A Decision Processes Account of the Differences in the Eyewitness Confidence-Accuracy Relationship between Strong and Weak Face Recognizers under Suboptimal Exposure and Delay Conditions

Jessica Nicole Gettleman

Paradise Valley, AZ

Bachelor of Arts, University of Missouri - Kansas City, 2017

A Thesis presented to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Master of Arts

Department of Psychology

University of Virginia

December, 2020

Committee:

Dr. Chad S. Dodson

Dr. Nicole M. Long

Abstract

When pristine testing conditions are used, an eyewitness's high-confidence identification from a lineup can be a reliable predictor of their identification accuracy (Wixted & Wells, 2017). Further, Grabman, Dobolyi, Berelovich, and Dodson (2019) found that high-confidence identifications are more predictive of accuracy for individuals with stronger than weaker face recognition ability. We extend this research by investigating why strong face recognizers make more informative confidence judgments and fewer high-confidence errors through the framework of two different accounts: the optimality account (Deffenbacher, 1980) and the decision processes account (e.g., Kruger & Dunning, 1999). The optimality account holds that differences in the predictive value of confidence ratings made by strong versus weak face recognizers are a result of differences in the quality of their memory representations for faces, indicating that confidence-accuracy calibration would be equated between these two groups when overall accuracy is equated, while the decision processes account attributes differences in calibration to strong face recognizers' superior metacognitive skills, which allow them to better evaluate their performance in the domain of face recognition. Therefore, to distinguish between these accounts, we manipulated exposure and retention interval to create conditions that produced comparable levels of identification accuracy between stronger and weaker face recognizers, and then examined their confidence-accuracy calibration. The decision processes account was supported, as differences in calibration between stronger and weaker face recognizers persisted even when overall identification accuracy was equated. Stronger face recognizers are better able to regulate their use of the confidence scale points with changes in identification accuracy.

Keywords: eyewitness identification, confidence, face recognition, calibration

A Decision Processes Account of the Differences in the Eyewitness Confidence-Accuracy Relationship between Strong and Weak Face Recognizers under Suboptimal Exposure and Delay Conditions

In what are known as the *Biggers* criteria, the Supreme Court outlined five determinants of the admissibility and reliability of an eyewitness identification. One criterion, the level of certainty demonstrated by the witness (*Neil v Biggers*, 1972), is frequently addressed in eyewitness memory research. Research generally supports the Court's inclusion of this criterion, as it has been shown that, under certain conditions, an eyewitness's high-confidence identification from a police lineup can be highly predictive of their accuracy (Wixted & Wells, 2017). When the identification is made from a fair lineup (i.e., where a naïve eyewitness is no more likely to select the suspect than a filler), this finding is robust to a variety of manipulations (e.g., the presence vs. absence of a weapon, same-race vs. cross-race identifications) and is consistent across both lab and field studies.

However, the predictive value of confidence ratings assigned to lineup identifications could differ depending on the face recognition ability of the individual witness. Research on face recognition ability indicates that the ability to recognize unfamiliar faces varies substantially across individuals (see Wilmer, 2017 for a review). Some individuals have developmental prosopagnosia, or face-blindness, and are often unable to recognize faces of close friends and family members, while others, termed "super-recognizers," are exceptionally skilled in this domain (Russell, Duchaine & Nakayama, 2009; Wan et al., 2017). Further, face recognition is an ability that is weakly correlated with measures of memory or general intelligence (Gignac, Shankaralingam, Walker, & Kilpatrick, 2016; Goldstein, Johnson, & Chance, 1979; Hildebrandt, Wilhelm, Schmiedek, Herzmann, & Sommer, 2011; Shakeshaft & Plomin, 2015; Wilhelm et al., 2010, Wilmer, 2017; Wilmer et al., 2012; Zhu et al., 2010) and is highly heritable (Shakeshaft & Plomin, 2015; Wilmer at al., 2010; Zhu et al., 2010).

Recently, Grabman, Dobolyi, Berelovich, and Dodson (2019) showed that confidence ratings were much more predictive of the accuracy of lineup identifications when they were provided by stronger than weaker face recognizers. Using a standard eyewitness lineup paradigm, these researchers found that weak face recognizers are more likely than strong face recognizers to make high-confidence misidentifications. We extend this research by providing an explanation of why face recognition ability is associated with the informativeness of confidence ratings. Specifically, we will distinguish between two potential accounts: the optimality account and the decision processes account.

Deffenbacher's (1980) optimality account argues that the predictive value of confidence increases when the encoding and retrieval conditions for the associated decision (e.g., the lineup identification) are increasingly optimal. This account suggests that face recognition ability moderates the confidence-accuracy relationship by influencing the quality of the information that is encoded and retrieved. Differences in the quality of the representation determine the predictive value of confidence. Stronger face recognizers have richer memory representations and consequently make confidence ratings with high predictive value. Likewise, weaker face recognizers have impoverished representations of faces and therefore their confidence ratings have little predictive value. Thus, the key prediction of this account is that if the quality of the memory representation determines the confidence-accuracy relationship, then stronger and weaker face recognizers should behave similarly – i.e., similar calibration of confidence to accuracy – under conditions that produce similar levels of identification accuracy.

An alternative account holds that differences in the confidence-accuracy relationship between weaker and stronger face recognizers are fueled by the different decision processes these two groups use to evaluate what they remember and how they assign confidence ratings. According to this decision account, expertise in a given domain allows individuals to evaluate their performance in that domain more precisely (Kruger & Dunning, 1999; see also Bol, Hacker, O'Shea, & Allen, 2005; Chi, 1978; Moreland, Miller, & Laucka, 1981). Therefore, the expertise or enhanced ability of stronger face recognizers gives their confidence ratings more predictive value than those of weaker face recognizers. As a result, this account predicts better calibration for stronger face recognizers even when identification accuracy is comparable between stronger and weaker face recognizers. In other words, the decision account suggests better ability in a particular domain will lead to better metacognitive ability in that domain, emphasizing the contribution of metacognition to the confidence-accuracy relationship.

In short, the optimality and decision accounts make different predictions about the confidence-accuracy relationship when identification accuracy is comparable between individuals of different face recognition ability. To distinguish between these two accounts, it is necessary to create conditions that produce comparable levels of identification accuracy between individuals of weaker and stronger face recognition ability. We do this by manipulating both: a) the delay between viewing target faces and making lineup identifications, and b) the frequency of participants' exposure to the target faces.

The finding that eyewitness memory performance worsens with increasing delay is well-established and widely replicated, using delays ranging from one hour to nine months (e.g., Brewer, Weber, & Guerin, 2019; Juslin, Olsson, & Winman, 1996; Lin, Strube, & Roediger, 2019; Palmer, Brewer, Weber, & Nagesh, 2013; Read, Lindsay, & Nichols, 1998; Sauer, Brewer, Zweck, & Weber, 2010). Despite these changes in accuracy with delay, many studies find the confidence-accuracy relationship remains strong, especially for responses made with high confidence (Semmler, Dunn, Mickes, & Wixted, 2018; Wixted, Read, & Lindsay, 2016).

In their field study, Sauer et al. (2010) had participants view a target person for 10 seconds from 10 meters away. Participants in the immediate condition were then presented with a target-present or target-absent lineup and asked to select the face of the person they just saw or indicate that the person was not present and then to rate their confidence in their decision. Participants in the delay condition took between 20 and 50 days to respond to an email, from which they completed this same task online. Sauer and colleagues found greater accuracy in the immediate than the delay condition both when a face was selected from a lineup (62% versus 47%, respectively) and when ‘not present’ was selected (82% versus 66%), indicating that the delay manipulation clearly affected performance. Further, though the confidence-accuracy relationship generally remained stable, there were notable differences between the immediate and delay conditions. Compared to the immediate condition, the delay condition exhibited greater overconfidence and decreased diagnosticity at each level of confidence. Finally, using delays ranging from 10 minutes to 10 days, Lin et al. (2019) found that while participants made fewer correct positive identifications after longer than shorter delays, there was a strong confidence-accuracy relationship that persisted even after the longest delay. Overall, confidence is still predictive of accuracy with longer delays, but the accuracy of identifications and probative value of confidence ratings decline. One limitation in this literature is a paucity of studies using three or more retention intervals, which is necessary to establish the trajectory of the forgetting curve. Including three retention intervals also permits a more forensically relevant examination of changes in eyewitness memory with delay as identifications are never made immediately after a crime, and the length of the delay between a crime and an identification varies widely. We use retention intervals of five minutes, one day, and one week.

Exposure duration is typically operationalized as the amount of time participants are exposed to the target face. Manipulations of five seconds (short exposure duration) versus 90 seconds (long exposure duration) are common, showing improved identification accuracy in the longer condition (Bornstein, Deffenbacher, Penrod, & McGorty, 2012; Palmer et al., 2013; see also Semmler et al., 2018). However, despite changes in overall accuracy, Palmer et al. (2013) found that the confidence-accuracy relationship does not change as a result of exposure duration. Following Dobolyi and Dodson (2013), we use repetition as a proxy measure for exposure duration, manipulating the number of times participants were shown the target face (one time vs. four times).

To summarize, the goal of the current experiment is to test two accounts of the effects of face recognition ability on the confidence-accuracy relationship. All extant studies that have compared the confidence-accuracy relationship between stronger and weaker face recognizers confound differences in accuracy with differences in face recognition ability. Because stronger face recognizers tend to be more accurate than weaker face recognizers, it has been difficult to determine whether differences in the confidence-accuracy relationship are due to differences in accuracy, differences in face recognition ability, or a combination of these factors. We solve this problem by manipulating delay and exposure to create a variety of levels of identification performance which allow us to determine whether differences in the confidence-accuracy relationship are best explained by differences in accuracy or differences in face recognition ability. There is a longstanding tradition of testing separate accounts of individual differences in one dimension of memory performance by controlling and matching performance on another potentially confounding dimension of memory (e.g., Carpenter & Schacter, 2018; Dodson, Bawa, & Krueger, 2007; Jacoby et al., 2005; Morcom, Li, & Rugg, 2007). When performance is equated, the optimality account argues that differences in face recognition ability will have little influence on the confidence-accuracy relationship. In contrast, the decision account predicts that stronger face recognizers will show a more robust confidence-accuracy relationship than weaker face recognizers even under conditions of comparable identification performance. In addition to allowing us to create performance-matched conditions, these delay and repetition manipulations are used, in concert with confidence ratings, face recognition ability, and another forensically relevant variable – decision time (e.g., Sauerland & Sporer, 2009) – to predict identification accuracy, as modeling the conjunctive effects of these estimator variables in a single model is a secondary goal of this study.

Method

Participants

Participants were 1046 individuals between the ages of 19 and 78 (*M* = 39.42, *SD* = 11.77; 50% female) who were recruited using Turkprime (Litman, Robinson, & Abberbock, 2017) to interface with Amazon's Mechanical Turk (mTurk) in exchange for payment. No consensus standards are available for a-priori power estimates for mixed effects logistic regression models, but given conservative recommendations of 50 responses per modeled variable (van der Ploeg, Austin, & Steyerberg, 2014) as well as findings that estimates are generally reliable for sample sizes greater than 30 with at least 10 responses per participant (McNeish & Stapleton, 2016), this sample size is deemed sufficient. The University of Virginia Institutional Review Board (IRB) approved this research.

Materials

Lineups. Participants viewed 12 lineups. These lineups consisted of formal photographs of six individuals with neutral facial expressions wearing identical maroon-colored t-shirts (Meissner, Brigham, & Butz, 2005), presented in a 2 x 3 array. All lineups were fair, meaning no face stood out and all faces were equally likely to be selected by a viewer naïve to the target (see Dobolyi & Dodson, 2013 for more details on lineup generation). Specifically, for all lineups, we calculated Tredoux's (1999) *E*, a score that ranges from 1 to the nominal size of the lineup, and estimates the functional size of a lineup. A perfect score is equal to the number of faces in the lineup: for example, a score of six for a six-person lineup. The average *E*-score for the 12 lineups is 4.48 (95% CI = 4.09, 4.87). To ensure the lineup identification task relied on face recognition (and not picture-matching), at encoding, participants saw different photos of potential lineup targets, depicting the targets in street clothing and with casual expressions (e.g., smiling).

Face Recognition Task. We administered the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) to assess participants' face recognition ability. This task includes 72 trials, presented in three increasingly difficult blocks, and is scored by taking a simple sum of the correct responses (scores range from 0-72, with chance = 24). In all blocks, participants are asked to select the face they previously viewed from among three choices, and lighting changes as well as visual noise are introduced in the latter blocks to increase difficulty. Figure 1 shows the distribution of CFMT scores from the present study. Internal reliability for the CFMT was high (Cronbach's α = .92).

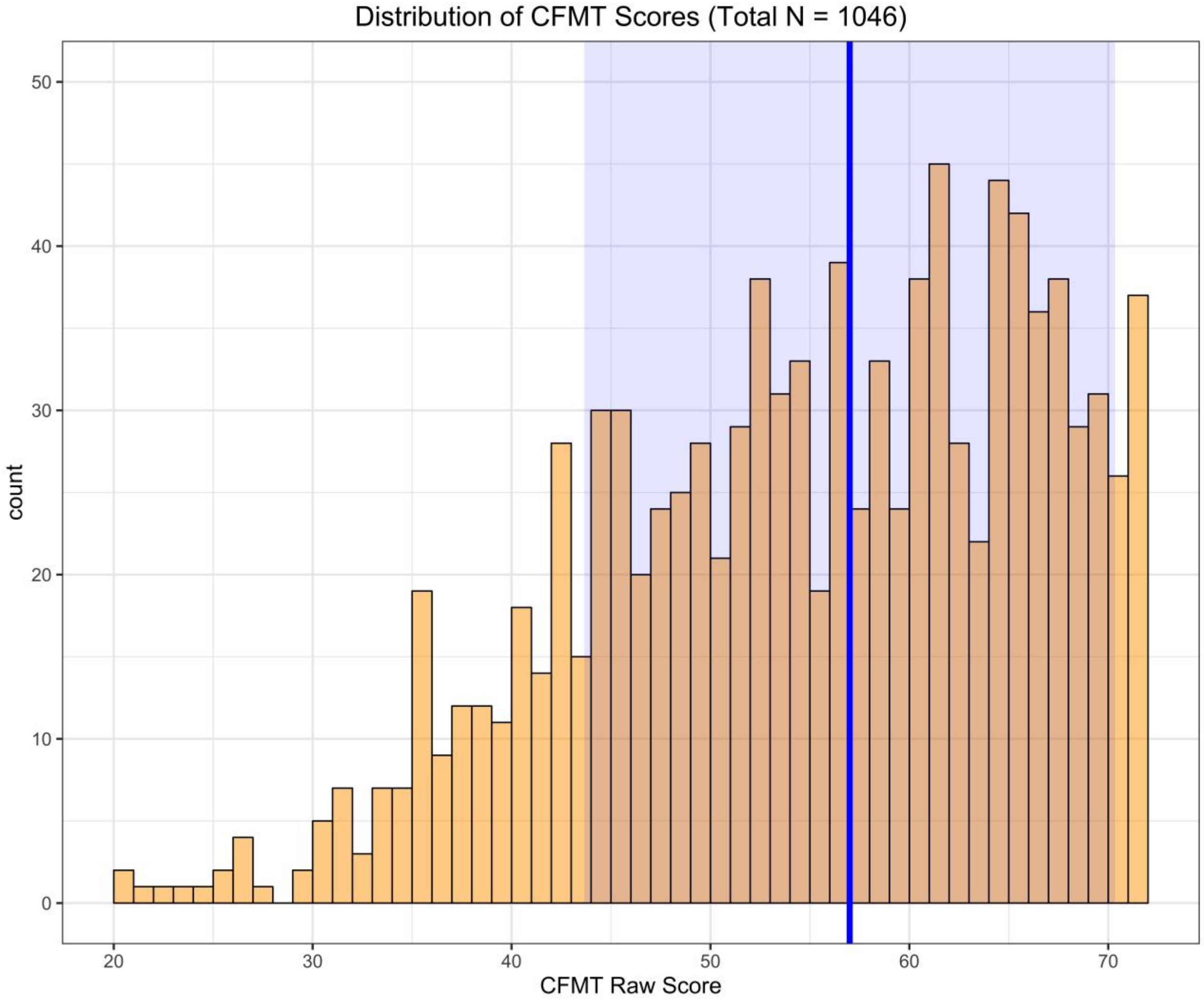

Figure 1. Distribution of CFMT scores for 1046 participants in the study. The blue line represents the median score (Median = 57), while the faded area surrounding represents ± 1 Median Absolute Deviation (MAD = 13.34).

Procedure

During encoding, participants viewed 12 faces: six of these faces one time each (weakly encoded faces) and the other six faces four times each (strongly encoded faces). Faces were displayed for three seconds with a one-second interstimulus interval. Participants were randomly assigned to one of four counterbalancing groups, which varied the presentation order of these faces as well as which faces were strongly and weakly encoded, with the following stipulations:

1) For faces shown four times each, the same face would appear twice in the first half of the encoding phase and twice in the second half and 2) the same face would not appear consecutively. Additionally, to control for primacy and recency effects, two filler faces appeared at the beginning of the encoding phase and another two filler faces appeared at the end but did not appear during the test phase.

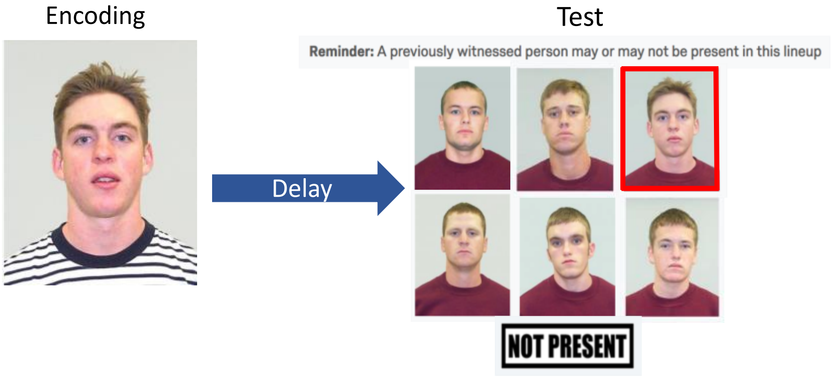
Encoding
Test
Reminder: A previously witnessed person may or may not be present in this lineup
Delay
NOT PRESENT

Figure 2. Example of the lineup identification task. Participants' task was to select the person from the encoding phase or to indicate that this person was 'Not Present' in the lineup.

Prior to the encoding phase, we instructed participants that they would "see a series of faces. Some faces you will see only once, and others you will see four times. Please pay close attention because after a delay we will ask you questions about who you saw." We further informed them that they would be randomly assigned to either a 5-minute, 1-day, or 1-week delay. As an attention check, before showing the encoding phase stimuli, we presented a question: "Some faces you will see only once, but how many times will you see the other faces?" Those responding anything other than '4' were asked to reread the instructions. Failing this check a second time resulted in termination of study procedures. (37 participants failed this check and are not included in the results or descriptive statistics).

Participants completed the lineup task (see Figure 2 for an example) after a delay of five minutes (n = 334), one day (n = 327), or one week (n = 385). We instructed them that they would see a series of lineups where a single face they viewed previously may or may not be present and that “previously viewed faces may look different in their lineup mugshots. This can be due to changes in lighting, clothing, facial hair, and/or other reasons.” Their task was to either identify the face they remembered from before, or to indicate that they did not recognize any of the faces in the lineup by selecting ‘not present.’ After making their selection, participants were instructed to indicate their confidence in the accuracy of their response using a 6-point scale ranging from 0% (guessing) to 100% (completely confident) in 20% increments.

To demonstrate the lineup task, we asked participants to pretend that they viewed a particular yellow smiley face. We then immediately presented a lineup of six colorful smiley faces. Only those who correctly selected the yellow smiley face proceeded to the test lineups. (66 participants failed this check and are not included in the results or descriptive statistics). In addition to the participants who were excluded due to failing these attention checks, 450 trials (3.6% of trials) were excluded based on decision time. Specifically, those with decision times of less than .1 seconds or outside three median absolute deviations were removed.

Half of the 12 test lineups contained an individual viewed during encoding (i.e., ‘target present’; TP), whereas the other half replaced this face with another person closely matched on descriptive characteristics (i.e., ‘target absent’; TA). Each lineup served as either a TP or TA lineup depending on its randomly assigned counterbalancing group. Participants completed the set of lineups associated with the counterbalancing group to which they were assigned at encoding. Lineup orders in all four counterbalancing groups followed two stipulations: 1) no more than three TP/TA lineups appeared consecutively and 2) no more than three lineups associated with strongly/weakly encoded faces appeared consecutively. After responding to all

12 lineups, participants completed the CFMT, followed by a short demographic survey that included questions about race, age, and sex.

Results

All of the deidentified data and the analysis code are located on the Open Science Framework (OSF): https://osf.io/tduzw/. The purpose of this experiment is to examine why face recognition ability influences the confidence-accuracy relationship for lineup identifications. Why, for instance, are weaker face recognizers more likely than stronger face recognizers to make high confidence misidentifications? To begin answering this question, we fit a logistic mixed effects model predicting identification accuracy from the fixed factors of Confidence, CFMT, Decision Time, Delay (5-minute, 1-day, 1-week), and Repetition (presenting target faces four times or one time), with random intercepts of Participant and Lineup. The random intercepts of Participant and Lineup instruct the model to consider the differential contributions of each participant and each lineup to the fixed effects. Specifically, the inclusion of the participant intercept partials out systematic error variance due to unmeasured characteristics of these individuals (e.g., fatigue, motivation). Similarly, the presence of the lineup intercept reduces the Type I error rate compared to regression or other models by considering the variability in the memorability of different items (Murayama, Sakaki, Yan, & Smith, 2014).

Continuous predictors (Confidence, CFMT, Decision Time) were centered and scaled prior to model fitting. We used separate models to analyze responses in which participants selected a face (chooser responses) and those in which participants selected ‘not present’ (nonchooser responses) because growing research shows that analyzing chooser and nonchooser data together with a single model can obscure effects within the nonchooser data, such as a reliable, albeit weak, relationship between confidence and nonchooser accuracy (e.g., Brewer & Wells, 2006; Grabman et al., 2019; Wixted & Wells, 2017).

With both the chooser and nonchooser models, we started by fitting full 5-way, 4-way, 3-way, 2-way, and main effects models using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015, version 1.1-17) in R *v.3.5.1* (R Core Team, 2018). We then selected the most parsimonious model from each start point using a backward stepwise elimination procedure based on Akaike's Information Criterion (AIC; Akaike, 1974). This method removed model terms that improved AIC without violating the principle of marginality (e.g., a two-way term could not be dropped if it was nested in a higher three-way term; Burnham & Anderson, 2002; Nelder, 1977). (See Appendix A for the chooser model selection table). Next, we selected the best-fitting of these reduced models as determined by lowest AIC. Likelihood ratio tests (LRTs) were performed using the *afex* package (Singmann, Bolker, Westfall, & Aust, 2018; version 0.23-0) to assess the significance of the best model's terms. The *effects* package (Fox, 2003; version 4.1-2) computed model estimates and 95% confidence intervals.

Finally, we used three different methods to assess absolute fit for the best models. First, we used the DHARMa package (Hartig, 2018; version 0.2.4) to perform Kolmogorov-Smirnov goodness-of-fit tests (KS tests), comparing the observed data to a cumulative distribution of 1,000 simulations from model estimates. Second, we used the same package to check for signs of model misspecification by examining residual plots for each predictor. Lastly, we used the MuMIn package (Barton, 2018; version 1.43.6) to calculate both the conditional ($R^2_{GLMM(c)}$) and marginal ($R^2_{GLMM(m)}$) pseudo-R^2 for fixed effects. The difference in these measures is that the conditional statistic includes variance accounted for by the random effects in the model (i.e., participant and lineup), whereas the marginal statistic is limited to fixed effects.

Chooser Accuracy

We first examined chooser accuracy, or identification accuracy when participants select a face (as opposed to 'not present') from a lineup. To include as much data as possible, following Dobolyi and Dodson (2018), we predicted chooser accuracy using the following types of responses: correct identifications from target-present lineups (TPc), foil identifications from target-present lineups (TPfa), and foil identifications from target-absent lineups (TAfa). In other words, the accuracy of these three types of responses served as the dependent variable for the chooser accuracy mixed effects model, in which each response was modeled separately. The frequencies of these types of chooser responses by repetition condition, delay, and confidence level are presented in Table 1, along with the frequencies of nonchooser responses: correct rejections from target-absent lineups (TAc) and incorrect rejections from target-present lineups (TPm).

Table 1

Frequency of Lineup Responses by Repetition, Delay, and Level of Confidence

			Confidence						
Repetition	***Delay***	***Response***	***0***	***20***	***40***	***60***	***80***	***100***	***Total***
4 times	**5 minutes**	TPc	9	38	69	93	145	198	552
		TPfa	9	44	48	61	24	6	192
		TAfa	12	107	124	92	65	17	417
		TAc	32	100	129	125	115	62	563
		TPm	15	42	54	52	47	23	233
	1 day	TPc	7	40	56	95	102	97	397
		TPfa	10	54	68	72	37	9	250
		TAfa	13	130	126	112	62	17	460
		TAc	34	85	123	112	74	44	472
		TPm	20	64	64	70	42	23	283
	1 week	TPc	11	97	87	75	70	35	375
		TPfa	22	112	94	60	32	6	326
		TAfa	43	196	125	93	39	11	507
		TAc	86	171	138	116	68	30	609
		TPm	75	101	110	61	50	22	419
1 time	**5 minutes**	TPc	6	40	60	68	52	39	265
		TPfa	6	73	71	66	48	13	277
		TAfa	20	102	114	101	55	11	403
		TAc	32	113	112	147	122	51	577
		TPm	22	99	106	87	79	44	437
	1 day	TPc	4	31	52	49	36	16	188

	TPfa	17	88	87	68	50	17	327
	TAfa	21	105	125	108	57	19	435
	TAc	39	91	124	111	92	38	495
	TPm	28	89	108	98	77	31	431
1 week	TPc	12	59	37	32	25	9	174
	TPfa	25	151	112	72	27	6	393
	TAfa	50	204	120	78	43	9	504
	TAc	73	157	141	124	69	38	602
	TPm	87	147	127	85	68	25	539

Note. TPc and TPfa refer to a correct response and a foil identification from a target-present lineup, respectively. TAfa and TAc refer to a foil identification and a correct response ('not present') from a target-absent lineup, respectively. TPm refers to a 'not present' response to a target-present lineup.

The best-fitting model of chooser accuracy contains several main effects and two-way interactions. In Wilkinson-Rogers (1973) notation, this model is as follows: Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:CFMT + DecisionTime:CFMT + (1|Participant) + (1|Lineup). This model adequately fits the data according to the two measures of absolute fit (KS D = .013, p = .211; pseudo-$R^2_{GLMM(m)}$ = .190; pseudo-$R^2_{GLMM(c)}$ = .278) as well as visual inspection of the residual plots. Overall, this chooser model is based on 6442 responses from 1019 participants.

LRTs showed significant main effects of Repetition, $\chi^2(1) = 112.51$, $p < .001$, Confidence, $\chi^2(1) = 333.12$, $p < .001$, Delay, $\chi^2(2) = 11.69$, $p = .003$, Decision Time, $\chi^2(1) = 8.84$, $p = .003$, and face recognition ability (i.e., CFMT score), $\chi^2(1) = 87.45$, $p < .001$. All of these main effects were moderated by two-way interactions, which are plotted in Figures 3 and 4 and described in more detail in the succeeding paragraphs. The two panels in Figure 3 show how chooser accuracy changes as a function of both the participant's confidence in their identification and (a) their face recognition ability (CFMT score) as well as (b) their delay condition. In both Figures, the lines represent the estimates from the mixed effects model, and the shading indicates the 95%

confidence interval. Note that the Repetition x Delay (p = .066) and Decision Time x CFMT (p = .078) interactions are non-significant.

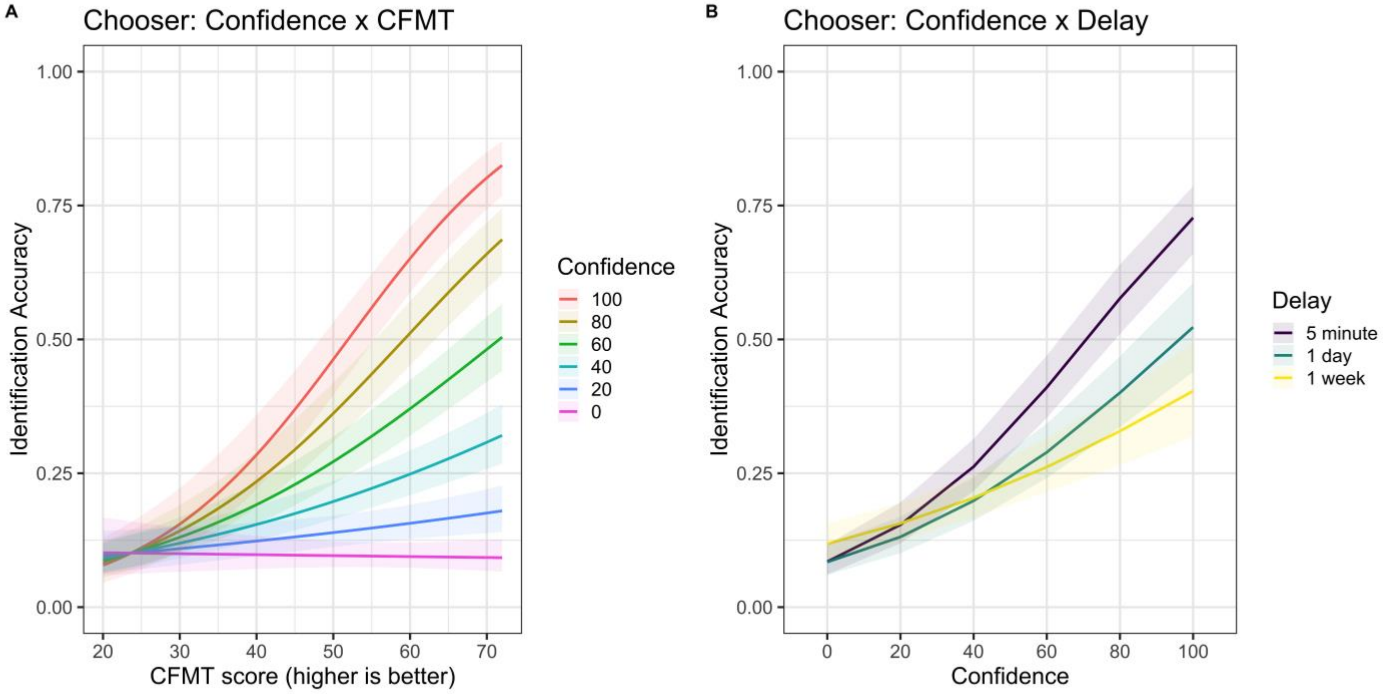

Figure 3. Two-way interactions between Confidence and (A) score on the Cambridge Face Memory Test (CFMT) and (B) Delay in the chooser model. Colored lines represent model estimates, with error shading indicating the 95% confidence interval. Notably, high confidence errors are more pronounced when participants are worse face recognizers (A) and when delay between encoding and recognition increases (B).

Figure 3a shows the interaction between Confidence and face recognition ability (CFMT score), $\chi^2(1) = 40.02$, $p < .001$. Confidence ratings are increasingly more predictive of chooser accuracy with increasing face recognition ability (i.e., individuals with higher CFMT scores). Specifically, for stronger face recognizers, high-confidence responses are much more likely to be correct than low-confidence responses. However, as shown in Figure 3a and replicating Grabman et al. (2019), the value of confidence as a predictor of lineup accuracy diminishes rapidly for individuals with worse face recognition ability. Those who are weaker face recognizers are much more likely to make a high confidence misidentification than individuals who are stronger face recognizers. Figure 3b shows that chooser accuracy generally worsens as delay increases. The interaction between Confidence and Delay, $\chi^2(2) = 27.70$, $p < .001$, indicates that the effect of delay on accuracy is greater for responses made with high versus low confidence.

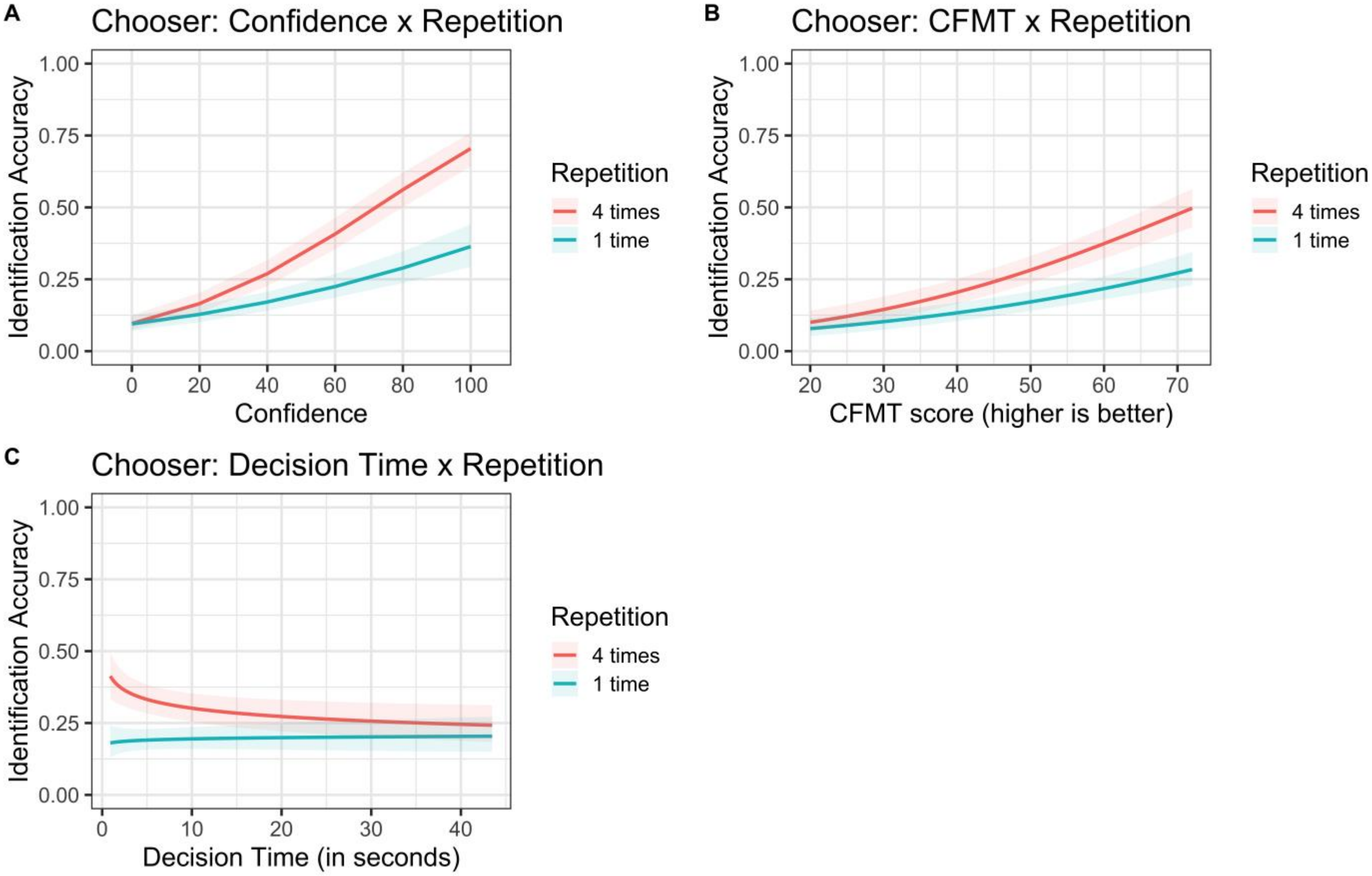

Figure 4. Two-way interactions between Repetition and (A) Confidence, (B) score on the Cambridge Face Memory Test (CFMT), and (C) Decision Time in the chooser model. Colored lines represent model estimates, with error shading indicating the 95% confidence interval.

Figure 4 shows generally higher accuracy for faces that had been seen four times than one time during the encoding phase. However, the separate panels reveal that the effects of repetition (4x vs. 1x) on chooser accuracy interact with confidence, face recognition ability, and decision time. In Figure 4a, the interaction between Confidence and Repetition, $\chi^2(1) = 28.05$, *p*

$< .001$, indicates that the benefit of having viewed the target face repeatedly is greater for identifications made with high, compared to low, confidence. Similarly, the interaction between face recognition ability (i.e., CFMT score) and Repetition, $\chi^2(1) = 3.88$, $p = .049$, in Figure 4b shows that stronger face recognizers benefit more from increased exposure to the target face than do weaker face recognizers.

The interaction between Decision Time and Repetition, $\chi^2(1) = 5.95$, $p = .015$, is represented in Figure 4c. For faces seen four times at encoding, consistent with the literature (Brewer & Wells, 2006; Sauer et al., 2010; Sauerland & Sporer, 2009), lineup identification accuracy worsens as decision time increases. For faces seen once only, accuracy is generally the same regardless of how long it takes to make the selection from the lineup. Given that chooser accuracy in the 1x faces condition (19.2%, 95% CI [15.8, 23.0]) was much worse than in the 4x faces condition (32.3%, 95% CI [27.6, 37.4]), it is likely that the overall poor performance for lineups associated with the 1x faces condition is interfering with the predictive ability of decision time.

Suspect-ID Accuracy

Mickes (2015; see also Wixted & Wells, 2017) has argued that identification accuracy should be measured as the ratio of correct identifications from target-present lineups (TPc) to the sum of this value plus foil identifications from target-absent (TAfa) lineups – a score known as suspect ID accuracy (i.e., TPc/[TPc+(TAfa/6)] for fair lineups). Foil responses to target-present lineups (TPfa) are excluded in suspect-ID accuracy since police know that target-present foils are innocent individuals. Therefore, suspect-ID accuracy duplicates the perspective of law enforcement by determining the likelihood that an identified individual is the guilty suspect (i.e., TPc) as opposed to an innocent filler (i.e., TAfa/6 with fair lineups).

We used a proxy measure for suspect-ID accuracy where all TA foil responses were retained as it is not feasible to divide the number of TA foil responses by six within the mixed effects model framework. Specifically, the dependent variable for our suspect-ID accuracy model differed from that for our chooser accuracy model in that it did not include TPfa responses. Using this accuracy measure, we fit a generalized linear mixed effects model predicting suspect-ID accuracy from the same fixed factors (Confidence, CFMT, Decision Time, Delay, and Repetition) and random intercepts (Participant and Lineup) as in the chooser accuracy model, using the same procedure. In Wilkinson-Rogers notation, the best-fitting model of suspect-ID accuracy is: Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:DecisionTime + Confidence:CFMT + DecisionTime:CFMT + Repetition:Confidence:DecisionTime + Confidence:DecisionTime:CFMT + (1|Participant) + (1|Lineup). (See Appendix B for the model selection table). This model adequately fits the data according to the two measures of absolute fit (KS $D = .015$, $p = .232$; pseudo-$R^2_{GLMM(m)} = .201$; pseudo-$R^2_{GLMM(c)} = .228$) as well as visual inspection of the residual plots. Overall, this model is based on 4677 responses from 1003 participants.

The suspect-ID analysis is generally consistent with the chooser analysis. Specifically, the main effects of Repetition, Confidence, Delay, Decision Time, and CFMT are significant in both the suspect-ID accuracy and chooser accuracy models, as are the interactions between Repetition and Confidence, Repetition and Decision Time, Confidence and Delay, and Confidence and CFMT. Additionally, the interaction of Decision Time and CFMT is non-significant in both models. However, the interaction between Repetition and Delay, indicating a greater decrease in accuracy with delay for lineups associated with faces seen once only than faces seen four times, is only significant in the suspect-ID accuracy model, $\chi^2(2) = 6.73$, $p = .034$, though it was marginally significant in the chooser accuracy model. Similarly, the interaction between Repetition and CFMT is only significant in the chooser accuracy model, $\chi^2(1) = 3.88$, $p = .049$, though it is a weak effect. All effects that are strong in one model (e.g., p < .01) are also strong in the other model. See Appendix C for an LRT table containing all model terms.

Face Recognition Ability and the Confidence-Accuracy Relationship

Our main question is whether the optimality or decision processes account best explains stronger face recognizers' more informative use of confidence. As a first step in answering this question, we examined calibration scores, which measure how well confidence ratings align with lineup identification accuracy. Excellent calibration, for example, is shown when individuals assign high, medium, and low confidence ratings to responses of high, medium, and low accuracy. We computed a chooser calibration score for each individual by taking the absolute value of the difference between (a) the actual accuracy at each level of confidence for a chooser response (i.e., either a correct identification or a false identification from either a target-present or target-absent lineup) and (b) the predicted accuracy of this response, as indicated by the confidence rating (e.g., Koriat & Goldsmith, 1996). This difference score is then weighted by the frequency of responses at each level of confidence. Calibration scores are error scores because they measure the deviation between predicted and actual accuracy and they range from 0 (perfect calibration) to 1.

We used multiple linear regression to examine the influence of face recognition ability (CFMT score, centered and scaled) and overall chooser accuracy (centered and scaled) on participants' calibration scores. Overall chooser accuracy (Accuracy) is computed for all participants who made at least one chooser response. The accuracy values (1s or 0s, indicating correct or incorrect) for all chooser (TPc, TPfa, and TAfa) responses for a participant are averaged in calculating that participant's overall chooser accuracy score. We included Accuracy as a continuous predictor, instead of Repetition and Delay, because Repetition and Delay were manipulated in order to create a range of conditions under which accuracy is comparable between strong and weak face recognizers. In order to distinguish between the optimality and decision processes accounts, it is necessary to remove differences in accuracy as a confounding variable.

As a reminder, the optimality account predicts similar calibration scores for stronger and weaker face recognizers when there are no differences in accuracy between these individuals. On the other hand, the decision account predicts better calibration for stronger than weaker face recognizers across all levels of accuracy. After performing a Yeo-Johnson power transformation (Yeo & Johnson, 2000) on calibration score to satisfy the homoscedasticity assumption of linear regression, we found a significant regression equation (F(3, 1015[1]) = 24.00, p <.001), with an adjusted-R^2 of .063. Both CFMT (β = -.12, t(1015) = -3.68, p <.001) and Accuracy (β = -.18, t(1015) = -5.59, p <.001) significantly predict Calibration, but these main effects were moderated by a significant interaction between CFMT and Accuracy (β = .08, t(1015) = 2.44, p = .015), which is plotted in Figure 5.

Figure 5 shows that changes in identification accuracy have a much greater effect on the calibration scores of weaker than stronger face recognizers. For stronger face recognizers, represented by the higher CFMT scores at the upper end of the x-axis, there is minimal separation between the lines representing the different accuracy levels. This indicates that

[1] Degrees of freedom are 1015 for this analysis because 27 participants made only nonchooser responses and thus were not included in this model.

calibration remains similar across the different levels of identification accuracy for those with strong face recognition ability. For weaker face recognizers, as shown by the separation in the lines at the lower end of the x-axis, declining identification accuracy is associated with worsening calibration scores. Put another way, at high levels of identification accuracy (represented by the purple and blue lines), there are similar calibration scores for individuals at all levels of face recognition ability. By contrast, lower levels of accuracy (represented by the red and yellow lines) are associated with greater differences in calibration scores between stronger and weaker face recognizers due to weak face recognizers' diminished calibration. On the surface, these results appear to provide support for both the optimality and the decision accounts. First, the optimality account is supported because calibration is similar between stronger and weaker face recognizers at higher levels of overall accuracy. And, the decision account is supported because stronger face recognizers are better calibrated at lower levels of overall accuracy.

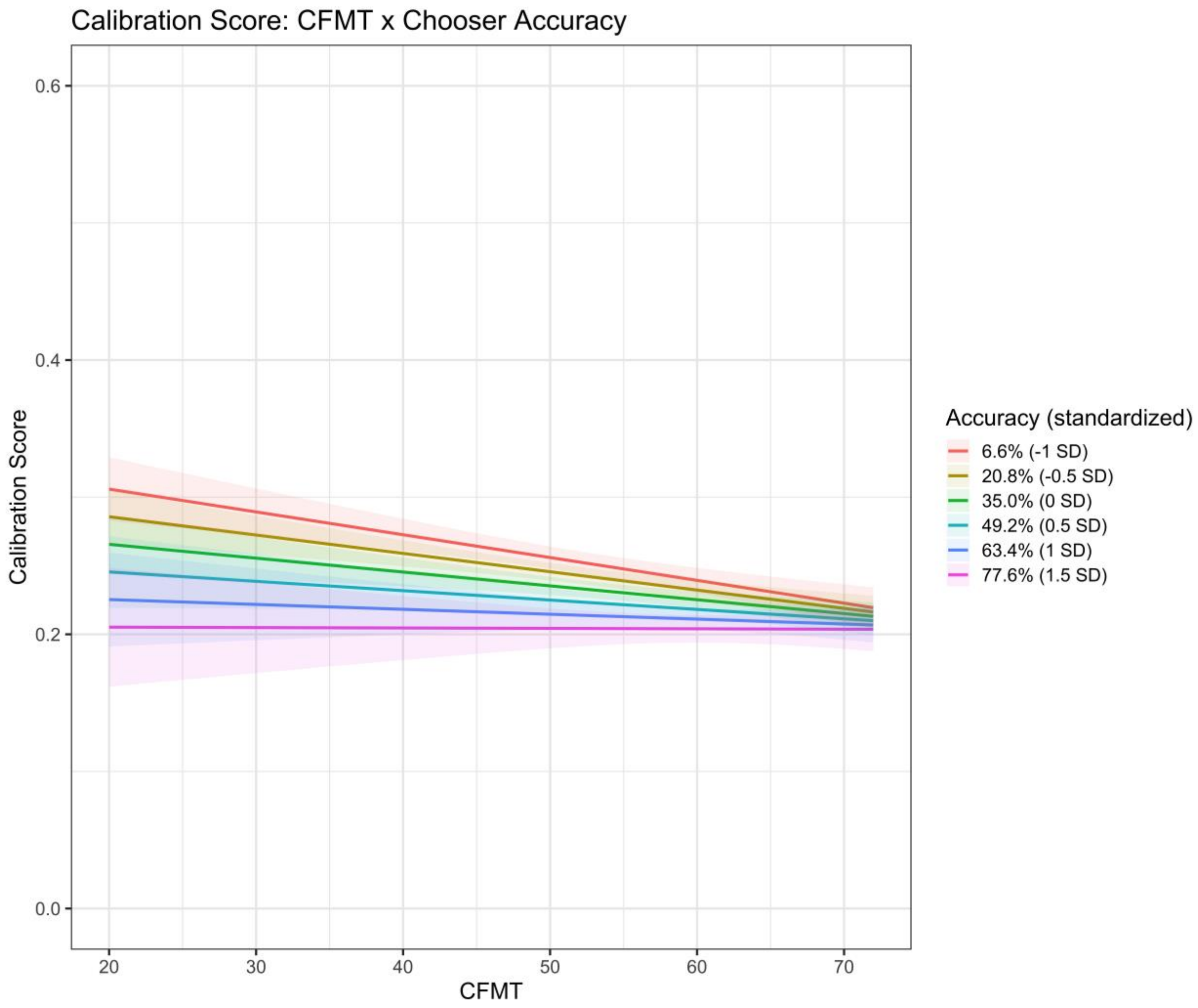

Figure 5. Interaction between CFMT and Chooser Accuracy in the CFMT x Accuracy model. Colored lines represent parameter estimates, with error shading indicating the 95% confidence interval. Standardized values indicate the number of standard deviations from the mean overall chooser accuracy percentage.

To further assess the influence of face recognition ability on the relationship between confidence and accuracy, we investigated the calibration curves that are the basis for the calibration scores. On the plots of the calibration curves, the y-axis represents the proportion of correct responses at each level of confidence displayed on the x-axis. To simplify matters, Figure 6 shows calibration curves for the top and bottom quartiles of accuracy, shown in the left and right facets, for individuals in the highest tertile, middle tertile, and lowest tertile of face recognition ability (CFMT score). The left side of Figure 6 shows that individuals at all levels of face recognition ability are comparably calibrated when identification performance is high. Interestingly, when identification accuracy is low, as shown on the right side of Figure 6, individuals at all levels of face recognition ability – including strong face recognizers – are vulnerable to making high confidence errors. In fact, when identification accuracy is low, the shape of the calibration curve is similar for individuals at all levels of face recognition ability. How then can stronger face recognizers be better calibrated than weaker face recognizers at low levels of accuracy? This is where it is important to examine the frequency of use of the different confidence ratings in Figure 6 because an individual's calibration score depends on (a) the deviation between the confidence rating and the actual level of performance at this rating; and (b) the frequency of use of the confidence rating.

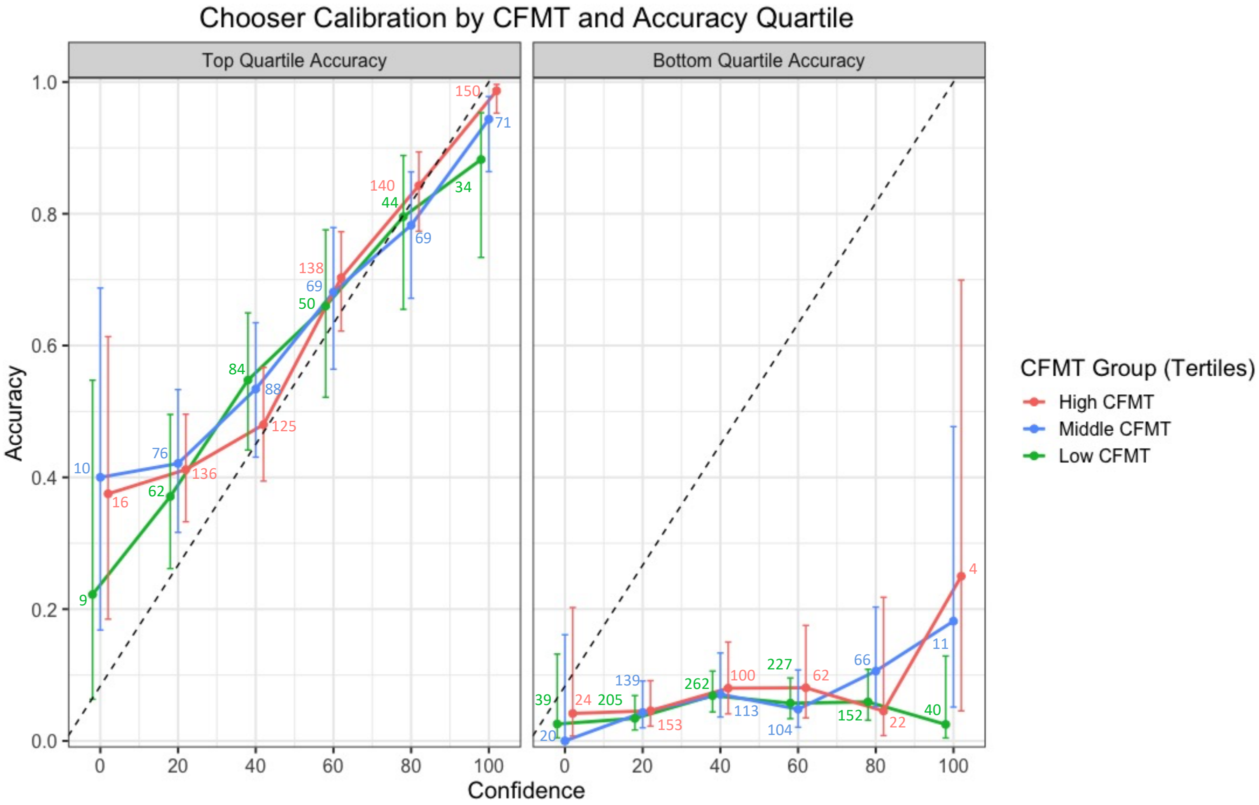

Figure 6. Confidence-accuracy curves as a function of Chooser Accuracy and CFMT Group (separated by tertiles). The diagonal line represents perfect calibration, and the numbers at each data point refer to the frequency of responses. Error bars indicate the 95% confidence interval.

In comparing the left and right facets of Figure 6, it is notable that stronger and weaker face recognizers differ dramatically in their frequency of use of the different confidence ratings, shown by the numbers at each data point in the figure. For example, under conditions that promote low levels of identification accuracy (right side of Figure 6), stronger face recognizers are nearly seven times more likely to use the lowest (0%/20%) than highest (80%/100%) levels of confidence. By contrast, weaker face recognizers use the different confidence scale points at a similar rate regardless of the level of accuracy of their identification.

Face Recognition Ability and Use of the Confidence Scale

To statistically analyze the relationship between face recognition ability, chooser accuracy, and the frequency of use of the different confidence ratings, we fit a cumulative link mixed model (CLMM) predicting the frequency of use of the confidence scale points from the fixed factors of CFMT (centered and scaled) and Accuracy (centered and scaled) with a random intercept of Participant. Frequency of use of the confidence ratings was tested as an ordinal response variable with each confidence rating (0, 20, 40, 60, 80, and 100) as a level. Models were fit using clmm2 in the *ordinal* package (Christensen, 2019; version 2019.4-25) in R *v.3.5.1* (R Core Team, 2018) and were compared with each other using AIC. LRTs assessed the significance of the terms of the best-fitting model: the model that contains the interaction and main effects and uses a flexible threshold. (See Appendix D for the model selection table). Note that equidistant threshold models were also evaluated, but flexible threshold models were preferred. The difference between flexible and equidistant thresholds is that the former assumes a relatively unstructured yet ordered response, where the difference between levels is not functionally equivalent, while the latter assumes equal spacing between levels.

LRTs showed significant main effects of face recognition ability (i.e., CFMT score), $\chi^2(1) = 5.80$, $p = .016$ and Accuracy, $\chi^2(1) = 113.26$, $p < .001$. These main effects were moderated by an interaction between face recognition ability and Accuracy, $\chi^2(1) = 26.24$, $p < .001$. The *effects* package (Fox, 2003; version 4.1-2) was used to compute model estimates and 95% confidence intervals for this interaction, which were subsequently plotted using a package currently in development to facilitate plotting of ordinal effects (Dobolyi, 2019).

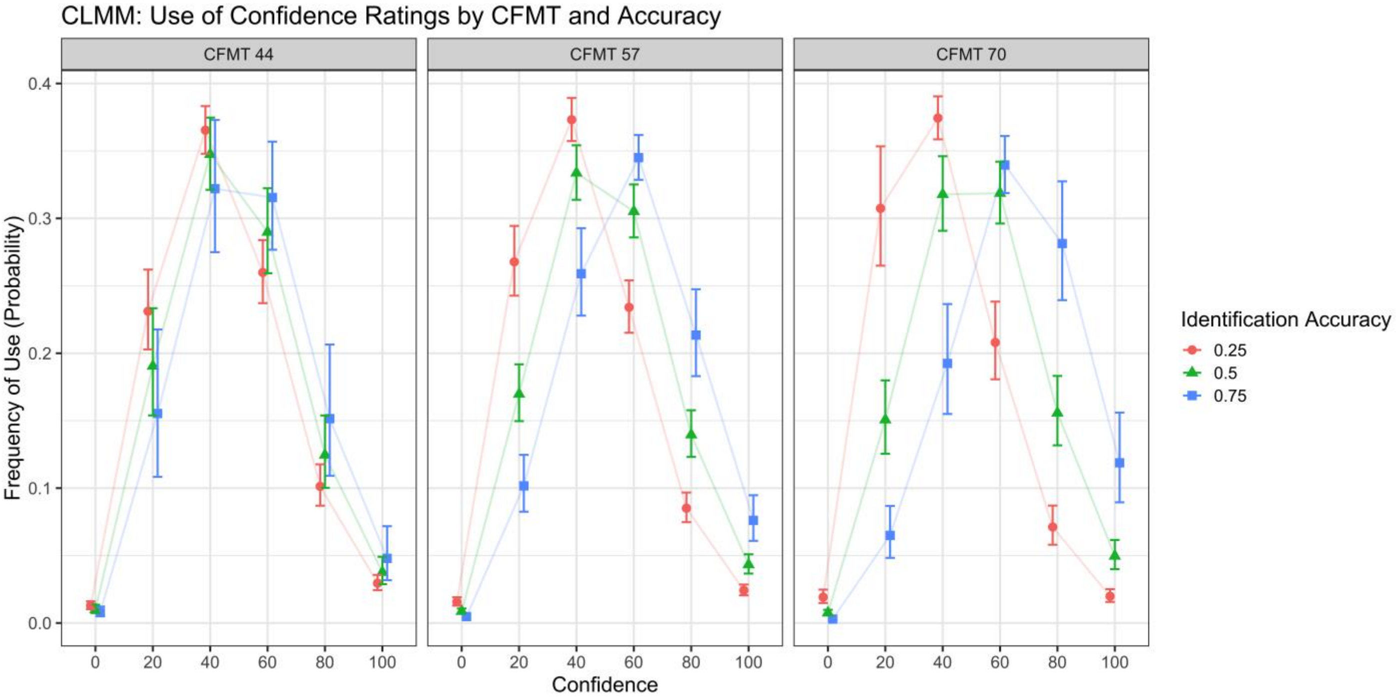

Figure 7. Interaction between CFMT and Chooser Accuracy in the CLMM model. The points indicate the predicted frequency of use of each confidence rating for the representative (i.e., median +/- one median absolute deviation) CFMT scores at the three levels of identification accuracy. Error bars indicate the 95% confidence interval.

Confirming and expanding upon the pattern depicted in Figure 6, Figure 7 shows that the frequency of use of the different levels of confidence (i.e., probability of use on the y-axis) changes with differing (a) levels of identification accuracy and (b) face recognition ability. Each facet presents the probability of use of the different confidence ratings at low (.25), moderate (.50), and high (.75) levels of identification accuracy for three representative levels of CFMT performance: the left facet shows one absolute deviation below the median (CFMT score of 44), the middle facet shows the median (CFMT score of 57), and the right facet shows one absolute deviation above the median (CFMT score of 70). Comparing the left and right facets, it is evident there is a dramatic difference in the use of confidence with changes in accuracy between weaker and stronger face recognizers. For example, the right facet shows that stronger face recognizers (represented by the CFMT score of 70) adjust their use of the confidence scale under conditions that induce higher or lower levels of accuracy. They use the highest levels of confidence (80%/100%) roughly four times more often when identification accuracy is high (the blue data points) than low (red data points). A similar pattern occurs for stronger face recognizers with the frequency of use of the lowest levels of confidence. By contrast, the left facet shows that weaker face recognizers (represented by the CFMT score of 44) fail to adjust their frequency of use of the confidence ratings under conditions that promote higher or lower levels of accuracy. They use the different levels of confidence at a similar rate regardless of their overall level of identification accuracy.

Overall, then, effectively regulating confidence is what allows stronger face recognizers to remain well-calibrated despite large changes in identification accuracy, whereas weaker face recognizers show a decreased ability to regulate confidence with declines in identification accuracy. Importantly, ineffectively regulating confidence is what causes weaker face recognizers to make more high confidence errors than stronger face recognizers under conditions when both groups show comparable levels of identification accuracy.

Nonchooser Accuracy

Finally, we investigated nonchooser accuracy, or the accuracy of lineup decisions for which participants selected ‘not present’ instead of identifying a face. For this analysis, we considered the accuracy of correct rejections from target-absent lineups (TAc) and incorrect rejections from target-present lineups (‘miss’; TPm). As shown earlier, Table 1 provides the frequencies of these types of responses by repetition condition, delay, and confidence level. The best-fitting model of nonchooser accuracy is represented in Wilkinson-Rogers notation as: Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:DecisionTime + (1|Participant) + (1|Lineup). (See Appendix E for the model selection table). According to KS tests (KS D = .011, p = .496) and visual inspection of the residual plots, this model adequately fits the data. However, the marginal pseudo-R^2 was considerably lower than in the chooser model (pseudo-$R^2_{GLMM(m)}$ = .019; pseudo-$R^2_{GLMM(c)}$ = .024). Given that two out of three absolute fit indices as well as our relative fit measure (i.e., AIC) supported proper model specification, we proceeded with this nonchooser model. Overall, this nonchooser model is based on 5660 responses from 992 participants.

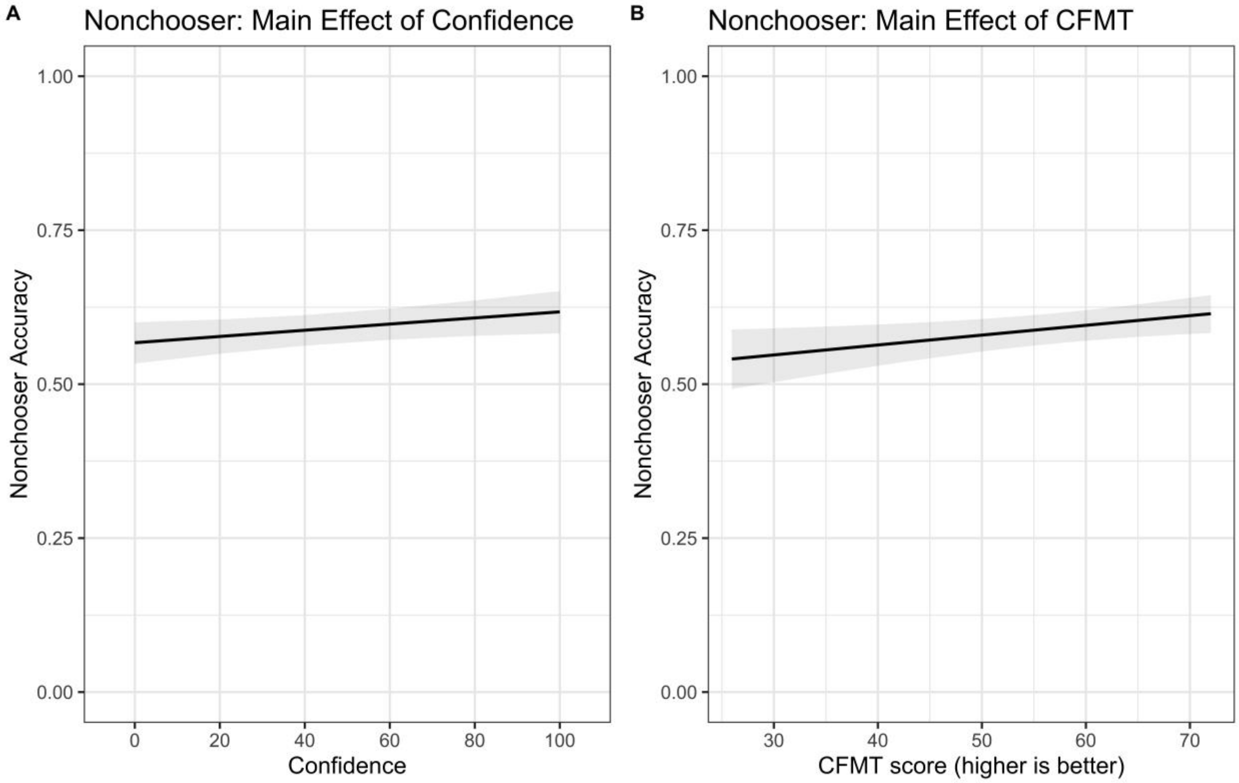

Figure 8. A) Confidence and B) CFMT main effects on Nonchooser Accuracy. Lines represent model estimates, with error shading representing the 95% confidence interval. Notably, performance improves with higher levels of confidence and greater face recognition ability.

Figure 8 shows the main effects of both Confidence and CFMT on nonchooser accuracy. Figure 8a indicates that nonchooser accuracy improves as participants express higher Confidence, $\chi^2(1) = 4.41$, $p = .036$. However, this effect is small, as is typical in nonchooser models (Dodson & Dobolyi, 2016; Sauerland & Sporer, 2009), since accuracy only improves by approximately 7% as ratings increase from 0% to 100% confidence. The main effect of CFMT, $\chi^2(1) = 8.90$, $p = .003$, represented in Figure 8b, shows that nonchooser accuracy increases with greater face recognition ability (i.e., higher CFMT score), though this effect is small numerically as accuracy improves by a total of 5% between the lowest and highest CFMT scores.

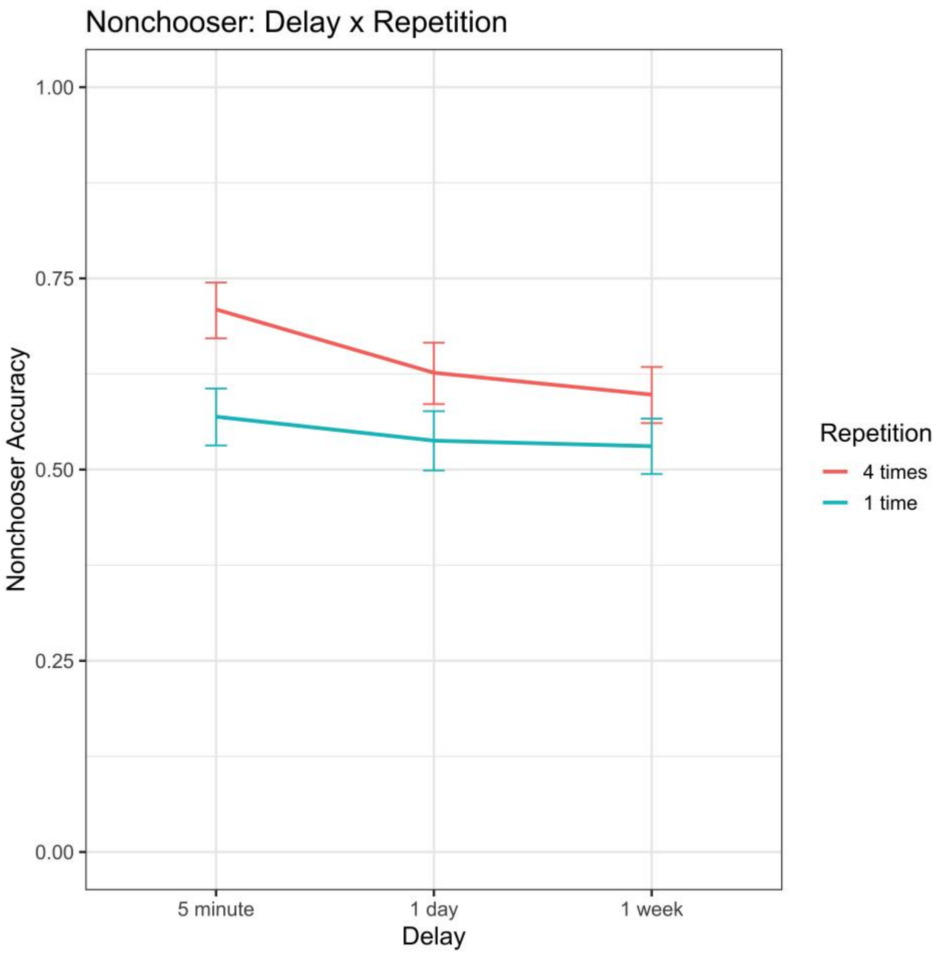

Figure 9. Interaction between Delay and Repetition in the nonchooser model. Colored lines represent model estimates, with error bars indicating the 95% confidence interval.

Lastly, we found a significant interaction between Repetition and Delay, $\chi^2(2) = 6.68$, *p* = .035, which is shown in Figure 9. Nonchooser accuracy worsened with increasing delay for lineups associated with both 4x faces and 1x faces; however, the effect of delay was stronger for lineups associated with 4x faces. The main effect of Decision Time (p = .241) as well as the Repetition x Decision Time (p = .148), Repetition x CFMT (p = .051), and Confidence x Decision Time (p = .102) interactions are non-significant.

Discussion

An eyewitness's confidence in an initial lineup identification is generally a reliable predictor of their accuracy (Wixted & Wells, 2017), and recent research has shown that this confidence-accuracy relationship is stronger for individuals with greater face recognition ability (Grabman et al., 2019). The question we answer in this paper is why stronger face recognizers show a more robust confidence-accuracy relationship than do weaker face recognizers. We address this question by distinguishing between two different accounts: the optimality account and the decision processes account. The optimality account attributes differences in the confidence-accuracy relationship between stronger and weaker face recognizers to stronger face recognizers' richer memory representations of faces, which improve their overall accuracy and give their confidence ratings more predictive value. The decision account posits that stronger face recognizers' superior ability in the domain of face recognition affords them superior metacognitive abilities in this domain, which allow them to better evaluate their performance and make more informed confidence judgments. One way of distinguishing between these two accounts is by creating conditions that produce comparable levels of identification performance between individuals of varying face recognition ability. The optimality account predicts a similar confidence-accuracy relationship between individuals of better and worse face recognition ability when their overall accuracy is similar because similar identification performance is produced by similar representations for the target face. By contrast, the decision account predicts that stronger face recognizers will show a stronger confidence-accuracy relationship, relative to weaker face recognizers, even under conditions of comparable identification performance.

Consistent with our past findings (Grabman et al., 2019; Grabman & Dodson, in press), we show that face recognition ability moderates the relationship between confidence and identification accuracy. Figure 3a shows that identification accuracy varies with confidence for stronger face recognizers, as identifications given high confidence ratings were far more likely to be correct than those made with low confidence. By contrast, accuracy was similar for weaker face recognizers across all levels of confidence. Therefore, those with worse face recognition ability are more likely to make high-confidence misidentifications, giving their confidence ratings less utility than those of better face recognizers. But, since stronger face recognizers are generally more accurate than weaker face recognizers, there is no information in Figure 3 that allows us to point to either the optimality account or the decision account to explain why the confidence-accuracy relationship is more robust in stronger than weaker face recognizers. However, by manipulating delay and exposure to create a variety of levels of identification performance, we are able to equate performance between stronger and weaker face recognizers, allowing us to determine whether differences in accuracy or differences in face recognition ability best explain differences in the confidence-accuracy relationship.

The key theoretical contribution of our results is that we dissociate two metacognitive processes that contribute to the calibration of confidence to accuracy: 1) the correspondence between confidence and accuracy is primarily influenced by overall accuracy; but 2) the frequency of use of the confidence scale is mainly affected by face recognition ability. Figure 6 shows that the correspondence between confidence and accuracy is primarily determined by the overall rate of identification accuracy. When overall accuracy is low (right panel of Figure 6), both stronger and weaker face recognizers exhibit a general pattern of overconfidence and poor correspondence between confidence and accuracy. It is important to emphasize that even strong face recognizers are vulnerable to making high confidence errors under conditions associated with low accuracy. But, when overall accuracy is high (left panel of Figure 6), then individuals of all levels of face recognition ability show a strong correspondence between accuracy and confidence. There is no evidence that stronger face recognizers show a stronger confidence-accuracy relationship than weaker face recognizers when overall identification accuracy is comparable between these groups. In short, in this paradigm, the relationship between confidence and accuracy appears to be mainly affected by the overall level of identification accuracy.

What distinguishes stronger from weaker face recognizers is the frequency of use of the different confidence ratings. For example, under conditions of low accuracy (Figure 6, right), stronger face recognizers make fewer high confidence (80%/100%) errors (26 total) than weaker face recognizers (192 total) due to their more informative use of the confidence scale. Specifically, Figure 7 shows that when overall identification accuracy is poor (i.e., the red line), stronger face recognizers much more frequently use the lower than the higher end of the confidence scale, and the opposite pattern is true when overall accuracy is high (i.e., the blue line). By contrast, weaker face recognizers use the different points on the confidence scale at generally the same rate regardless of whether overall accuracy is poor or high.

Differences in the frequency of use of the confidence ratings do not appear to be caused by differences in the way that stronger and weaker face recognizers interpret the meaning of particular points on the confidence scale. In other words, the placement and criteria used for making a particular confidence rating appear to be similar for all individuals, regardless of their face recognition ability. Consider the pattern when all individuals show comparably high accuracy, as represented by the blue lines in Figure 7. Is it possible that stronger face recognizers (right panel in Figure 7) more often use the top half of the confidence scale than do weaker face recognizers (left panel in Figure 7) because stronger face recognizers use a more liberal criteria for these confidence ratings – i.e., they require less or less precise information to give a confidence rating of 60, 80 or 100? If stronger and weaker face recognizers used different criteria and interpreted the confidence scale differently, then this would affect the ratio of correct identifications and false alarms for these levels of confidence, which in turn would affect accuracy at these levels of confidence. But, Figure 6 shows that when overall identification accuracy is comparable between stronger and weaker face recognizers, there is also comparable accuracy at each level of confidence. Therefore, we see little evidence that the meaning of a particular confidence point is different for individuals of differing face recognition ability.

How then are stronger face recognizers able to regulate their use of the confidence scale more effectively than weaker face recognizers such that they more often use higher and lower points on the confidence scale when identification accuracy is high and low, respectively? We speculate that the decision process that distinguishes stronger from weaker face recognizers involves an awareness of the likely accuracy of a particular response. This awareness may be thought of as a kind of second-order metacognitive judgment. Drawing on the work of Jang, Wallsten, and Huber (2012; see also Fleming and Daw, 2017), the type of evidence that is the basis for an identification decision is not necessarily identical to the evidence that contributes to a confidence rating. Specifically, face recognition ability determines metacognitive awareness of the likely accuracy of a response, which in turn contributes to the frequency of use of confidence ratings. Stronger face recognizers likely have an increased awareness that a response is more likely to be correct after a five-minute delay than after a week's delay, and consequently, more frequently use the top half of the confidence scale in the former than the latter condition. By contrast, weaker face recognizers' diminished metacognitive awareness contributes to their use of the confidence scale points at the same rate regardless of whether overall identification performance is excellent (75% accuracy), middling (50% accuracy), or poor (25% accuracy).

This difference in regulating use of the confidence scale is what allows stronger face recognizers to remain relatively well-calibrated, as shown in Figure 5, despite large changes in identification accuracy. By contrast, failure to appropriately regulate confidence ratings leads weaker face recognizers to become much worse calibrated with declining identification accuracy. Therefore, changes in identification accuracy cannot explain the differences in calibration between individuals of varying levels of face recognition ability. Instead, differences between stronger and weaker face recognizers in the frequency of use of the confidence scale points – fueled, we believe, by the relative insight into a decision's likely accuracy – are responsible for changes in the confidence-accuracy relationship with face recognition ability.

Our results are difficult to explain with standard signal detection models of memory. The standard signal detection model assumes that a confidence rating is based on the same information that is used to make a memory judgment. Consequently, this model predicts that when stronger and weaker face recognizers show (a) comparable overall levels of identification performance and (b) comparable accuracy at each level of confidence (i.e., similar confidence-accuracy relationship), then they should also show a similar frequency of use of the different confidence ratings – contrary to what we observed. Instead of the standard signal detection model, our data are better explained by models that decouple the information that drives a memory decision from the information that contributes to a confidence rating.

Other studies have shown that differences in the use of the confidence scale impact the confidence-accuracy relationship. Carpenter et al. (2019) administered perceptual tasks and asked participants to rate their confidence after each task decision using a four-point scale. They then implemented training sessions, providing feedback to roughly half of participants regarding their task performance while the other half received feedback on their metacognitive calibration, or how well their confidence judgments predicted their accuracy. After calibration training, participants were more likely to use the highest confidence rating when evaluating correct decisions compared to before training. Additionally, Koriat and Goldsmith (1996) found that the increased ability to distinguish correct from incorrect answers (i.e., better monitoring effectiveness) on recall tests than on recognition tests was a result of recall participants' increased polarization, or their greater likelihood of using both extremes (compared to only the high extreme) of the confidence scale. Though neither finding shows differences in the use of the entire confidence scale between groups (as we show between stronger and weaker face recognizers), these studies do indicate that changes in the use of confidence ratings are responsible for differences in the confidence-accuracy relationship.

Although our results clearly favor the decision processes account, it is important to note that this study is the first (to our knowledge) to document the contribution of expertise to calibration in the domain of face recognition ability. Improved confidence-accuracy calibration with expertise has been shown before, but specific to other domains, such as academic performance (Bol et al., 2005; Moreland et al., 1981; Shaughnessy, 1979), chess ability (Chi, 1978), tennis skills (McPherson & Thomas, 1989), physics problems (Chi, Glaser, & Rees, 1982), and social aptitude (Fagot & O'Brien, 1994), some of which require intentional and consistent practice to reach "expert" status, while face recognition ability is generally considered an innate skill (though see Balas & Saville, 2017 for contrasting evidence).

For example, Shaughnessy (1979) asked students to evaluate their memory for items they had studied on four exams throughout a semester. For a total of 30 four-alternative multiple-choice questions per exam, students gave confidence ratings for their responses using a four-point scale ranging from "definitely incorrect" to "definitely correct" immediately after answering each question. Those who performed better on the exams were also better calibrated. To explain this difference in calibration, Shaughnessy suggested that individual differences in memory-monitoring ability, above and beyond differences due to the strength of the memory trace, lead to more accurate memory assessments for students with better exam scores. He further proposed that since memory-monitoring ability is a skill, it can be improved through instruction, which can help poor students improve their calibration and memory self-assessments despite their inferior memory ability. This explanation is in line with the decision account as well as our results regarding metamemory judgments for face recognition performance, given that we matched memory performance for stronger and weaker face recognizers and still found a difference in calibration between the two groups. Overall, more research regarding the role of proficiency in determining the predictive value of confidence judgments for face recognition tasks is necessary, but our results strongly suggest that there is a causal relationship between expertise and calibration.

Further, as mentioned previously, the primary purpose of the delay and exposure (repetition) manipulations was to establish conditions under which the performance of stronger and weaker face recognizers could be matched; however, the effect of delay on identification accuracy is also of interest. Replicating previous work (Juslin et al., 1996; Palmer et al., 2013; Read et al., 1998; Sauer et al., 2010), we observed a decrease in identification performance with increasing delay. Importantly, we used three retention intervals, allowing us to trace the trajectory of forgetting with respect to face-relevant information. The decline in accuracy between the 5-minute and 1-day delays was much greater than that separating the 1-day and 1-week delays, suggesting that the first 24 hours after viewing a face (or suspect) are critical.

Contrary to previous work (Palmer et al., 2013; Sauer et al., 2010; see also Wixted et al., 2016), we did not find that high confidence is a protective factor against the effects of delay, using either chooser accuracy or our proxy measure for suspect-ID accuracy. A possible explanation for this discrepancy is the difference in overall performance between our study and others in the literature. Overall performance in our study was significantly lower than in prior studies, with mean chooser accuracy scores of 38.1%, 27.5%, and 23.6% in the 5-minute, 1-day, and 1-week delay conditions, respectively. In contrast, prior work has found chooser accuracy scores of 62% and 47% (Sauer et al., 2010), 60.2% and 53.7% (Palmer et al., 2013), and 69% and 64% (Juslin et al., 1996) for the immediate and delay conditions. Differences in stimulus presentation could be responsible for this discrepancy in overall accuracy as previous work has used target persons (e.g., Palmer et al., 2013; Sauer et al., 2010) or mock-crime videos (e.g., Juslin et al., 1996), while we used photographs of target faces, which are likely less memorable (Goldstein, Chance, Hoisington, & Buescher, 1982). Additionally, these differences in stimulus presentation led to differences in the number of separate faces participants encoded; participants in our study encoded 12 faces, while participants only encoded either one or two faces in Juslin et al. (1996), Palmer et al. (2013), and Sauer et al. (2010). Therefore, encoding multiple faces (and viewing some of these faces for only one 3-second interval) could also have contributed to the relatively low level of overall accuracy we observed. However, using mock-crime videos and immediate versus two-week delays and with a mean decision accuracy score of 66%, Brewer et al. (2019) also found that high confidence was not protective against the effects of delay.

Both face recognition ability and confidence predict the accuracy of nonchooser (i.e., 'not present') responses. Consistent with our previous findings (Grabman et al., 2019), stronger face recognizers were more accurate than weaker face recognizers in correctly rejecting lineups, indicating that better face recognition ability improves lineup performance when the target face is absent from the lineup as well as when it is present. Nonchooser accuracy was also higher for high confidence than low confidence responses, which is consistent with previous work (e.g., Grabman et al., 2019; Wixted & Wells, 2017). However, it is important to emphasize that both face recognition ability and confidence are much stronger predictors of chooser accuracy than nonchooser accuracy.

With respect to ecological validity, a possible limitation of our study is that the use of photographs of target faces is less realistic than showing participants a live person or mock-crime video. In our study, participants viewed photographs of multiple faces and were later shown one lineup associated with each face (12 lineups total) and asked to respond to each lineup by identifying an individual they viewed previously or indicating that none of these individuals were present in the lineup. In contrast, the real-world eyewitness experience typically involves witnessing a crime and subsequently responding to a single lineup constructed around the suspect for this crime (except in crimes with multiple perpetrators). However, the advantages of using 12 different target faces include increased stimulus generalizability as well as more data overall. Additionally, as is the case for actual eyewitnesses, participants viewed different pictures of the target individual at encoding and test, ensuring our paradigm relied on face recognition rather than simple picture-matching.

Another potential limitation is our assumption that equating performance between stronger and weaker face recognizers equates the quality of the memory representation. The Grabman et al. (2019) data speak to this point. This paper used an eyewitness identification paradigm that is nearly identical to the one used in the current paper. The one difference is that in Grabman et al. (2019), participants were asked to provide a justification for their identification decisions. Participants' justifications can be separated into two broad categories: those indicating reliance on familiarity to make identification decisions (e.g., "This person looks familiar") versus those referencing recollective information in making these decisions (e.g., "I remember his nose"). Grabman et al. (2019) did not find an interaction between CFMT (face recognition ability) and Justification-type when modeling the chooser accuracy data, which suggests there is no difference in reliance on familiarity versus recollection between stronger and weaker face recognizers. Because there is no evidence that reliance on familiarity versus recollection differs according to face recognition ability, we believe our assumption is reasonable that when identification performance is comparable between stronger and weaker face recognizers, then the underlying memory representation is also comparable. Future studies should explore the relationship between face recognition ability and confidence-accuracy calibration using different stimulus presentations and retention intervals.

In conclusion, we distinguished between two accounts of the differences in the eyewitness confidence-accuracy relationship between individuals with stronger and weaker face recognition ability by identifying an explanation for the superior calibration of stronger face recognizers. Our results support the decision processes account, suggesting that the metacognitive abilities afforded to stronger face recognizers by their expertise, or talents, in the domain of face recognition ability are responsible for the discrepancy in the use of confidence between stronger and weaker face recognizers. These metacognitive skills allow stronger face recognizers to consider factors affecting their identification accuracy when making associated confidence judgments. Without this same awareness, weaker face recognizers fail to appropriately adjust their use of the confidence scale with changes in accuracy. Taken together, we have shown that the correspondence between confidence and accuracy is primarily influenced by the overall level of identification accuracy, whereas the frequency of use of the confidence scale is primarily influenced by face recognition ability. Therefore, the "certainty of the witness" criterion put forth by the Supreme Court for assessing the reliability of an identification could be equally applicable to all potential eyewitnesses, if those lacking in face recognition ability were able to improve their confidence judgments by considering and applying information regarding the likely accuracy of their identifications.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19,* 716-723. http://dx.doi.org/10.1109/TAC.1974.1100705

Balas, B., & Saville, A. (2017). Hometown size affects the processing of naturalistic face variability. *Vision Research*, *141*, 228-236. http://dx.doi.org/10.1016/j.visres.2016.12.005

Barton, K. (2018). *MuMIn: Multi-model inference*. R package version 1.43.6. https://CRAN.Rproject.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-17. https://CRAN.Rproject.org/package=lme4

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, *73,* 269-290. http://dx.doi.org/10.3200/JEXE.73.4.269-290

Bornstein, B. H., Deffenbacher, K. A., Penrod, S. D., & McGorty, E. K. (2012). Effects of exposure time and cognitive operations on facial identification accuracy: A meta-analysis of two variables associated with initial memory strength. *Psychology, Crime & Law*, *18*, 473-490. http://dx.doi.org/10.1080/1068316X.2010.508458

Brewer, N., Weber, N., & Guerin, N. (2019). Police lineups of the future? *American Psychologist*. Advance online publication. http://dx.doi.org/10.1037/amp0000465

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11-30. http://dx.doi.org/10.1037/1076-898X.12.1.11

Burnham, K.P., & Anderson, D.R. (2002). *Model selection and inference: A practical information-theoretic approach.* New York: Springer. http://dx.doi.org/10.1007/b97636

Carpenter, A. C., & Schacter, D. L. (2018). Flexible retrieval mechanisms supporting successful inference produce false memories in younger but not older adults. *Psychology and Aging, 33*, 134-143. http://dx.doi.org/10.1037/pag0000210

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*, 51-64. http://dx.doi.org/10.1037/xge0000505

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73-96). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 17-76). Hillsdale, NJ: Erlbaum.

Christensen, R. H. B. (2019). *ordinal - Regression Models for Ordinal Data*. R package version 2019.4-25. http://www.cran.r-project.org/package=ordinal/

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior, 4*, 243-260. http://dx.doi.org/10.1007/BF01040617

Dobolyi, D. G. (2019). *ordinalEffects: Improved Effect Plots for Ordinal Models*, v0.1.0. http://dx.doi.org/0.5281/zenodo.3464589

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. http://dx.doi.org/10.1037/a0034596

Dobolyi, D. G., & Dodson, C. S. (2018). Actual versus perceived eyewitness accuracy and confidence and the featural justification effect. *Journal of Experimental Psychology: Applied*, *24*, 543-563. http://dx.doi.org/10.1037/xap0000182

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging*, *22*, 122-133. http://dx.doi.org/10.1037/0882-7974.22.1.122

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology*, *30*, 113-125. http://dx.doi.org/10.1002/acp.3178

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*, 576-585. http://dx.doi.org/10.1016/j.neuropsychologia.2005.07.001

Fagot, B. I., & O'Brien, M. (1994). Activity level in young children: Cross-age stability, situational influences, correlates with temperament, and the perception of problem behaviors. *Merrill Palmer Quarterly*, *40*, 378-398.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*, 91-114. http://dx.doi.org/10.1037/rev0000045

Fox, J. (2003). Effect displays in R for generalized linear models. *Journal of Statistical Software, 8*, 1–27. http://dx.doi.org/10.18637/jss.v008.i15

Gettleman, J. N., Grabman, J. H., Dobolyi, D., & Dodson, C. (2020). Explaining differences in the eyewitness confidence-accuracy relationship between strong and weak face recognizers. *Open Science Framework.* http://dx.doi.org/10.17605/osf.io/tduzw

Gignac, G. E., Shankaralingam, M., Walker, K., & Kilpatrick, P. (2016). Short-term memory for faces relates to general intelligence moderately. *Intelligence*, *57*, 96-104. http://dx.doi.org/10.1016/j.intell.2016.05.001

Goldstein, A. G., Chance, J. E., Hoisington, M., & Buescher, K. (1982). Recognition memory for pictures: Dynamic vs. static stimuli. *Bulletin of the Psychonomic Society*, *20*, 37-40. http://dx.doi.org/10.3758/BF03334796

Goldstein, A. G., Johnson, K. S., & Chance, J. (1979). Does fluency of face description imply superior face recognition? *Bulletin of the Psychonomic Society*, *13*, 15-18. http://dx.doi.org/10.3758/BF03334999

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*, *8*, 233-243. http://dx.doi.org/10.1016/j.jarmac.2019.02.002

Grabman, J. H., & Dodson, C. S. (in press). Stark individual differences: Face recognition ability influences the relationship between confidence and accuracy in a recognition test of Game of Thrones actors. *Journal of Applied Research in Memory and Cognition.* http://dx.doi.org/10.1016/j.jarmac.2020.02.007

Hartig, F. (2018). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. R package version 0.2.4. https://CRAN.Rproject.org/package=DHARMa

Hildebrandt, A., Wilhelm, O., Schmiedek, F., Herzmann, G., & Sommer, W. (2011). On the specificity of face cognition compared with general cognitive functioning across adult age. *Psychology and Aging*, *26*(3), 701-715. http://dx.doi.org/10.1037/a0023056

Jacoby, L. L., Bishara, A. J., Hessels, S., & Toth, J. P. (2005). Aging, subjective experience, and cognitive control: Dramatic false remembering by older adults. *Journal of Experimental Psychology: General*, *134*, 131-148. http://dx.doi.org/10.1037/0096-3445.134.2.131

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review, 119*, 186-200. http://dx.doi.org/10.1037/a0025960

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. http://dx.doi.org/10.1037/0278-7393.22.5.1304.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490-517. http://dx.doi.org/10.1037/0033-295X.103.3.490

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121-1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121

Lin, W., Strube, M. J., & Roediger, H. L. (2019). The effects of repeated lineups and delay on eyewitness identification. *Cognitive Research: Principles and Implications*, *4*. http://dx.doi.org/10.1186/s41235-019-0168-1

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433-442. http://dx.doi.org/10.3758/s13428-016-0727-z

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*, 295-314. http://dx.doi.org/10.1007/s10648-014-9287-x

McPherson, S. L., & Thomas, J. R. (1989). Relation of knowledge and performance in boys' tennis: Age and expertise. *Journal of Experimental Child Psychology*, *48*, 190-211. http://dx.doi.org/10.1016/0022-0965(89)90002-7

Meissner, C. A., Brigham, J. C., & Butz, D. (2005). Memory for own- and other-race faces: A dual-process approach. *Applied Cognitive Psychology*, *19*, 545-567. http://dx.doi.org/ 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*, 93-102. http://dx.doi.org/10.1016/j.jarmac.2015.01.003

Morcom, A. M., Li, J., & Rugg, M. D. (2007). Age effects on the neural correlates of episodic retrieval: Increased cortical recruitment with matched performance. *Cerebral Cortex*, *17*, 2491-2506. http://dx.doi.org/10.1093/cercor/bhl155

Moreland, R., Miller, J., & Laucka, F. (1981). Academic achievement and self-evaluations of academic performance. *Journal of Educational Psychology*, *73*, 335-344. http://dx.doi.org/10.1037/0022-0663.73.3.335

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1287-1306. http://dx.doi.org/10.1037/a0036914

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

Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society: Series A (General)*, *140*, 48-63. http://dx.doi.org/10.2307/2344517

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. http://dx.doi.org/10.1037/a0031602

Read, J. D., Lindsay, D. S., & Nichols, T. (1998). The relation between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? In C. P. Thomson, D. Bruce, J. D. Read, D. Hermann, D. Payne, & M. P. Toglia (Eds.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107-130). Mahwah, NJ: Erlbaum.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*, 252-257. http://dx.doi.org/10.3758/PBR.16.2.252

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

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. http:dx.doi.org/10.1037/a0014560

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied*, *24*, 400-415. http://dx.doi.org/10.1037/xap0000157

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*, 12887-12892. http://dx.doi.org/10.1073/pnas.1421881112

Shaughnessy, J. J. (1979). Confidence judgment accuracy as a predictor of test performance. *Journal of Research in Personality*, *13*, 505-514. http://dx.doi.org/10.1016/0092-6566(79)90012-6

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *afex: Analysis of factorial experiments*. R package version 0.23-0. https://CRAN.Rproject.org/package=afex

Tredoux, C. (1999). Statistical considerations when determining measures of lineup size and lineup bias. *Applied Cognitive Psychology*, *13*, S9-S26. http://dx.doi.org/10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-1

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*, 137. http://dx.doi.org/10.1186/1471-2288-14-137

Wan, L., Crookes, K., Dawel, A., Pidcock, M., Hall, A., & McKone, E. (2017). Face-blind for other-race faces: Individual differences in other-race recognition impairments. *Journal of Experimental Psychology: General*, *146*, 102-122. http://dx.doi.org/10.1037/xge0000249

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces: One element of social cognition. *Journal of Personality and Social Psychology*, *99,* 530-548. http://dx.doi.org/10.1037/a0019972

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, *22*, 392-399. http://dx.doi.org/10.2307/2346786

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*, 225-230. http://dx.doi.org/10.1177/0963721417710693

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*, 360-392. http://dx.doi.org/10.1080/02643294.2012.753433

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., … & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*, 5238-5241. http://dx.doi.org/10.1073/pnas.0913053107

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition*, *5*, 192-203. http://dx.doi.org/10.1016/j.jarmac.2016.04.006

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

Yeo, I-K., & Johnson, R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*, 954-959. http://dx.doi.org/10.1093/biomet/87.4.954

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., … & Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology*, *20*, 137-142. http://dx.doi.org/ 10.1016/j.cub.2009.11.067

Appendix A

Table A1

The Model Selection Table for the Chooser Accuracy Mixed Effects Model

Model Formula	df	AIC	ΔAIC
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:CFMT + DecisionTime:CFMT + (1 | Participant) + (1 | Lineup)	18	6638.49	0.00
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:DecisionTime + Confidence:CFMT + Delay:CFMT + DecisionTime:CFMT + Repetition:Confidence:Delay + Repetition:Confidence:DecisionTime + Repetition:Confidence:CFMT + Repetition:Delay:CFMT + (1 | Participant) + (1 | Lineup)	27	6639.23	0.74
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:DecisionTime + Confidence:CFMT + Delay:CFMT + DecisionTime:CFMT + Repetition:Confidence:Delay + Repetition:Confidence:DecisionTime + Repetition:Confidence:CFMT + Repetition:Delay:CFMT + (1 | Participant) + (1 | Lineup)	27	6639.23	0.74
Accuracy ~ (Repetition + Confidence + Delay + DecisionTime + CFMT)^5 + (1 | Participant) + (1 | Lineup)	50	6668.90	30.41
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + (1 | Participant) + (1 | Lineup)	9	6739.34	100.85

Note. Model formulae are in Wilkinson-Rogers notation (Wilkinson & Rogers, 1973). The table displays the model formula, degrees of freedom, and AIC for the best-fitting model from each start point (i.e., five-way full interaction, four-way interaction, three-way interaction, two-way interaction, and main effects only) in order of the overall best- to worst-fitting model based on AIC. The change in AIC between the current and overall best-fitting model is also included. Two of the model formulae in this table are the same (those for the three-way and four-way interaction models) since the best-fitting model from both of these start points converged on the identical model.

Appendix B

Table B1

The Model Selection Table for the Suspect-ID Accuracy Mixed Effects Model

Model Formula	df	AIC	ΔAIC
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:DecisionTime + Confidence:CFMT + DecisionTime:CFMT + Repetition:Confidence:DecisionTime + Confidence:DecisionTime:CFMT + (1 | Participant) + (1 | Lineup)	21	5490.09	0.00
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:DecisionTime + Confidence:CFMT + DecisionTime:CFMT + Repetition:Confidence:DecisionTime + Confidence:DecisionTime:CFMT + (1 | Participant) + (1 | Lineup)	21	5490.09	0.00
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Confidence + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:Delay + Confidence:CFMT + DecisionTime:CFMT + (1 | Participant) + (1 | Lineup)	18	5490.24	0.15
Accuracy ~ (Repetition + Confidence + Delay + DecisionTime + CFMT)^5 + (1 | Participant) + (1 | Lineup)	50	5522.74	32.65
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + (1 | Participant) + (1 | Lineup)	9	5580.69	90.60

Note. Model formulae are in Wilkinson-Rogers notation (Wilkinson & Rogers, 1973). The table displays the model formula, degrees of freedom, and AIC for the best-fitting model from each start point (i.e., five-way full interaction, four-way interaction, three-way interaction, two-way interaction, and main effects only) in order of the overall best- to worst-fitting model based on AIC. The change in AIC between the current and overall best-fitting model is also included. Two of the model formulae in this table are the same (those for the three-way and four-way interaction models) since the best-fitting model from both of these start points converged on the identical model.

Appendix C

Table C1

The Likelihood Ratio Table for the Suspect-ID Accuracy Model

Model Term	df	LR (χ2)	p	
Repetition	1	63.96	<.001	***
Confidence	1	278.45	<.001	***
Delay	2	7.85	.020	*
DecisionTime	1	7.83	.005	**
CFMT	1	64.72	<.001	***
Repetition:Confidence	1	21.7	<.001	***
Repetition:Delay	2	6.73	.034	*
Repetition:DecisionTime	1	6.19	.013	*
Repetition:CFMT	1	2.2	.138	
Confidence:Delay	2	19	<.001	***
Confidence:DecisionTime	1	3.19	.074	.
Confidence:CFMT	1	38.15	<.001	***
DecisionTime:CFMT	1	3.3	.069	.
Repetition:Confidence:DecisionTime	1	3.69	.055	.
Confidence:DecisionTime:CFMT	1	2.06	.151	

Note. Signif. codes: '***' <.001 '**' .01 '*' .05 '.' .10.

Appendix D

Table D1

The Model Selection Table for the Use of Confidence Ratings Cumulative Link Mixed Model

Model Formula	Threshold Type	df	AIC	ΔAIC
Confidence ~ CFMT * Overall.Accuracy + (1 | Participant)	Flexible	9	19015.08	0.00
Confidence ~ CFMT + Overall.Accuracy + (1 | Participant)	Flexible	8	19039.33	24.25
Confidence ~ CFMT * Overall.Accuracy + (1 | Participant)	Equidistant	6	19608.29	593.21
Confidence ~ CFMT + Overall.Accuracy + (1 | Participant)	Equidistant	5	19637.68	622.60

Note. Model formulae are in Wilkinson-Rogers notation (Wilkinson & Rogers, 1973).

Appendix E

Table E1

The Model Selection Table for the Nonchooser Accuracy Mixed Effects Model

Model Formula	df	AIC	ΔAIC
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:DecisionTime + (1 | Participant) + (1 | Lineup)	14	7590.05	0.00
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:DecisionTime + (1 | Participant) + (1 | Lineup)	14	7590.05	0.00
Accuracy ~ Repetition + Confidence + Delay + DecisionTime + CFMT + Repetition:Delay + Repetition:DecisionTime + Repetition:CFMT + Confidence:DecisionTime + (1 | Participant) + (1 | Lineup)	14	7590.05	0.00
Accuracy ~ Repetition + Confidence + Delay + CFMT + (1 | Participant) + (1 | Lineup)	8	7593.94	3.89
Accuracy ~ (Repetition + Confidence + Delay + DecisionTime + CFMT)^5 + (1 | Participant) + (1 | Lineup)	50	7634.56	44.51

Note. Model formulae are in Wilkinson-Rogers notation (Wilkinson & Rogers, 1973). The table displays the model formula, degrees of freedom, and AIC for the best-fitting model from each start point (i.e., five-way full interaction, four-way interaction, three-way interaction, two-way interaction, and main effects only) in order of the overall best- to worst-fitting model based on AIC. The change in AIC between the current and overall best-fitting model is also included. Three of the model formulae in this table are the same (those for the two-way, three-way, and four-way interaction models) since the best-fitting model from all of these start points converged on the identical model.