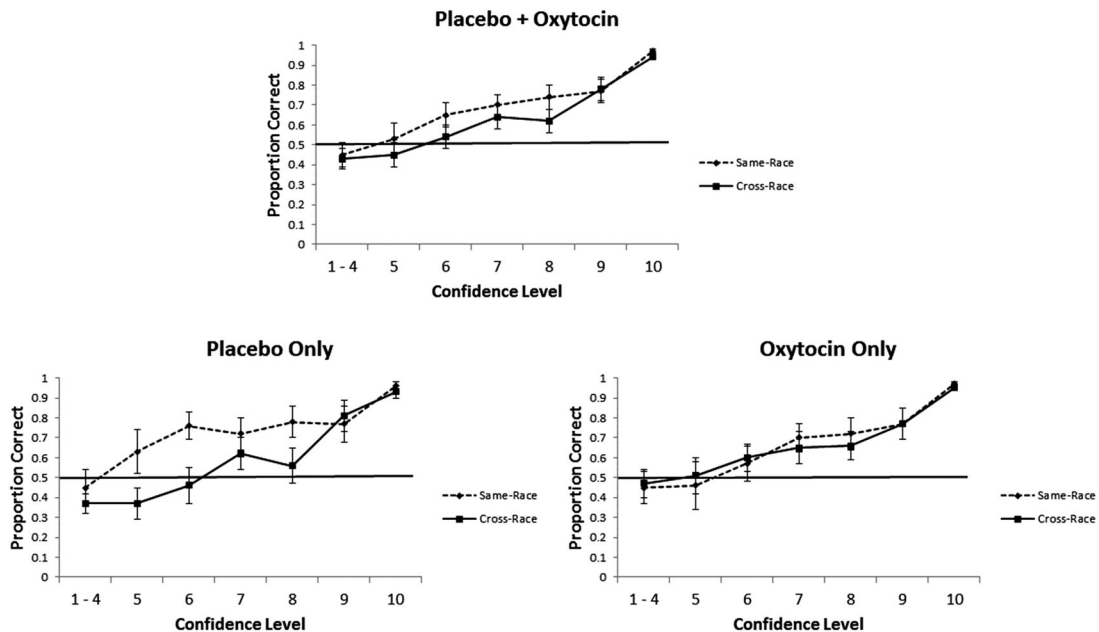


Figure 4. Confidence-specific accuracy assessed with confidence-accuracy characteristic (CAC) curves in Experiment 2. CAC curves for the placebo and oxytocin groups are presented separately and together. Chance performance is denoted by the horizontal line. Error bars represent standard error.

confidence-specific accuracy. The effect size for the highest level of confidence appears to be fairly large, but that is because the error bars associated with the proportion correct measures at high levels of accuracy are very small. The actual difference in high-confidence accuracy for same- and cross-race judgments is negligible, as shown in Figure 4 (proportions are .97 and .94, respectively). Thus, there may be a trend for high-confidence same-race judgments to be associated with higher accuracy than high-confidence cross-race judgments; however, the difference is negligible.

Using a Bonferroni correction of $\alpha = .0125$, dependent *t* tests on the CA relationship for “new” responses indicated that same- and cross-race confidence-specific accuracy did not differ at any of the seven confidence levels in Experiment 2 ($ps > .15$). The CA relationship for “new” responses in Experiment 2 is presented in Figure 3 (top right panel).

Discussion

As predicted, there was a reliable CA relationship for same- and cross-race faces in Experiment 2. The overall impression from these results is the same as

that of Experiment 1 and corresponds to what Dodson and Doholji (2015) reported using line-ups: Confidence is highly predictive of accuracy, face race may be slightly predictive as well (though it was not significant here), but the direction of any face race effect (when it exists) is that cross-race confidence-specific accuracy is slightly lower than same-race confidence-specific accuracy. These conclusions hold even though discrimination accuracy was substantially higher for same- than for cross-race faces. This is because discrimination accuracy and confidence-specific accuracy are different measures. To see whether these results replicate on another dataset, in Experiment 3 we performed secondary analyses of data published by O’Brien and Wasson (2015).

Experiment 3: O’Brien and Wasson (2015)

Method

In Experiment 1 of their study, O’Brien and Wasson (2015) tested 86 Caucasian college students for recognition memory of 16 White (same-race) and 16 Black (cross-race) faces, half old and half new. The study was a 3 (face presentation condition: individual, 3-face group without arrow, 3-face group with arrow) ×

2 [race of target face: same-race (White faces), cross-race (Black faces)] mixed factorial design. Face presentation was varied between subjects; race of target face was varied within subjects. The experiment was conducted in a classroom to groups of 16–20 participants. Each group was randomly assigned to an experimental condition. The faces were presented on a screen at the front of a room. In the study phase, participants were randomly assigned to view the 16 target faces (8 White and 8 Black) in one of three presentation conditions, varied between subjects. Each three-face group contained one target face and two foil faces of the same race as the target face (i.e., a homogeneous group). Throughout this study, participants were never tested on the two foil faces. In both of the group presentation conditions (with and without an arrow), each target appeared equally often in the first, second, or third position across all participants. In the individual presentation condition, each target face was presented for 5 s. In the group presentation conditions, the three-face group was presented for 15 s. The test phase immediately followed the study phase. In the test phase, each of the 32 test faces was presented individually and was randomly arranged. Participants made an old/new judgment for each along with a confidence rating on a scale of 1 (*not at all confident*) to 5 (*very confident*) on a paper packet. Participants were instructed to use the full range of the scale. To make Experiment 3 comparable to Experiments 1 and 2, only participants who were presented with faces individually rather than in a group were included in this analyses (final $N = 29$).

Results

There was a significant difference in discrimination accuracy between same- and cross-race faces as measured by the signal detection measure d' , 1.52 and 0.88, respectively, $F(1, 28) = 7.00$, $p = .01$, $d = 0.49$. Similar to Experiments 1 and 2, the manipulation of race of target face in Experiment 3 was successful in eliciting a CRE.

A 2 (race of target face: same-race, cross-race) \times 4 (confidence: 1 & 2, 3, 4, 5) repeated measures ANOVA was conducted on the proportion correct data. We collapsed across Levels 1 and 2 for all analyses because there were too few observations at the two lowest levels of confidence. The mean proportion correct for each condition is plotted in Figure 5. Consistent with the previous two studies, results from the Experiment 3 ANOVA indicated that

there was not a significant main effect of race of target face, $F(1, 28) = 0.76$, $p = .39$, $\eta_p^2 = .03$, power = .13, but there was a significant main effect of confidence level, $F(3, 26) = 33.52$, $p < .001$, $\eta_p^2 = .80$, power = 1.00. Confidence accounted for a much larger amount of variance in proportion correct than race of target face, $\eta_p^2 = .80$ and $\eta_p^2 = .03$, respectively. With a Bonferroni correction of $\alpha = .008$, the proportion correct at the highest level of confidence was significantly higher than the proportion correct at each of the other three lower levels of confidence (all $ps < .001$). In contrast, the proportion correct at the lowest confidence level (1 and 2) did not significantly differ from both Confidence Levels 3 and 4 (all $ps < .05$), and the proportion correct at Confidence Levels 3 and 4 did not significantly differ ($p = .77$).

The weak relationship between confidence and accuracy at the three lower levels of confidence in Experiment 3 is consistent with results reported by Weber and Brewer (2003), who reported that confidence is a reliable predictor of accuracy at or above chance levels but not below chance. Here, chance performance is a proportion correct of .50. Participants in Experiment 3 had slightly lower levels of confidence-specific accuracy than those in Experiments 1 and 2, and in Experiment 3, confidence-specific accuracy for same- and cross-race faces was around chance performance for the three lower levels of confidence.

In contrast to Experiment 1 but similar to Experiment 2, there was not a significant interaction between race of target face and confidence level, $F(3, 26) = 0.06$, $p = .98$, $\eta_p^2 = .01$, power = .06. Nevertheless, we conducted dependent t tests on the proportion correct data at each of the four levels of confidence. After removing participants who did not make at least two ratings at any specific confidence level, the sample sizes slightly differed across t tests (total $N = 29$). With a Bonferroni correction of $\alpha = .0125$, the proportion correct for same- and cross-race identifications did not significantly differ at Confidence Level 5, $t(27) = 0.26$, $p = .40$, power = .06, at Confidence Level 4, $t(26) = -0.95$, $p = .82$, power = .15, at Confidence Level 3, $t(24) = -0.49$, $p = .69$, power = .08, or at Confidence Levels 2 and 1, $t(18) = -0.23$, $p = .59$, power = .06. Although there was a significant interaction between face race and confidence in Experiment 1 but not in Experiment 2 or 3, pairwise comparisons in all three studies indicated that at all confidence levels, confidence-specific accuracy did not differ significantly between same- and cross-race faces.

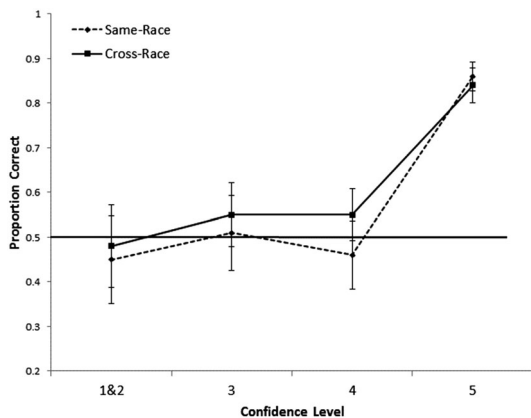


Figure 5. Confidence-specific accuracy assessed with confidence-accuracy characteristic (CAC) curves in Experiment 3. Chance performance is denoted by the horizontal line. Error bars represent standard error.

The sample size in Experiment 3 was small because we included only participants who were presented with individual faces. The small sample size may have resulted in low power to detect significant differences; therefore, it is useful to compare the effect size for each statistical test. Figure 2 (bottom left panel) illustrates the comparison of effect sizes between d' (discrimination accuracy) and confidence-specific accuracy. The effect size at all levels of confidence was small compared to the effect size for the difference in same- and cross-race discrimination accuracy, which replicates results of Experiments 1 and 2.

Using a Bonferroni correction of $\alpha = .0125$, dependent t tests on the CA relationship for “new” responses indicated that in Experiment 3, same- and cross-race confidence-specific accuracy did not significantly differ at the lower confidence levels ($ps > .02$) but did differ at the highest level of confidence ($p < .001$). High-confidence same-race judgments were associated with higher accuracy than high-confidence cross-race judgments. Furthermore, there was a less reliable CA relationship for “new” than for “old” responses. This may be attributed to overall lower accuracy in Experiment 3 than in Experiments 1 and 2. The CA relationship for “new” responses in Experiment 3 is presented in Figure 3 (bottom left panel).

Discussion

Together, the results from Experiments 1, 2, and 3 indicate that the CA relationship did not differ significantly as a function of race of face. In contrast to Experiments

1 and 2, in Experiment 3, confidence was not a reliable indicator of accuracy at the lower portion of the confidence scale, where performance was at or below chance. These results are consistent with those of Deffenbacher (1980) and Weber and Brewer (2003); the CA relationship is less reliable when performance is at or below chance. Observers who are merely guessing should not be more confident in a guess that resulted in a correct identification than a guess that resulted in an incorrect identification. Furthermore, observers who are guessing (i.e., whose memory signal is very weak) would have a more lax criterion for responding “old” and are thus predicted to be less confident in their responses than observers who make recognition judgments relying on more information in memory (i.e., strong feelings of familiarity or specific episodic information). Thus, accuracy is more likely to fluctuate around chance levels at lower levels of confidence than at higher levels of confidence. In fact, Weber and Brewer (2003) reported that although confidence reliably predicted accuracy in the upper portion of their 100-point confidence scale where performance was above chance levels, the CA relationship was less reliable for the lower portion of the 100-point scale because the lower portion was associated with chance performance or guessing.

This hypothesis was examined in Experiment 4 where overall recognition performance was relatively low compared to Experiments 1, 2, and 3. Discrimination accuracy for same- and cross-race faces, as measured by d' , was (a) 1.10 and 0.75, respectively, in our Experiment 1 by Nguyen and Pezdek (2015), (b) 1.89 and 1.56, respectively, in our Experiment 2 by Blandón-Gitlin et al. (2014), and (c) 1.52 and 0.88, respectively, in our Experiment 3 by O'Brien and Wasson. Discrimination accuracy for same- and cross-race faces, as measured by d' , was lower in our Experiment 4 by Pezdek et al. (2012), 0.50 and 0.16, respectively. The d' values close to zero indicate chance performance. Thus it was predicted that in Experiment 4, confidence would be a reliable indicator of accuracy only when recognition accuracy was above chance.

Experiment 4: Pezdek et al. (2012)

Method

In Experiment 1 of their study, Pezdek et al. (2012) tested 44 Caucasian college students for recognition memory of 16 White (same-race) and 16 Black

(cross-race) faces, half old and half new. The study was a 2 (face presentation condition: individual, group) \times 2 [race of target face: same-race (White faces), cross-race (Black faces)] within-subjects factorial design. To make Experiment 4 comparable to the previous three studies, only responses to faces presented individually, rather than in a group, were included in Experiment 4 analyses. Each of the eight study faces (four White and four Black) in the individual face presentation condition was presented for 5 s in the study phase, with an interstimulus interval of 1.5 s. In the test phase, participants were presented with 32 faces and made an old/new recognition judgment for each along with a confidence rating on a scale of 1 (*not at all confident*) to 5 (*very confident*). Participants were instructed to use the full range of the scale. The total number of test faces per race of face was smaller in this study (four) than in Experiments 1 (24), 2 (25), and 3 (eight); therefore, the mean proportion correct for each individual may be less stable.

Results

There was a significant difference in discrimination accuracy between same- and cross-race faces as measured by d' , 0.50 and 0.16, respectively, $F(1, 43) = 5.21$, $p = .03$, $d = 0.34$. Similar to the previous three studies, the manipulation of race of target face in Experiment 4 was successful in eliciting a CRE based on measures of discrimination accuracy.

For the CAC analysis, a correction to the number of false alarms was made in this study because the face presentation manipulation was within subjects; therefore, a separate false-alarm rate could not be calculated for faces presented individually. For each participant, we divided the number of false alarms per confidence level by two to account for the fact that there were twice as many possible false alarms as there were possible hits. Again, because there were too few observations at the two lowest levels of confidence, we collapsed across Levels 1 and 2 for all analyses.

A 2 (race of target face: same-race, cross-race) \times 4 (confidence: 1 & 2, 3, 4, 5) repeated measures ANOVA was conducted on the proportion correct data. The mean proportion correct for each condition is plotted in Figure 6. Results from the ANOVA indicated no significant main effect of race of target face, $F(1, 43) = 0.63$, $p = .43$, $\eta_p^2 = .01$, power = .12, but there was a significant main effect of confidence

level, $F(3, 41) = 10.40$, $p < .001$, $\eta_p^2 = .43$, power = 1.00. Confidence accounted for a much larger amount of variance in proportion correct than race of target face ($\eta_p^2 = .43$; $\eta_p^2 = .01$, respectively). With a Bonferroni correction of $\alpha = .008$, all pairwise comparisons among the four levels of confidence were not significant ($ps > .10$). Although not statistically significant, replicating Experiments 1, 2, and 3, the highest level of confidence was associated with higher mean accuracy than the other three levels of confidence.

This ANOVA yielded a significant interaction between race of target face and confidence level, $F(3, 41) = 4.79$, $p = .01$, $\eta_p^2 = .26$, power = .87. We conducted dependent t tests on the proportion correct data for each of the four levels of confidence. After removing participants who did not make at least two ratings at any confidence level, the sample sizes slightly differed across t tests (total $N = 44$). A Bonferroni correction of $\alpha = .0125$ was used. In contrast to Experiments 1, 2, and 3, in Experiment 4 a dependent t test indicated that the proportion correct was significantly higher for same- than for cross-race faces at Confidence Level 5, $t(29) = 3.77$, $p < .001$, power = .95. The proportion correct did not significantly differ between same- and cross-race faces at Confidence Level 4, $t(33) = -0.17$, $p = .57$, power = .05, at Confidence Level 3, $t(34) = 0.77$, $p = .22$, power = .12, or at Confidence Levels 1 and 2, $t(20) = -0.86$, $p = .80$, power = .13. These pairwise comparisons mirror the significant interaction between race of target face and confidence level: Same- and cross-race confidence-specific accuracy did not differ at each of the three lower levels of confidence but same-race confidence-specific accuracy was significantly higher than cross-race confidence-specific accuracy at the highest level of confidence. Figure 2 (bottom right panel) illustrates the effect size comparison for d' and confidence-specific accuracy for each level of confidence. In contrast to the previous three studies, the effect size for differences between same- and cross-race confidence-specific accuracy at the highest level of confidence was larger than the effect size for d' .

Using a Bonferroni correction of $\alpha = .0125$, dependent t tests on the CA relationship for "new" responses indicated that in Experiment 4, same- and cross-race confidence-specific accuracy did not significantly differ at the lower confidence levels ($ps > .02$) but did differ at the highest confidence level ($p = .002$). High-confidence same-race judgments were associated with higher accuracy than high-confidence cross-race judgments. Furthermore, there was a less

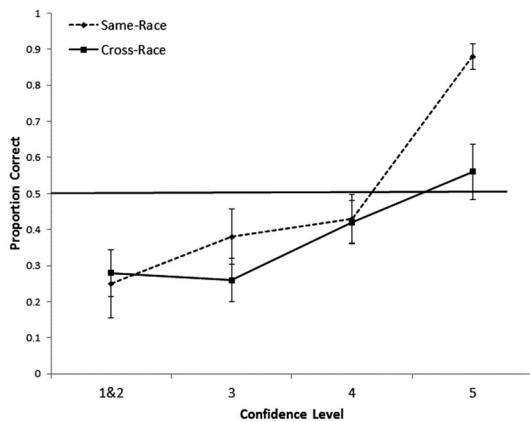


Figure 6. Confidence-specific accuracy assessed with CAC curves in Experiment 4. Chance performance is denoted by the horizontal line. Error bars represent standard error.

reliable CA relationship for “new” than for “old” responses. Similar to Experiment 3, this difference may be attributed to lower overall performance in Experiment 4 than in Experiments 1 and 2. The CA relationship for “new” responses in Experiment 4 is presented in Figure 3 (bottom right panel).

Discussion

In Experiment 4, confidence-specific accuracy did not differ as a function of race of face at the lower levels of confidence, but same-race confidence-specific accuracy was significantly higher than cross-race confidence-specific accuracy at the highest level of confidence. In contrast to the previous three studies, in Experiment 4, high levels of confidence were indicative of high levels of accuracy for same-race faces but only slightly above chance levels for cross-race faces. The absence of a reliable CA relationship for cross-race faces may be attributed to overall chance performance, resulting in a significant difference in trustworthiness (i.e., the proportion of correct “old” responses) for same- and cross-race faces at the highest level of confidence. Furthermore, slightly above-chance cross-race accuracy at the highest level of confidence may also be indicative of liberal responding. In contrast to the previous three studies, results from Experiment 4 suggest that high-confidence same-race identifications may be more trustworthy than high-confidence cross-race identifications when observers are merely guessing in the cross-race condition.

General discussion

Much of the research on same- and cross-race face memory posits that the CRE is a result of qualitative differences at encoding (Goldinger et al., 2009; Hills & Lewis, 2006; Hugenberg et al., 2010; Meissner et al., 2005). Encoding of same-race faces is more likely to involve effortful elaboration of specific diagnostic information than encoding of cross-race faces. Thus, observers are better able to rely on recollective processes and retrieve specific episodic information during recognition of same- than recognition of cross-race faces, resulting in higher overall discrimination accuracy for same- than for cross-race faces. According to Wixted and Mickes’s (2010) continuous dual-process model of recognition memory, however, memories based more on recollective processes do not always have higher old/new discrimination accuracy than memories based more on familiarity processes. They reported a dissociation between quality (content) and quantity (strength) of memory, typically measured with remember/know judgments and confidence ratings, respectively. When comparing strong, high-confidence recollection- and familiarity-based memories that are similar in memory strength (i.e., they had the same confidence rating), there was no significant difference in old/new discrimination accuracy (Wixted & Mickes, 2010). These results suggest that the reliability of the CA relationship may not depend on the type of processing (i.e., recollection or familiarity) because confidence ratings are indicative of the strength of the memory rather than the quality of the memory.

If, according to results from Wixted and Mickes (2010), strong, high-confidence memories, which can be recollection or familiarity based, do not differ in recognition accuracy, then high-confidence same-race judgments and high-confidence cross-race judgments may not differ in accuracy. Results from the four CAC analyses indicate that confidence judgments made for “old” responses immediately following the initial test phase reliably predict recognition accuracy when performance is above chance levels (Experiments 1, 2, and 3) but not when performance is at chance levels (Experiment 4). These results are consistent with findings of Deffenbacher (1980) and Weber and Brewer (2003). When there is no memory signal, observers merely guess, and, consequently, confidence is not predicted to be indicative of accuracy. Furthermore, confidence-specific accuracy for same- and cross-race faces did

not significantly differ at each level of confidence when overall performance was above chance levels (Experiments 1, 2, and 3) but significantly differed when overall performance was at chance levels for cross-race faces (Experiment 4). In other words, when observers are not performing at chance levels, a high-confidence cross-race identification can be as trustworthy as a high-confidence same-race identification.

Applying these results to the legal system, it is important to note that the eyewitness cases that are more likely to proceed to trial are those with high-confidence eyewitnesses. Accordingly, our findings would suggest that judges and jurors should not be hasty to judge cross-race identifications as unreliable, especially when eyewitness confidence was obtained close in time to the identification, and performance is above chance levels. This is so even though it is also true that overall same- and cross-race recognition accuracy differs considerably as measured by d' . In other words, the existence of a CRE on d' (which is what the CRE typically refers to) does not necessarily imply that confidence-specific accuracy differs significantly for same- and cross-race faces. Although there was slightly higher accuracy for high-confidence same- than cross-race judgments in our studies (which would have been significant if a less stringent correction for multiple testing was used), the difference in accuracy at high confidence is nonetheless small. As shown in [Figure 2](#), when recognition memory performance is above chance, the magnitude of the CRE is smaller at high levels of confidence than when measured by d' .

Our results are similar to those reported by Dodson and Dobolyi (2015). Although they reported statistically significant higher calibration scores for same- than for cross-race judgments, the size of this difference was small ($d = .31$ for Black participants and $d = .10$ for White participants). Furthermore, the proportion correct for high-confidence same- and cross-race judgments were similar, 80% and 77%, respectively. Dodson and Dobolyi had a large sample size ($N = 1,656$) and thus more power to detect even small effects. Some of our reported comparisons may have been statistically significant if we had used a more liberal alpha correction or had a larger sample size. However, the focus should be on the effect sizes, which are fairly similar across our two studies. Although there may be a real difference between same- and cross-race face accuracy across all levels of confidence, differences in confidence-

specific accuracy were small compared to differences in discrimination accuracy. As confidence-specific accuracy is a more informative measure to triers of fact than discrimination accuracy (Mickes, 2015), triers of fact should not be quick to judge a cross-race identification as unreliable compared to a same-race identification.

Across all four studies, confidence was a stronger predictor of accuracy than face race. In fact, results from the mixed-effects logistic regression indicated that when confidence is equated, face race is no longer a significant predictor of accuracy. Furthermore, our results suggest that the CA relationship does not differ as a function of face race. These four studies, together with the results of Palmer et al. (2013) and Dodson and Dobolyi (2015), suggest that when confidence is collected immediately after the initial recognition memory test, and performance is above chance levels, observers are able to metacognitively adjust their subjective confidence ratings to account for poor encoding conditions. For example, witnesses may know that they only caught a glimpse of the perpetrator's face and thus do not rate their confidence as high unless they are absolutely sure. Thus, confidence can be a reliable indicator of accuracy for both same- and cross-race faces when overall performance is above chance levels, and confidence is collected immediately after the initial identification. This is one explanation for the difference in same- and cross-race discrimination accuracy in light of the similarity in same- and cross-race confidence-specific accuracy reported in these four studies.

The low recognition memory performance for cross-race faces in our Experiment 4, utilizing data from Pezdek et al. (2012; cross-race mean $d' = 0.16$), may explain the less reliable CA relationship for cross-race than same-race faces; high confidence was indicative of high accuracy for same-race faces but only indicative of slightly above-chance performance for cross-race faces. This level of overconfidence for cross-race faces at the highest confidence level may be indicative of a high false-alarm rate and liberal responding. This liberal responding in confidence ratings may have resulted from the within-subjects manipulation of individual versus group presentation of faces by Pezdek et al. (2012). In that study, discrimination accuracy for cross-race faces, but not same-race faces, decreased to chance levels when target faces were presented in a group (with two companion faces) rather than individually. This suggests that cross-race target faces presented in a

group were not encoded well, and therefore all appeared to be new faces when later presented in the test phase, when, in fact, half of the test faces were old, and half were new. When taking a multi-item recognition test, participants may not have wanted to respond “new” on all test faces, and, thus, incorrectly responded “old” to some new cross-race test faces. Participants may have incorrectly rated some new test faces with high confidence due to strong feelings of familiarity if the target and test faces shared similar characteristics.

These considerations suggest that it will be important to test these issues using an eyewitness show-up paradigm where observers only make one judgment. In the four studies reported, observers made multiple judgments. Observers may be able to improve their ability to calibrate their confidence ratings as testing proceeds and, as a result, inflate the reliability of the CA relationship (Brewer, 2006). The same concern might apply to the findings reported by Dodson and Dobolyi (2015), who reported very similar results to ours using a line-up procedure. In their study, participants each made 12 memory judgments. Despite these limitations, results from these studies help us better understand the recognition memory framework and provide support for the dissociation between quality (content) and quantity (strength) of a memory trace (Wixted & Mickes, 2010). The absence of a CRE with confidence-specific accuracy suggests that it is important to consider both quality and quantity when assessing recognition accuracy.

One might argue that because these four studies were not fully crossed with a second racial group of participants, the difference between discrimination accuracy for same- and cross-race faces may be due to the nature of the particular stimuli used. For example, perhaps overall, the White faces were more memorable than the Black faces. In all four studies, however, response bias, measured with the signal detection measure C , was more conservative (i.e., less likely to say “old”) for same- than for cross-race faces. This is consistent with previous CRE research that reports that observers are more liberal in responding to cross- than to same-race faces (Meissner & Brigham, 2001). In addition, Meissner et al. (2005) tested both White and Black participants on a recognition memory task using photos from the same database as the one that we used and did not find a significant main effect of face race on discrimination accuracy. This suggests that the White faces in

this stimulus set are not more memorable than the Black faces. Furthermore, this alternative explanation would not account for the difference in the magnitude of the CRE between discrimination and confidence-specific accuracy. Finally, it is worth noting that Wixted, Mickes, Dunn, Clark, and Wells (2016) recently reported evidence of a strong CA relationship in real eyewitnesses tested using photo line-ups in the Houston Police Department. Most eyewitness identifications in the jurisdiction are cross-race identifications, yet high confidence accuracy in that study was estimated to be very high (greater than 95% correct).

In addition to the theoretical contribution of this research, there are many applications of these results to eyewitness memory. An estimated 76% of DNA exonerated cases can be attributed to eyewitness misidentifications (Garrett, 2011). However, and consistent with the results of our study, Garrett (2011) reviewed trial materials for 161 DNA exonerated cases and reported that 57% of innocent suspects were misidentified by eyewitnesses who expressed low levels of certainty during their initial identification. The results from our studies and others (Mickes, 2015; Palmer et al., 2013; Weber & Brewer, 2003) provide evidence that under certain conditions, the effects of estimator variables like race of face and exposure duration do not affect the reliability of the CA relationship. In other words, when confidence is collected immediately after the initial identification, and performance is above chance, a high-confidence cross-race identification can be as trustworthy as a high-confidence same-race identification. At the very least, under these conditions, the magnitude of the CRE (assessed by Cohen’s d) is attenuated when confidence is taken into account.

Notes

1. One concern about running ANOVAs on the proportion correct data is that the two independent variables of race of target face and confidence may be non-orthogonal. To account for this possibility, for each participant in each of the four studies, we ran a χ^2 test of independence on the number judgments made for same- and cross-race faces at each level of confidence. Based on these analyses, we removed participants using a cut-off of $\alpha = .05$ on the χ^2 test of independence and then reran the ANOVAs. The number of participants removed per study was as follows: Experiment 1 = 2, Experiment 2 = 2, Experiment 3 = 1, and Experiment 4 = 2. With these participants removed, there were no differences in the statistical significance of any main effects or

interactions in all four studies. Next, to decrease our chances of Type II error, we removed participants using a cut-off of $\alpha = .10$ on the χ^2 test of independence and re-ran the ANOVAs. The number of additional participants removed per study was as follows: Experiment 1 = 7, Experiment 2 = 6, Experiment 3 = 2, and Experiment 4 = 2. With these additional participants removed, there were no differences in the statistical significance of any main effects or interactions in Experiments 2, 3, and 4. The main effect of race was significant ($p = .02$) in Experiment 1; therefore, we re-ran the t tests comparing the proportion correct at each confidence level after removing the 7 participants who had significant χ^2 test of independence using $\alpha = .10$. The statistical significance of the t tests did not differ with these participants removed. Given that the results of the ANOVAs and t tests did not differ after removing these participants, in the manuscript, we report the ANOVA results on all participants.

- The CA relationship was examined separately for the placebo and oxytocin conditions and was similar to the CA relationship for both groups combined. Before an alpha correction, there were statistically significant differences in same- and cross-race confidence-specific accuracy at confidence levels of 5 and 6 for the placebo group alone and at level 10 for the oxytocin group alone ($ps < .05$). There were no other significant differences. To be consistent with the other reported studies, we used a Bonferroni correction of $\alpha = .007$, and these differences are no longer statistically significant. More critically, examining the mean proportions from the oxytocin group at level 10 indicated that high-confidence same-race judgments had an average accuracy of .97 whereas high-confidence cross-race judgments had an average accuracy of .95. There may be a real difference between high-confidence same- and cross-race judgments, but the difference is small.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, 74(2), 81–99. doi:10.1037/h0029531
- Blandón-Gitlin, I., Pezdek, K., Saldivar, S., & Steelman, E. (2014). Oxytocin eliminates the own-race bias in face recognition memory. *Brain Research*, 1580, 180–187. doi:10.1016/j.brainres.2013.07.015
- Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology*, 11, 3–23. doi:10.1348/135532505X79672
- Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, 4(4), 243–260. doi:10.1007/BF01040617
- Deffenbacher, K. A., Bornstein, B. H., McGorty, E. K., & Penrod, S. D. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied*, 14(2), 139–150. doi:10.1037/1076-898X.14.2.139
- Dodson, C. S., & Dobolyi, D. G. (2015). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology*. doi:10.1002/acp.3178
- Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory and Cognition*, 24, 523–533. doi:10.3758/BF03200940
- Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, 111(2), 524–542. doi:10.1037/0033-295X.111.2.524
- Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press.
- Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(5), 1105–1122. doi:10.1037/a0016548
- Hills, P. J., & Lewis, M. B. (2006). Reducing the own-race bias in face recognition by shifting attention. *The Quarterly Journal of Experimental Psychology*, 59(6), 996–1002. doi:10.1080/17470210600654750
- Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review*, 117(4), 1168–1187. doi:10.1037/a0020463
- Ingram, K. M., Mickes, L., & Wixted, J. T. (2012). Recollection can be weak and familiarity can be strong. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(2), 325–339. doi:10.1037/a0025483
- Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, 7(4), 330–334. doi:10.1037/h0028434
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87(3), 252–271. doi:10.1037/0033-295X.87.3.252
- 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, and 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. doi:10.1002/acp.1097
- Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, 4(2), 93–102. doi:10.1016/j.jarmac.2015.01.003
- Nguyen, T. B., & Pezdek, K. (2015, June). Does memory for same- and cross-race faces benefit from more "on-time" and "off-time?". Paper presented at the meeting of the Society of Applied Research in Memory and Cognition, Victoria, BC.
- O'Brien, M., & Wasson, C. (2015, May). *Social-cognitive processes and recognition of same- and cross-race faces*. Poster presented at the meeting of the Association for Psychological Science, New York, NY.
- Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, 19(1), 55–71. doi:10.1037/a0031602

- Peirce, J. W. (2007). PsychoPy—Psychophysics software in python. *Journal of Neuroscience Methods*, 162(1-2), 8–13. doi:10.1016/j.jneumeth.2006.11.017
- Pezdek, K., O'Brien, M., & Wasson, C. (2012). Cross-race (but not same-race) face identification is impaired by presenting faces in a group rather than individually. *Law and Human Behavior*, 36(6), 488–495. doi:10.1037/h0093933
- 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.
- 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(3), 490–499. doi:10.1037/0021-9010.88.3.490
- Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, 117(4), 1025–1054. doi:10.1037/a0020874.
- Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, 4(4), 329–334. doi:10.1016/j.jarmac.2015.08.007
- Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences*, 113, 304–309.