Investigating the Reliability of Those Who Provide (and Those Who Interpret) Eyewitness Confidence Statements

Jesse Howard Grabman

Charlottesville, Virginia

BA, University of Virginia, 2013

A Predissertation Research Project 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, 2019

Readers:

Dr. Chad S. Dodson

Dr. James P. Morris

Introduction

On the morning of May 7, 2000, 15-year old Brenton Butler was walking to retrieve a job application from the local Blockbuster video. Two hours earlier, a ‘skinny black male’ approached Mary and James Stephens outside their hotel and demanded Mary’s purse. Standing about three feet from the couple, the man pulled out a pistol and shot Mary dead before running away. Two police officers saw Butler and pulled him aside thinking he vaguely matched the perpetrator’s description. As Butler talked to a detective, from fifty-feet away James Stephens indicated that this was the teenager who shot his wife. Taken aback, the officers brought Stephens closer, and he confirmed that “he was sure of it, he would not put an innocent man in jail” (De Lestrade, 2001). Butler was tried as an adult based on this eyewitness testimony, and later acquitted due to investigators coercing him into a false confession. Ultimately, forensic evidence proved a different man committed the crime.

Judges in the United States are advised to use certainty as an indicator of eyewitness reliability (*Neil vs. Biggers,* 1972). And, increasing evidence shows that high confidence at the time of the initial identification is a strong predictor of accuracy, so long as proper lineup administration procedures are followed (Wixted & Wells, 2017). This strong relationship between high confidence and accuracy is documented in many laboratory studies, using a variety of manipulations (e.g. weapon vs. no weapon, other-race identifications) and stimuli (e.g., identifications after viewing photos of faces, videos, and/or staged crimes). Moreover, a recent field study suggests that these findings extend to real-world identifications (Wixted, Mickes, Dunn, Clark, & Wells, 2016).

However, as the Butler case demonstrates, high eyewitness confidence is not always reliable. In this thesis, I present research from our lab that raises important caveats to the growing consensus about a strong relationship between eyewitness confidence and accuracy. This includes lightly adapted versions of two published first-authored articles (Grabman, Dobolyi, Berelovich, & Dodson, 2019; Grabman & Dodson, 2019), as well as results from a recently submitted first-authored manuscript.

Part I shows that individual differences in face recognition ability influence the rate of high confidence errors. Specifically, weaker face recognition ability corresponds to increased rates of high confidence errors in both a controlled eyewitness experiment using criminal lineups (Study 1A), and in an uncontrolled 'real-world' face recognition task of actors from the popular television show *Game of Thrones* (Study 1B). Part II shows that the probative value of eyewitness confidence statements depends on evaluators (e.g., police officers, judges, jurors) properly interpreting the level of certainty the witness intended to convey. In three experiments (Study 2A – C), participants systematically misinterpreted witnesses' verbal confidence statements when they knew the identity of the suspect in a criminal lineup – a situation that is common in criminal justice decisions. Taken together, these studies suggest a degree of caution is warranted when using eyewitness confidence as an indicator of accuracy.

Introduction References

De Lestrade, J. X. (2001). *Murder on a Sunday Morning*.

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*(2), 233–243. https://doi.org/10.1016/j.jarmac.2019.02.002

Grabman, J. H., & Dodson, C. S. (2019). Prior knowledge influences interpretations of eyewitness confidence statements: 'The witness picked the suspect, they must be 100% sure'. *Psychology, Crime and Law*, *25*(1), 50–68. https://doi.org/10.1080/1068316X.2018.1497167

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences*, *113*(2), 304–309. https://doi.org/10.1073/pnas.1516814112

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*(1), 10–65. https://doi.org/10.1177/1529100616686966

Part I: Investigating the influence of face recognition ability on the confidence-accuracy relationship in eyewitness memory.

Study 1A: Predicting High Confidence Errors in Eyewitness Memory: The Role of Face Recognition Ability, Decision-Time, and Justifications (*Grabman et al., 2019*)

How confident can we be about eyewitness confidence? A growing consensus suggests that identifications by highly confident witnesses are generally accurate (Wixted & Wells, 2017). However, the question is whether there are variables that systematically influence the accuracy of high confidence identifications. In the sections that follow we briefly review research on three factors that form the foundation of the first study: (a) the speed of a lineup identification, (b) the basis for an identification from a lineup, and (c) face recognition ability. We focus primarily on face recognition ability as no one (to our knowledge) has investigated the influence of this factor on high confidence misidentifications.

Many studies find that lineup-identification accuracy worsens as decision-times increase when individuals choose a face from a lineup, though this association is weaker for non-identifications (e.g., Brewer & Wells, 2006; Dobolyi & Dodson, 2018; Dodson & Dobolyi, 2016; Dunning & Stern, 1994; Sauer, Brewer, Zweck, & Weber, 2010). But, growing evidence shows that high confidence errors also change as a function of the speed of lineup decisions. For example, Sauerland and Sporer (2009) found that confident (90 -100%) and fast (< 6s) identifications produced greater identification accuracy (97.1%) than confident, but slow, identifications (60.4%) (for similar results, see Brewer & Wells, 2006). Similarly, modeling decision-times continuously, Dodson and Dobolyi (2016) observed that accuracy greatly diminished for highly confident responses (100%) as decision-times increased. Taken together, these results suggest that, even under pristine lineup administration conditions, highly confident identifications may be reliable only insofar as the decision is made quickly.

In addition to decision-time, highly confident eyewitnesses can differ in the basis for their identification of someone from a lineup. In the only study to examine this issue, Dobolyi and Dodson (2018) asked individuals to justify their level of confidence in a response to a lineup. A content analysis showed that nearly 50% of all lineup-identifications were justified by referring to a single or multiple observable features about the suspect (e.g., "I remember his eyes and nose"). Moreover, 20% of all identifications were accompanied by a reference to familiarity (e.g., "He's familiar"), with the remaining identifications based on either an expression of recognition (e.g., "I recognize him") or a reference to an unobservable feature (e.g., "He looks like my cousin") or a mixture of these justification-types. For the present purposes, the key point is that high confidence misidentifications increased when identifications referenced familiarity as compared to the other justification types. However, the period between encoding and test was short (5-minutes), meaning that it is unclear whether this relationship holds for longer delays.

Finally, research conclusions about the confidence-accuracy relationship are currently based on and apply to the average individual. This focus on the average person, however, neglects individual differences which may account for some of the high-confidence errors that appear even when investigators follow proper procedures. The ability to recognize unfamiliar faces varies considerably from person to person (see Wilmer, 2017 for review). At the low end are those with prosopagnosia ('face-blindness'), while other individuals exhibit exceptional skill ('super-recognizers') (Ramon, Bobak, & White, 2019; Russell, Yue, Nakayama, & Tootell, 2010; Wan et al., 2017). Face recognition ability is highly heritable (Wilmer et al., 2010; Zhu et al., 2010) and distinct from other cognitive markers such as verbal and visual recognition ability, and general intelligence (e.g., for reviews, see Wilmer, 2017; Wilmer et al., 2012).

Although a few studies have shown that measures of face recognition predict eyewitness identification performance (Andersen, Carlson, Carlson, & Gronlund, 2014; Bindemann, Avetisyan, & Rakow, 2012; Morgan et al., 2007), no one has examined how heterogeneity in face recognition ability impacts the rate of high confidence misidentifications. One hypothesis about this relationship stems from Deffenbacher's (1980) optimality account, which holds that confidence will be a stronger predictor of accuracy under more than less ideal conditions at encoding, storage and retrieval. By this account, face recognition ability should influence the quality (optimality) of what is encoded and retrieved, which in turn will influence the relationship between confidence and accuracy. In short, poor face recognizers should be more prone than strong face recognizers to make high confidence misidentifications. Alternatively, Semmler, Dunn, Mickes, and Wixted's (2018) constant likelihood ratio account argues that, regardless of changes in overall accuracy, people assign confidence ratings so as to maintain the relationship between confidence and accuracy. Even though poor face recognizers will show worse accuracy than strong face recognizers, this account argues that there will be few changes in the predictive value of confidence – a high confidence identification will be comparably accurate across all levels of face recognition ability.

In sum, the purpose of this study is to investigate factors that potentially increase the rate of high confidence misidentifications, namely (a) decision-time, (b) justifications, and (c) face recognition ability. We examine these variables in concert with two other forensically relevant factors: the other-race effect (e.g., Meissner & Brigham, 2001) and retention interval (Wixted, Read, & Lindsay, 2016).

Methods

Participants

The study was administered online on respondents' personal laptop or desktop computers using Amazon's Mechanical Turk (mTurk). The 569 participants comprising the results ranged in age from 18 to 50 years (*M* = 31.66, *SD* = 6.08), were primarily female (68.5%), and all self-reported their race as White/Caucasian. Though no consensus standards are available for a-priori power estimates for mixed effects logistic regression models, this sample size was deemed sufficient in light of conservative recommendations of 50 responses per modeled variable (Van Der Ploeg, Austin, & Steyerberg, 2014), and findings that estimates are generally reliable for sample sizes greater than 30 with at least 10 responses per participant (McNeish & Stapleton, 2016). All participants received payment for completing the study. The University of Virginia Institutional Review Board approved this research.

Materials

Lineups. Participants viewed the same six Black and six White lineups as used in Dobolyi & Dodson (2013, 2018). These lineups consisted of a formal "head and shoulders" photograph of six individuals arranged in a 2 x 3 grid, wearing a maroon colored t-shirt, and exhibiting neutral facial expressions (see Figure 1A.1 for an example). All lineups met the criteria that no face is substantially more likely to be chosen by a naïve viewer based on a description of the perpetrator (i.e. lineups were 'fair'; see Dobolyi & Dodson, 2013 for more details on lineup generation). To avoid a simple picture-matching strategy, at encoding participants saw different photos of potential lineup targets wearing varied street clothing and casual expressions (e.g., 'smiling').

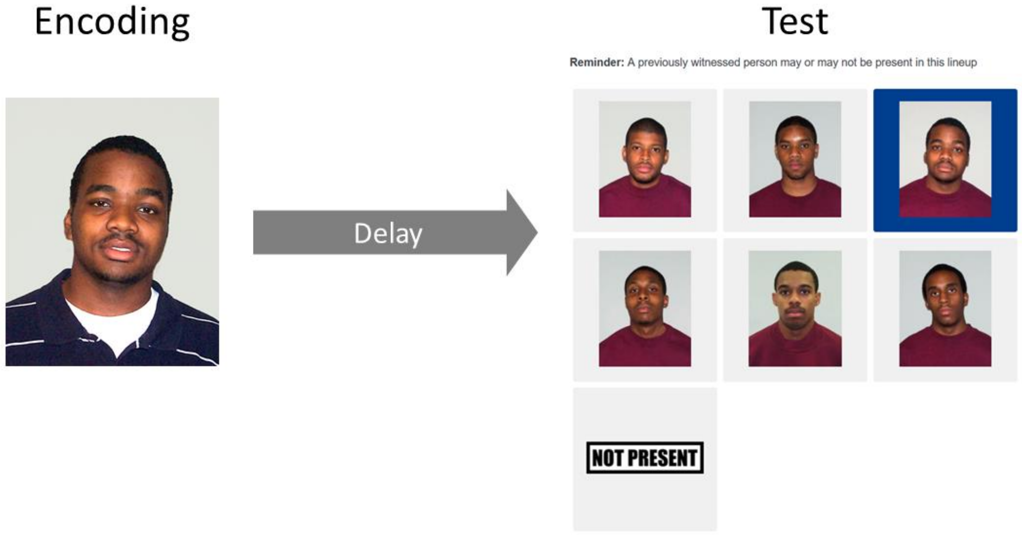

Encoding
Test
Reminder: A previously witnessed person may or may not be present in this lineup
Delay
NOT PRESENT

Figure 1A.1. Example of the identification task. Participants' task was to select the person from the encoding phase, or to indicate that they were "Not Present" in the lineup.

Face Recognition Task. We administered the Cambridge Face Memory Test (CFMT) (Duchaine & Nakayama, 2006) to assess participants' face recognition ability. In this task, respondents attempt to memorize six faces in three separate orientations. For each trial, previously viewed faces must be selected from an array of the target face and two foils. The test phase proceeds across 72 trials in three increasingly difficult blocks. Past research shows that a simple sum of correct responses is a reliable indicator of poor to above average recognition ability, with performance ranging from 0-72 correct selections (Cho et al., 2015). Figure 1A.2 shows the distribution of CFMT scores from the present study.

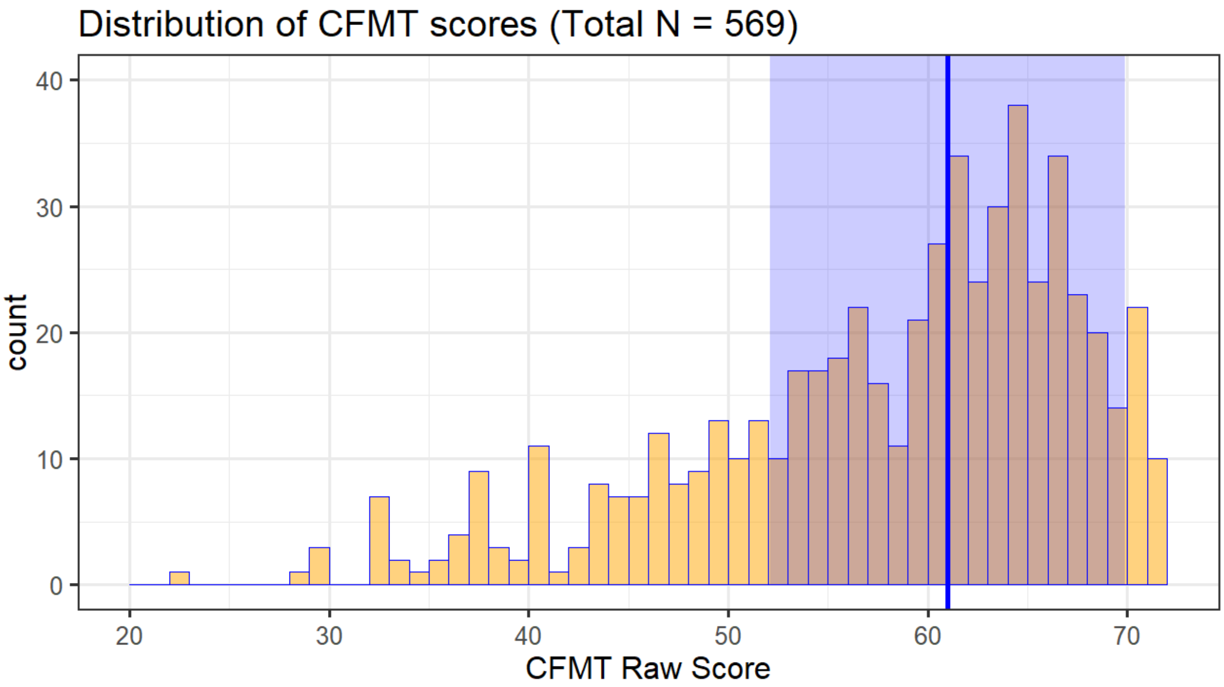


Figure 1A.2. Distribution of CFMT score for 569 participants in the study. The blue line represents the median score (Median = 61), while the faded area surrounding represents ± 1 Median Absolute Deviations (MAD = 8.9).

Procedure

Procedurally, the study is similar to Dobolyi & Dodson (2018), except for two key differences. First, all participants completed the CFMT at the end of the lineup memory task. Second, we assigned roughly half of participants (n = 277) to a 5-minute delay between the encoding and test phases, while the remaining participants were tested a day later (n = 292). Prior to the encoding phase, we instructed participants that they would "see a series of faces. These faces will repeat 3 times. Please pay close attention because after a delay we will ask you questions about who you saw." We further informed them that some participants would be randomly assigned to a 5-minute delay, whereas others would be prompted to return after a one-day delay. As an attention check, before showing the stimuli we asked, "how many times will the faces repeat?" Those responding anything other than '3' were asked to reread the instructions.

Failing this check a second time resulted in termination of study procedures (9 participants failed this check and are not included in the results or summary statistics).

After passing the check, participants viewed six Black and six White faces as a block three times in a randomized order. This order followed the stipulations that: 1) The same face would not appear at the end of one block and begin the subsequent block (i.e., none would be shown 'back to back') and 2) faces of the same race would be shown a maximum of two consecutive times. Faces appeared for three seconds with a one second interstimulus interval. Additionally, to control for primacy and recency effects, four filler faces (two Black, two White) appeared at both the beginning and end of the encoding phase, but did not appear during the test phase.

Participants completed the lineup task after either five minutes of working on an online word search, or roughly one day later upon seeing the prompt to begin the next phase of the experiment (see Figure 1A.1 for an example of the task). We instructed them that they would see a series of lineups where a single face they viewed previously may or may not be present. Their task was either to 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, we asked participants, "in their own words, [to] please explain how certain [they] are in [their] response" by typing into a text box. This was followed by a prompt to "please provide specific details about why" they made this expression of certainty. Finally, we asked them to indicate their confidence using a 6-point scale ranging from 0% (not at all certain) to 100% (completely certain) in 20% point increments.

To check comprehension, and to demonstrate the 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, after reading “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” (33 participants failed this check and are not included in the results or summary statistics).

In the test phase, half of the lineups (3 Black, 3 White) 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 condition. One of two predetermined lineup presentation orders were randomly assigned to each participant, with both following the criteria that 1) no more than two TP/TA lineups appeared consecutively, 2) no more than two lineups of the same race appeared consecutively, and 3) lineups appeared in different serial position across the two presentation orders. Finally, after finishing the lineups, participants completed the CFMT, followed by a short demographic survey that included questions on race, age, and sex.

Results

Data Preparation

The dataset is comprised of 7,248 lineup responses (12 lineups/participant x 604 participants), and is available on the Open Science Framework (OSF) (https://osf.io/j25yc). We divided the data into six roughly equal-sized groups of participants, and assigned each group to two research assistants to code justifications for lineup responses. The coding scheme was nearly identical to Dobolyi & Dodson (2018), categorizing justifications based on familiarity (F; e.g., “he looks familiar.”), single observable feature (O; e.g., “I remember his nose.”), multiple observable features (Omany; e.g., ‘I remember his nose and eyes.’), single unobservable feature

(U; e.g., ‘he looks like my cousin.’), multiple unobservable features (Umany; e.g. ‘He looks like my cousin, and another guy I know.’), and recognition (R; e.g., ‘I recall seeing this guy before.’). However, whereas Dobolyi & Dodson (2018) assigned combinations of justification types into a general ‘mixed’ category, we coded these responses into categories representing either familiarity + observable (FO; e.g., ‘his nose looks familiar’), or observable + unobservable (OU; e.g., ‘my friend’s eyes look like that’). The coding scheme for ‘not present’ responses is the same as for identifications, except that statements referred to the absence of a justification category, such as ‘none of the faces look familiar’ (coded as F) or ‘I don’t recognize any of them’ (coded as R). Statements that did not fit any category were coded as unknown.

Overall interrater agreement was high, with matching categorizations for 80.5% of lineup justifications. Across the pairs of raters, agreement ranged from 71.6% - 85.5%, with Cohen’s Kappas indicating acceptable agreement across coders (range Cohen’s κ = .66 - .83). To maximize the number of available responses, a third research assistant (masked to the other raters’ categorizations) coded statements where there was disagreement. We accepted any categorizations where at least two out of the three raters agreed on the statement. Due to the cross-race manipulation, we removed 20 participants who did not self-report their race as White/Caucasian. Additionally, we removed 15 participants based on not providing any justifications (N = 1), giving the same justification for all 12 lineups (e.g., “it was the same face as before”; N = 11), or providing nonsensical answers (e.g., “they’re all white guys wearing the same t-shirt”; N = 3).

As we planned on investigating decision-times in several analyses, we log transformed decision-times for each lineup, and calculated a median absolute deviation score. We removed decision-times shorter than .100 ms (n = 14 responses), as well as responses longer than 3 deviations above the median (roughly one minute) (n = 183 responses). We then eliminated responses where justifications could not be categorized (n = 845 responses). We also observed minimal numbers of OU (n = 27 responses) and Umany (n = 8 responses) categorizations, therefore we did not analyze these trials. Finally, we noticed many respondents mentioned that one of the Black target faces resembled a celebrity in the news during the experiment. Given that the study aims to examine responses to unfamiliar faces, this would be a major confound, and we removed responses to this lineup (n = 491 responses). In total, we examined 5,272 responses from 569 participants.

Table 1A.1 provides a breakdown of the frequency of justifications across confidence levels for chooser responses (i.e., selecting a face from the TP or TA lineup) and non-chooser responses (i.e., responding ‘not present’). Justifications for chooser decisions most frequently referenced one or more observable features, either in the context of familiarity with these features (FO = 10.7%), or otherwise (O1 + Omany = 31.7%). In contrast, non-chooser decisions most commonly referred to not recognizing any faces in the lineup (R = 65.1%) or that faces were unfamiliar (F = 31.9%).

We analyzed chooser responses and non-chooser responses with separate models because the infrequent use of many of the justification-types for non-chooser responses meant that it was impracticable to use the same model for both response-types. For each model of the ‘chooser’ and ‘non-chooser’ data, we used multi-model comparisons (Burnham & Anderson, 2002) to obtain the best generalized linear mixed effects model among the fixed factors: Justification Type, Lineup Race (Same Race, Other Race), Delay (5 minute, Day), Confidence, Decision-time and CFMT score. Participant ID served as a random intercept. Continuous predictors (confidence, decision-time, CFMT) were centered and scaled prior to model fitting.

Response	Lineup Race	Justification	**Confidence**						Total
			0	20	40	60	80	100	
Chooser	Same Race	F	14	92	90	86	49	14	345
		FO	7	42	53	49	25	6	182
		O1	2	31	47	55	80	68	283
		Omany	1	7	23	45	55	42	173
		R	13	60	66	87	80	100	406
		U1	0	3	8	21	22	35	89
	Other Race	F	13	97	88	71	56	10	335
		FO	2	28	26	32	18	6	112
		O1	1	22	41	56	53	58	231
		Omany	2	14	28	49	41	50	184
		R	10	48	59	66	66	95	344
		U1	0	5	5	9	18	26	63
	Total		65	449	534	626	563	510	2747
Non-Chooser	Same Race	F	31	78	84	109	109	39	450
		FO	1	1	1	3	1	0	7
		O1	0	4	2	3	5	3	17
		Omany	0	1	0	4	4	2	11
		R	51	118	170	220	230	126	915
		U1	0	1	0	1	1	0	3
	Other Race	F	24	39	82	99	79	33	356
		FO	0	0	3	0	2	2	7
		O1	0	1	2	8	4	6	21
		Omany	0	0	1	0	3	1	5
		R	73	109	120	176	168	83	729
		U1	0	0	0	1	2	1	4
	Total		180	352	465	624	608	296	2525

Table 1A.1. Frequency of responses in the intersection of lineup race, justification type, and confidence level for both Chooser and Non-Chooser decisions.

To begin, we started by fitting full 6-way, 5-way, 4-way, 3-way, 2-way, and main effects models using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014, version 1.1-21) in R *v.3.5.1* (R Core Team, 2018). Next, a backward stepwise elimination procedure based on Akaike's Information Criterion (AIC) selected the most parsimonious model from each start point. This method removed model terms that demonstrated any improvement in AIC, so long as this did not violate principles of marginality (e.g. a two-way term could not be dropped if it was nested in a higher three-way term). We then selected the best fitting of these reduced models as determined by AIC. Significance testing was performed on final model terms using likelihood ratio tests calculated by the *afex* package (Singmann, Bolker, Westfall, & Aust, 2018, version 0.21-2). The *effects package* (Fox, 2003, version 4.0-2) computed model estimates and 95% confidence intervals.

Finally, while there are no consensus standards for assessing absolute fits for generalized linear mixed effects models, we examined fits for final models using three methods. First, we used the DHARMa package (Hartig, 2018, version 0.2.0) 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 examined residual plots based on deviations between simulated and observed values to check for signs of model misspecification (i.e., ensuring errors are uniformly distributed for each predicted value). And third, we calculated marginal pseudo-R^2 ($R^2_{GLMM(m)}$) for fixed-effects, using the MuMIn package (Barton, 2018, version 1.42.1; see also Nakagawa & Schielzeth, 2013). This statistic includes variance accounted for by fixed effects in the model, while partialing out variance from the random effect structure (i.e., participant intercept).

Chooser model.

We sought to include as much data as possible in the analysis of identification accuracy and so, following Dobolyi and Dodson (2018), we modeled this score as the rate of correct identifications from target-present lineups (TPc) relative to the sum of this score and the rates of foil identifications from target-present (TPfa) and target-absent (TAfa) lineups (i.e., TPc/[TPc+TPfa+TAfa]).

Written in Wilkinson-Rodgers (1973) notation, the best-fitting model of identification accuracy consists of several main effects and two-way interactions: Accuracy ~ LineupRace + Confidence + Delay + DecisionTime + CFMT + Justification + Confidence:LineupRace + Confidence:Delay + Confidence:DecisionTime + Confidence:CFMT + Confidence:Justification + DecisionTime:CFMT + DecisionTime:Justification + CFMT:Justification + (1|Participant). The absolute fit indices indicate that this model adequately fit the data (KS *D* = .017, *p* = .410; pseudo-$R^2_{\text{GLMM(m)}}$ = .365), as did visual inspection of the residual plots.

Likelihood ratio tests showed significant main effects of lineup-race, $\chi^2(1) = 6.08$, p = .014, delay, $\chi^2(1) = 11.75$, p = .001, confidence, $\chi^2(1) = 20.20$, p < .001, face-recognition ability (i.e., CFMT score), $\chi^2(1) = 20.96$, p < .001, and justification-type, $\chi^2(1) = 14.49$, p = .013. The effect of delay reflects higher accuracy in the 5-minute (44.4%, 95% CI [39.6, 49.2]) compared to the one-day condition (33.4%, 95% CI [29.4, 37.7]). Other significant effects were all moderated by two-way interactions, which we describe below. The main effect of Decision-time (p = .294), and the interactions between Confidence and Delay (p = .096), Decision-time and CFMT (p = .155), and CFMT and Justification (p = .054) are non-significant. The four panels in Figure 1A.3 show how identification accuracy changes as a function of both the participant’s level of confidence in their identification and (a) their face recognition ability

(CFMT score), (b) their decision-time, (c) the lineup-race and (d) the justification for their decision, respectively. In each of these figures, the lines represent the mixed-effects model's estimates, with the shading representing the 95% confidence interval.

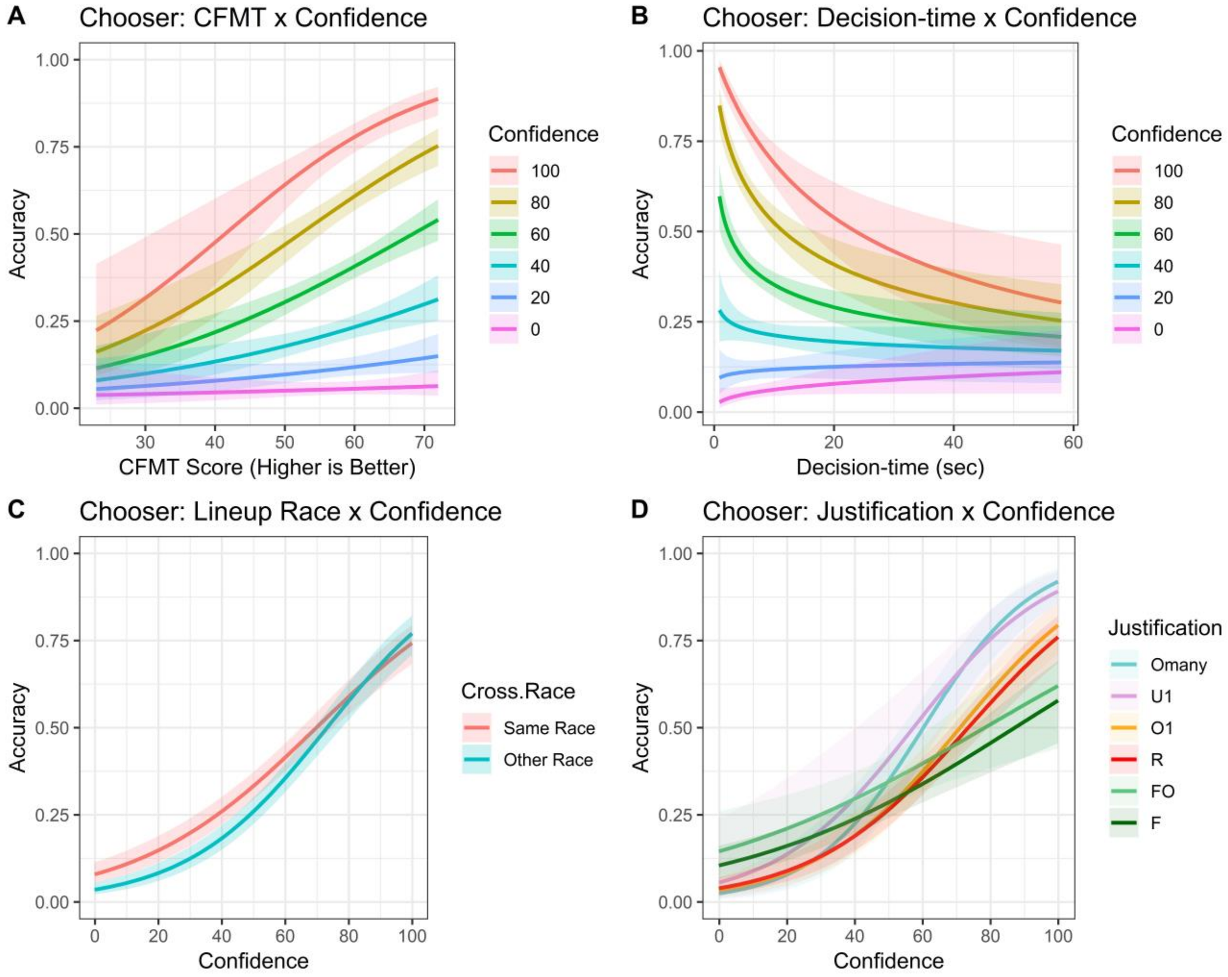


Figure. 1A.3. Two-way interactions between Confidence and (A) CFMT, (B) Decision-time, (C) Lineup Race, and (D) Justification type in the chooser model. Lines represent model estimates, with error shading representing the 95% confidence interval. Notably, high confidence errors are more pronounced when participants are worse face recognizers (A), take longer to make a decision (B), and/or use F/FO as the basis for selecting a face (D).

Figure 1A.3a shows the interaction between face recognition ability (CFMT score) and confidence, $\chi^2(1) = 4.54$, p = .033. Poor face recognizers (i.e., individuals with lower CFMT scores) are less able than strong face recognizers to use confidence ratings to distinguish between correct and incorrect identifications. But, the result that we want to emphasize involves high confidence responses. Figure 1A.3a clearly shows that when individuals are 100% confident in their identification there is a drop-off in accuracy with steadily decreasing CFMT scores. Poor face recognizers are much more prone to make high confidence misidentifications than are strong face recognizers.

Figure 1A.3b shows that relatively fast and highly confident identifications are more accurate than slower and less confident identifications, replicating past research (Dodson & Dobolyi, 2016; Sauerland & Sporer, 2007, 2009). But, the interaction between Decision-time and Confidence, $\chi^2(1) = 17.48$, $p < .001$, reflects the strong increase in high confidence errors that occurs with longer decision times. Although the highest confidence responses (i.e., the solid red line in Figure 1A.3b) are close to 100% accurate when they occur within a few seconds, the accuracy of these highest confidence identifications decreases to roughly 50% when decision-time is delayed to 20s. There is no comparable drop off in accuracy with increasing decision-time for moderate to low confidence responses. Essentially, highly confident but slow identifications are vulnerable to being wrong.

The interaction between confidence and lineup-race is shown in Figure 1A.3c, $\chi^2(1) = 6.12$, $p = .013$. Identification accuracy is worse for cross-race than same-race lineups when individuals are of moderate to low confidence in their identification than when they are highly confident – an effect that is consistent with past studies (e.g., Dodson & Dobolyi, 2016; Nguyen & Pezdek, 2017; Wixted & Wells, 2017). Put another way, highly confident identifications are less influenced by the cross-race effect.

Figure 1A.3d shows that identification accuracy depends on both confidence and the justification for the identification, as reflected by the interaction between these factors, $\chi^2(5) =$

28.14, p < .001. Consistent with Dobolyi & Dodson (2018), there is a stronger relationship between confidence and accuracy –shown by a steeper line in Figure 1A.3d – when individuals refer to observable (O1 + Omany; e.g., I remember his eyes) or unobservable (U1; e.g., He looks like my cousin) features about the suspect than when they refer to familiarity (F; e.g., He's familiar). Moreover, there are more high confidence errors when individuals provide a familiarity (F) or a familiarity-observable justification (FO, e.g., His chin is familiar) than when they provide any of the other justification-types.

Finally, Figure 1A.4 shows that the predictive value of the different justification-types is stronger at faster than at slower decision-times, as reflected by the interaction between decision-time and justification-type, $\chi^2(5) = 12.01$, p = .035. For clarity, we removed the Unobservable (U1) category from the figure because of the lack of data at the longer decision-times for this justification. References to many observable features (Omany) are associated with identifications that are over 80% accurate when the identification is made quickly. But, as seen in Figure 1A.4, the accuracy associated with this justification-type drops below 40% when this identification is made slowly (> 10 s).

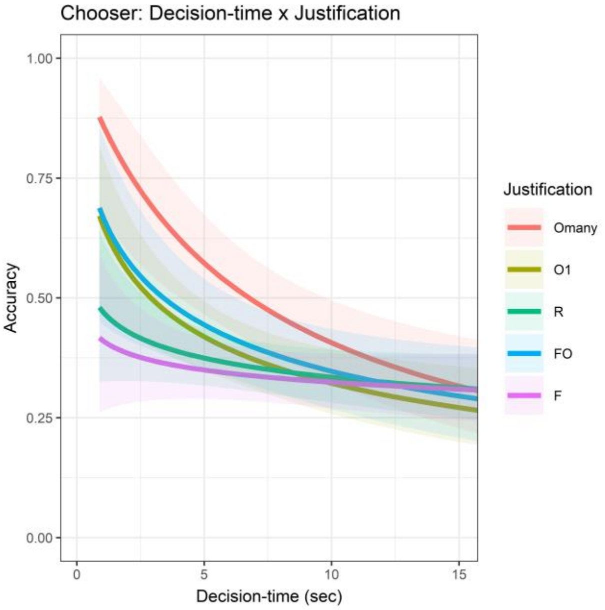


Figure. 1A.4. Interaction pattern between Decision-time and Justification type. Lines represent model estimates, with error shading representing the 95% confidence interval. Discerning accuracy seems to be more useful for fast responses than slow responses, where there is little differentiation between the justification types.

Non-Chooser model.

Non-chooser accuracy is modeled as the rate of correct rejections from target-absent lineups (TAc), relative to the sum of this score and the number of incorrect rejections from target-present lineups ('miss'; TPm) (i.e., (i.e., TAc/[TAc+TPm]). As shown in Table 1A.1, nearly all justifications (97.0%) for a Not Present response were based on the lack of either Familiarity (F) or Recognition (R), consistent with Dobolyi & Dodson (2018). Consequently, our modeling analysis consisted of these two justification-types as there is too little data to include the other justification-types.

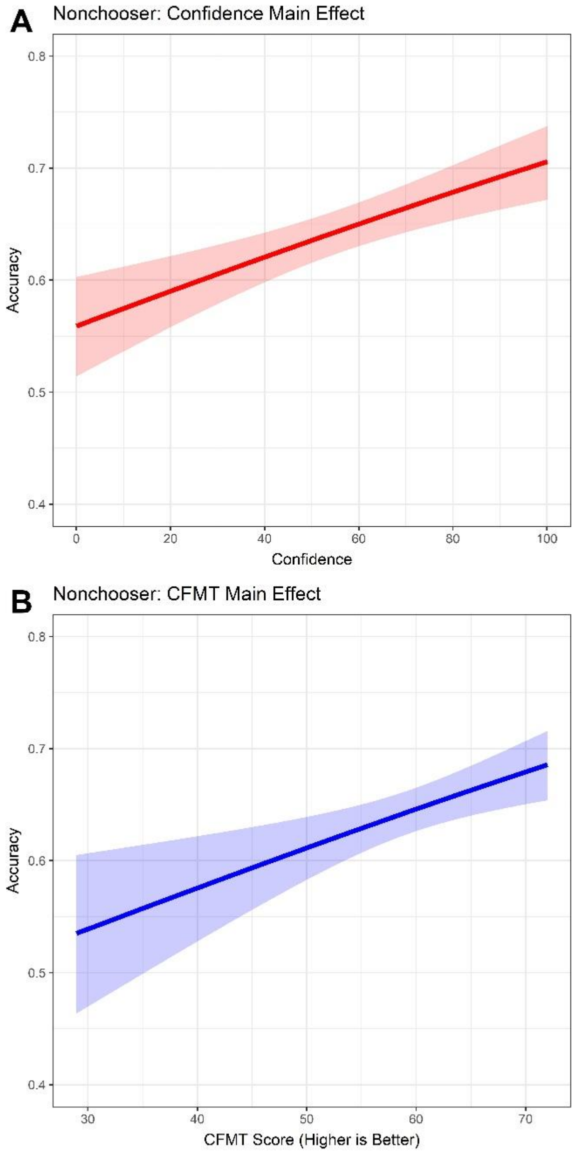


Figure. 1A.5. A) Confidence and B) CFMT main effects on non-chooser accuracy. Lines represent model estimates, with error shading representing the 95% confidence interval. Notably, performance improves with higher levels of confidence, and greater face recognition ability.

The best-fitting model of non-chooser accuracy is represented in Wilkinson-Rodgers notation as: Accuracy ~ LineupRace + Confidence + Delay + DecisionTime + CFMT + Justification + Confidence:CFMT + DecisionTime:CFMT + (1|Participant). Visual inspection of the residual plots and KS tests showed that this model fit the data (KS *D* = .014, *p* = .758). However, the marginal pseudo-R^2 was considerably lower than in the Chooser model (pseudo-$R^2_{GLMM(m)}$ = .019). Given that our relative fit measure (i.e., AIC) and two out of three absolute fit indices supported proper model specification, we proceeded with this non-chooser model.

We found the expected relationship between delay and accuracy, with participants exhibiting higher accuracy in the 5-minute condition (66.5%, 95% CI [63.7, 69.1]) than the one-day condition (62.2, 95% CI [59.4, 64.9]), $\chi^2(1) = 4.78$, $p = .029$.

Additionally, non-chooser accuracy improved as participants expressed more Confidence, $\chi^2(1) = 18.20$, $p < .001$. As presented in Figure 1A.5, accuracy steadily rises as confidence increases, improving by nearly 15% from 0% to 100% confidence. This finding conflicts with multiple previous studies examining confidence and non-chooser accuracy (e.g., Dobolyi & Dodson, 2018; Sauerland & Sporer, 2009). We speculate on the reasons for this discrepancy in the Study 1A Discussion.

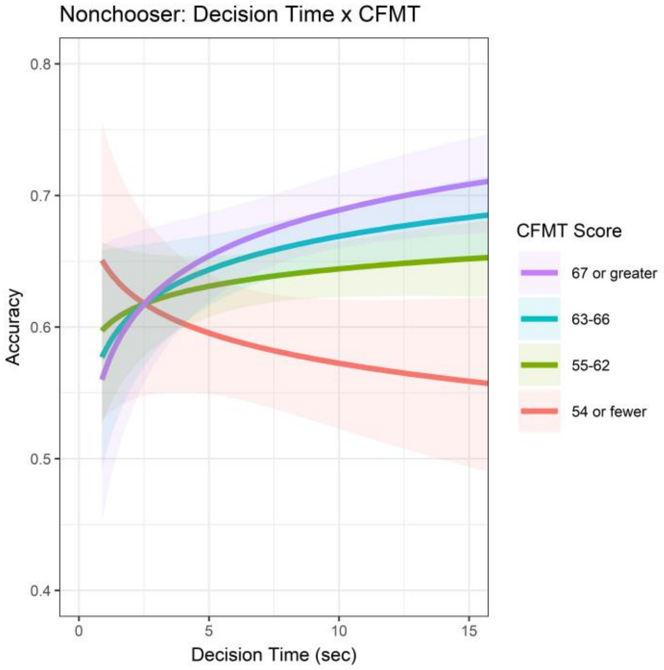


Fig. 1A.6. Two-way interaction between decision-time and CFMT score. Lines represent model estimates for the 0-25th, 25-50th, 50-75th, and >75th percentiles of CFMT performance. Error shading represents the 95% confidence interval. Performance is comparable across face recognition ability for fast decisions, but poor face recognizers show worse accuracy over time.

The main effect of CFMT, $\chi^2(1) = 10.30$, $p = .001$, reflects improved non-chooser accuracy with stronger face recognition ability. As shown in Figure 1A.5, those with the median CFMT score (i.e., 61) show worse non-chooser performance (~65%) than do those with scores only one median deviation higher (i.e., 70) (~68%). However, this finding is qualified by a weak interaction between face recognition ability and decision-time, $\chi^2(1) = 4.58$, $p = .032$. This interaction suggests that performance is comparable across face recognition ability for quick decisions, but poorer recognizers show worse accuracy with increasing decision-time (see Figure 1A.6).

Finally, we found a significant main effect of justification category, $\chi^2(1) = 4.41$, $p = .036$. Familiarity-based rejections (67.3%, 95% CI [63.9, 70.4]) were more accurate than were those based on recognition (62.9%, 95% CI [60.5, 65.2]), although numerically the size of this difference is small. The main effect of decision-time (p = .137) and the interaction between confidence and CFMT (p = .091) are both non-significant.

Suspect-Id Model

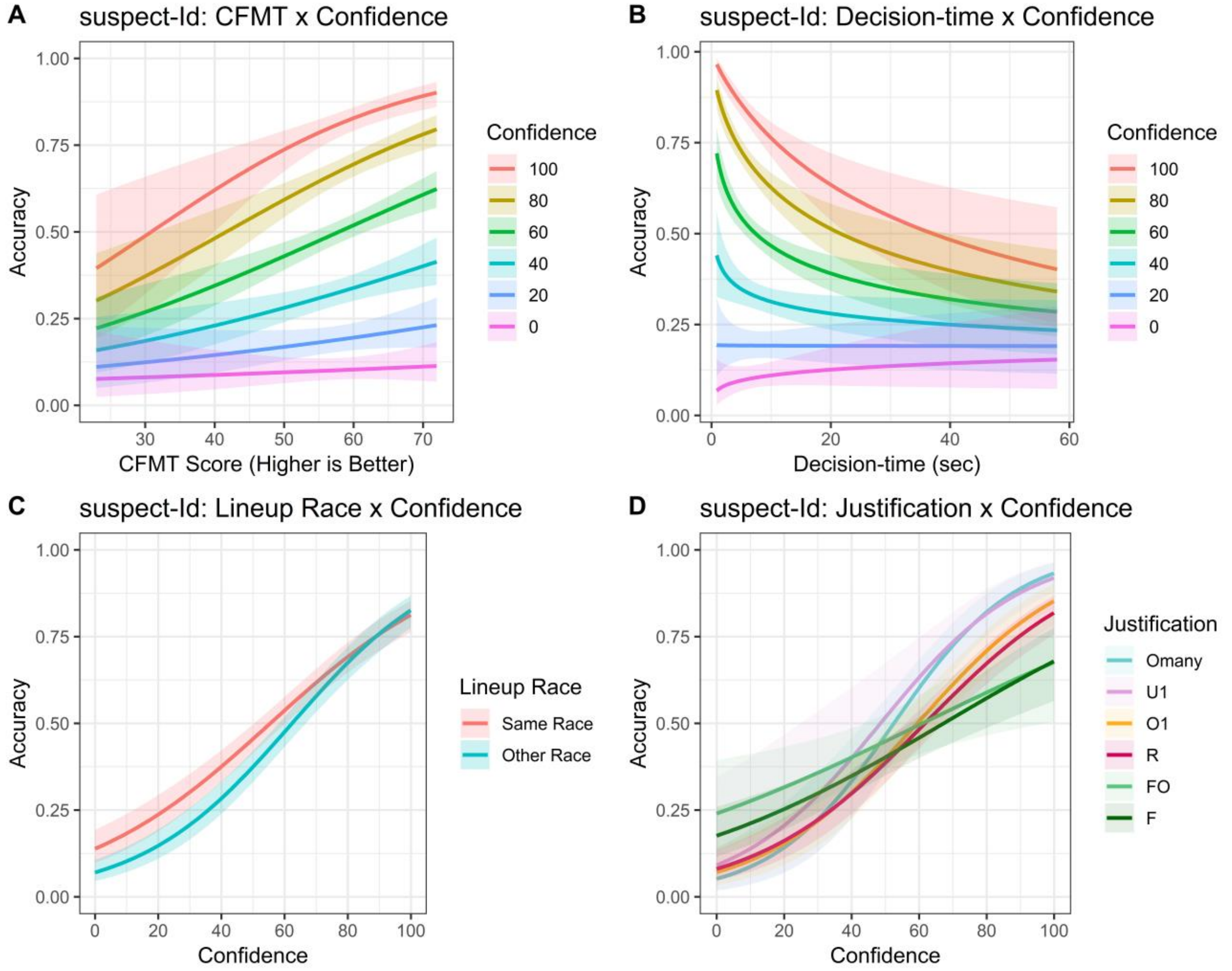
Mickes (2015; see also Wixted & Wells, 2017) has argued that identification accuracy should be measured as the rate of correct identifications relative to the sum of this value and foil identifications from target-absent lineups – a score known as suspect ID accuracy (i.e., TPc/[TPc+(TAfa/6)] for fair lineups). The reason why responses to foils from target-present lineups (TPfa) are excluded in suspect-ID accuracy is because police know that target-present foils are innocent individuals. Thus, suspect-ID accuracy duplicates the perspective of law enforcement: given that an individual has been identified, what is the probability that this individual is the guilty suspect (i.e., TPc) and not an innocent suspect (i.e., TAfa/6 with fair lineups).

Because our modeling procedure does not allow for the suspect-Id adjustment without a substantial loss of TAfa responses (e.g., removal of 5/6 of the false alarm responses), we analyzed a quasi-suspect-Id accuracy score: the ratio of correct responses to target present lineups [i.e., TPc] over the sum of TPc and false alarms to target absent lineups [i.e. TPc/(TPc + TAfa)].

We examined suspect-Id accuracy using the same backward stepwise procedure detailed in the main document. Written in Wilkinson-Rodgers notation, the best fitting model of suspect-Id accuracy consists of several main effects and two-way interactions: Accuracy ~ LineupRace + Confidence + Delay + DecisionTime + CFMT + Justification + LineupRace:Confidence + Confidence:DecisionTime + Confidence:CFMT + Confidence:Justification + DecisionTime:CFMT + DecisionTime:Justification + (1|Participant). Both computed absolute fit indices supported that this model adequately explained the data (KS $D = .013$, $p = .812$, pseudo-$R^2_{GLMM(m)} = .353$), as did visual inspection of the residual plots.

Likelihood ratio tests showed comparable patterns to the identification accuracy model. There were significant main effects of lineup-race, $\chi^2(1) = 4.42$, p = .036, delay, $\chi^2(1) = 6.07$, p = .014, confidence, $\chi^2(1) = 16.04$, p < .001, CFMT, $\chi^2(1) = 32.39$, p < .001, and justification-type, $\chi^2(5) = 14.07$, p = .015. As expected, the main effect of delay reflects better accuracy in the 5-minute (56.8%, 95% CI [52.3, 61.1]) than the 1-day (49.2%, 95% CI [44.9, 53.6]) condition.

Crucially, we highlight the similar interactions patterns between confidence and (a) CFMT, $\chi^2(1) = 3.13$, p = .077, (b) decision-time, $\chi^2(1) = 12.92$, p < .001, (c) lineup-race, $\chi^2(1) = 4.08$, p = .043, and (d) justification-type, $\chi^2(5) = 24.37$, p < .001. As seen in Figure 1A.7a-d, these suspect-Id results are consistent with the identification accuracy model. Specifically, high confidence is associated with more errors for (a) poor face recognizers, (b) slower decision times, and (d) F/FO justifications, but also diminished other-race effects (c). All other effects are non-significant (ps > .071).



1A.7. Suspect-Id interactions between Confidence and (A) CFMT, (B) Decision-time, (c) Lineup Race, and (D) Justification-type. Lines represent model estimates, with error shading representing the 95% confidence interval.

Study 1A Discussion

Recent research suggests that high confidence eyewitness identifications are generally reliable (Wixted & Wells, 2017). Our study adds important caveats to this assessment. We document three factors that are systematically related to high confidence misidentifications: (a) the speed of the decision, (b) the basis for an identification from a lineup, and (c) face recognition ability.

Decision-time is strongly related to high confidence misidentifications. Consistent with past studies (e.g., Brewer & Wells, 2006; Dodson & Dobolyi, 2016; Sauerland & Sporer, 2007, 2009), we observed that fast and confident identifications – presented in Figure 1A.3b -- are many times more accurate than fast and unconfident identifications. But, the key point is that there is a sharp increase in high confidence errors with longer decision times. Whereas highest confidence (100%) identifications made in the initial seconds are nearly always accurate, these identifications fall to nearly 75% accuracy when decision-time increases to 6 seconds and after 20 seconds these reports are roughly 50% accurate (see Brewer & Wells, 2006; Sauerland & Sporer, 2009 for a similar pattern). As Dodson and Dobolyi (2016) suggest, participants appear to adopt an increasingly liberal criterion for making high confidence identifications with increasing decision-time – causing an increase in high confidence errors.

Additionally, consistent with Dobolyi & Dodson (2018), familiarity justifications are more frequently associated with high confidence misidentifications than are justifications that refer to either an expression of recognition, or (un)observable feature(s) about the suspect. Moreover, this relationship persisted across a longer delay than previously studied, and after accounting for the effects of face recognition ability. With both the Department of Justice (Yates, 2017) and the National Academy of Sciences (National Research Council, 2014) advising law enforcement to note the exact wording of an eyewitness's identification, our finding provides investigators with an additional layer of information with which to assess witness credibility.

Finally, for the first time, we show that the Cambridge Face Memory Test predicts the likely accuracy of high confidence identifications. Poor face recognizers are much more vulnerable than strong face recognizers to make high confidence misidentifications. Even when individuals are 100% confident, Figure 1A.3a shows that the average face recognizer (i.e., median CFMT score of 61) is much more likely than the strongest face recognizers (i.e., CFMT score of 72) to make a high confidence misidentification – with below-average face recognizers even more vulnerable to making high confidence errors.

This finding supports the 'optimality' account, wherein the predictive value of a confidence statement is directly tied to the quality of the face representation (Deffenbacher, 1980). As poorer face recognizers encode less robust representations of target faces, high confidence is a less reliable indicator of accuracy than for better recognizers. However, as a counterpoint to the optimality account, many studies find that eyewitnesses adjust their use of high confidence ratings to maintain impressive levels of accuracy in non-ideal encoding conditions, such as lengthy retention intervals, and increased viewing distances (Semmler et al., 2018; Wixted & Wells, 2017). Further research will be necessary to disentangle these accounts, especially studies incorporating measures of individual differences.

An additional question that needs further clarification is why poor face recognizers use high confidence ratings for (presumably) weak face representations. As the present experiment was not designed to answer this question, we can only speculate. However, a large body of literature shows that people can severely overestimate their competence when they perform poorly on a task, and correspondingly exhibit overconfidence (e.g., Kruger & Dunning, 1999;

Lichtenstein & Fischhoff, 1977). These errors occur most frequently in content areas that people lack knowledge, and/or receive minimal feedback on performance. Although it seems like there should be *consistent* feedback on face recognition ability (e.g., embarrassingly introducing oneself to a person met the night before), there is an ongoing debate about the degree to which people have insight into their face recognition ability (Bobak, Mileva, & Hancock, 2018; Gray, Geoffrey, & Richard, 2017). It is conceivable that poor recognizers underestimate the extent of their deficiency, and/or place undue emphasis on non-diagnostic memory signals.

With respect to non-identifications, we highlight two factors that were related to the accuracy of a “not present” response. First, stronger face recognizers (i.e., higher CFMT scores) were more accurate at correctly rejecting lineups than were poorer face recognizers, presumably because their more robust representations of previously seen faces allowed them to recognize when a target individual was absent from a lineup.

Second, contrary to research that has observed little relationship between confidence and non-chooser accuracy (e.g., Dodson & Dobolyi, 2016; Sauerland & Sporer, 2009), we found that confidence in non-chooser decisions was informative, such that highly confident rejections were more often correct than were low confidence rejections. But, consistent with previous findings, confidence is a stronger predictor of chooser accuracy than non-chooser accuracy (e.g., Brewer & Wells, 2006). We believe that the conflicting findings about confidence and non-chooser accuracy between this study and previous work stems from our decision to model chooser and non-chooser responses separately. To illustrate this point, we followed past studies and constructed a single model of chooser and non-chooser accuracy and found that confidence did not significantly predict non-chooser accuracy. However, there are qualitative differences between chooser and non-chooser decisions, as evidenced by changes in the relative use of justification categories, which suggests individuals may adjust how they use the confidence scale in these two situations. Reinforcing the impact of the modeling procedure, Wixted and Wells (2017) isolated non-chooser responses from a dataset provided by Wetmore et al. (2015), and similarly found that high confidence rejections were more accurate than were those made with lower confidence.

In sum, existing research on eyewitness identification has focused on the average individual and has shown that a participant's confidence rating about an identification is informative of its accuracy (Wixted & Wells, 2017). We show that high confidence identifications do not protect against the increase in errors that accompany poorer face recognition ability, increasing decision-time or the use of familiarity as a justification for a response. Taken together, this study suggests that the justice system should take both individual differences and confidence into account when determining the likely accuracy of an eyewitness decision.

Study 1A References

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36-40.

Barton, K. (2018) MuMIn: Multi-model inference. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014*). lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-21.

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96-103.

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2018). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*, https://doi.org/10.1177/1747021818776145.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11-30.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.

Cho, S. J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., Van Gulick, A. E., ... & Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological assessment*, *27*(2), 552-566.

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

De Lestrade, J. X. (2001). *Murder on a Sunday Morning*. Docurama.

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*(4), 345-357.

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

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and Eyewitness Identifications: The Cross-Race Effect, Decision Time and Accuracy. *Applied Cognitive Psychology*, *30*(1), 113-125.

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*(4), 576-585.

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67*(5), 818.

Fox, J. (2003). Effect displayes in R for generalized linear models. *Journal of Statistical Software*, *8*(15), 1-27.

Gray, K. L., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population. *Royal Society open science*, *4*(3), https://doi.org/10.1098/rsos.160923.

Hartig, F. (2018). *DHARMa: Residual diagnostics for hierarchical (mulit-level/mixed) regression models*. R package version 0.2.0.

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*(6), 1121-1134.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183. doi:10.1016/0030-5073(77)90001-0

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*(2), 295-314.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3-35.

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

Morgan III, C. A., Hazlett, G., Baranoski, M., Doran, A., Southwick, S., & Loftus, E. (2007). Accuracy of eyewitness identification is significantly associated with performance on a standardized test of face recognition. *International Journal of Law and Psychiatry*, *30*(3), 213-223.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: The National Academies Press.

Nguyen, T. B., Pezdek, K., & Wixted, J. T. (2017). Evidence for a confidence–accuracy relationship in memory for same-and cross-race faces. *The Quarterly Journal of Experimental Psychology*, *70*(12), 2518-2534.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic bulletin & review*, *16*(2), 252-257.

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*(4), 337-347.

Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime & Law*, *13*(6), 611-625.

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*(1), 46-62.

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied, 24*(3), 400-415.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *afex: Analysis of factorial experiments*. R package version 0.21-2.

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*(1), 102.

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*(1), 8-14.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Applied Statistics*, *22*, 392-399. doi: 10.2307/2346786

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225-230.

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*(5-6), 360-392.

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*(11), 5238-5241.

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences*, *113*(2), 304-309.

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*(2), 192-203.

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*(1), 10-65.

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*(1), 137.

Yates, S.Q. (2017, Jan 6). Memorandum for heads of department law enforcement components all department prosecutors. *Subject: Eyewitness identification: Procedures for conducting photo arrays*. https://www.justice.gov/archives/opa/press-release/file/923201/download.

Study 1B. Stark Individual Differences: Face Recognition Ability Influences the Relationship Between Confidence and Accuracy in a Recognition Test of Game of Thrones Actors (Grabman & Dodson, *submitted*)

Most people have experienced the embarrassment of greeting a stranger as if they were a recent acquaintance. Whether we risk this social faux pas depends on our certainty that we previously encountered this individual. In higher stakes contexts, eyewitness confidence has profound effects on the criminal justice system. Juror decisions are strongly influenced by confidence (Brewer & Burke, 2002), and judges are instructed to use certainty as an indicator of whether to admit the witness's testimony in court (*Neil vs. Biggers*, 1972). The question is how probative confidence is of face recognition accuracy.

In an influential review of the eyewitness literature, Wixted and Wells (2017) found that high confidence identifications are generally accurate. This relationship holds over changes in retention interval (i.e., the amount of time between study and test) (see Wixted, Read, et al., 2016 for a review), exposure duration (i.e., the amount of time a face is viewed at encoding) (e.g., Palmer, Brewer, Weber, & Nagesh, 2013), and a variety of other manipulations (see Wixted & Wells, 2017 for a review). However, there is a compelling need for studies of the confidence-accuracy relationship which capture the richness of the real-world face viewing experience.

The fact that the average person can recognize thousands of unique faces (Jenkins, Dowsett, & Burton, 2018) masks aspects of this task that are remarkably complex. Faces are encountered in a myriad of contexts, often with considerable changes in lighting, orientation, and other characteristics (e.g., hair, age, clothing, etc.). While the majority of people can easily recognize family members and friends in a variety of situations, this task is far more challenging for unfamiliar faces (Kramer, Young, & Burton, 2018). As some examples of this difficulty, growing literature suggests that minimal disguises (such as sunglasses) can impair face recognition accuracy (Mansour, Beaudry, & Lindsay, 2017; Nguyen & Pezdek, 2017; Righi, Peissig, & Tarr, 2012; Terry, 1994). Moreover, studies in the face matching literature (i.e., indicating whether two simultaneously presented faces are the same person or different people), show that subtle changes in viewing conditions (e.g., photos of the same person taken with different cameras) can substantially decrease matching decision accuracy (see Young & Burton, 2017 for a review).

Given the complexity of real-world face recognition, claims about the value of high confidence are complicated by multiple factors. First, participants in past studies generally knew that they were in an experiment, which potentially alters their face encoding strategies. Second, exposure durations are shorter than those experienced in everyday life (e.g., 90-seconds), and retention-intervals rarely longer than a few weeks (though see Read, Lindsay, & Nicholls, 1998 for an exception). Third, most studies use single-trial designs, which limits conclusions to the small group of people presented. Finally, there is typically a single context for encoding faces, whereas in practice we must learn to recognize people (often encountered more than once) in varied environments.

Additionally, a largely ignored aspect of the confidence-accuracy relationship in the eyewitness literature is heterogeneity in unfamiliar face recognition ability (Duchaine & Nakayama, 2006). Skill in this domain ranges from people with developmental prosopagnosia (i.e., face blindness), who may have difficulties recognizing even close family members (J. J. S. Barton & Corrow, 2016), to super-recognizers who are actively recruited to police departments for their face-recognition prowess (Ramon, Bobak, & White, 2019; Russell, Duchaine, & Nakayama, 2009). These differences are highly heritable (Shakeshaft & Plomin, 2015; Wilmer et al., 2010; Zhu et al., 2010), and only weakly associated with general intelligence (Gignac, Shankaralingam, Walker, & Kilpatrick, 2016; Shakeshaft & Plomin, 2015; Wilhelm et al., 2010; Zhu et al., 2010).

Multiple studies show that higher face recognition ability predicts increased accuracy in eyewitness identification tasks (Andersen, Carlson, Carlson, & Gronlund, 2014; Bindemann, Avetisyan, & Rakow, 2012; Morgan et al., 2007). But, only our group has investigated whether this skill influences the probative value of confidence in face recognition tasks. In contrast to previous research documenting a robust confidence-accuracy relationship across a wide range of manipulations, we found that weaker face recognizers are far more likely to make high confidence errors than are stronger recognizers (Grabman, Dobolyi, Berelovich, & Dodson, 2019).

However, there are several aspects that limit the real-world applicability of Grabman et al (2019). Participants viewed static images of faces at encoding and test, which fails to capture the experience of encountering moving people in varied contexts. Moreover, the study used relatively short exposure durations (3 repetitions of 3-seconds) and retention-intervals (up to 1 day). It is possible that the impact of face recognition ability on the confidence-accuracy relationship is minimal with longer exposures or delays. Finally, the stimulus set consisted solely of young adult males, which further limits generalizability.

Given the paucity of studies of the confidence-accuracy relationship under real-world viewing conditions, there are two aims for the current study. The first aim is to determine if the results from a more naturalistic setting mirror those of the carefully designed experiments cited in Wixted and Wells (2017). The second aim is to assess whether differences in face recognition ability influence the confidence-accuracy relationship using a design that addresses each of the short-comings of our previous study (Grabman et al., 2019).

To accomplish these aims, we leveraged a dataset published by Devue, Wride, and Grimshaw (2019), accessed using the Open Science Framework (OSF) (https://osf.io/wg8vx). In this study, participants viewed the first six seasons of the popular television show *Game of Thrones* (GoT) as the series aired, then completed a recognition task of 90 pictures of actors (not in character) intermixed with 90 strangers. Importantly, participants viewed the show for personal entertainment, meaning that all faces are incidentally encoded. Moreover, as Devue et al. (2019) note, there are several additional aspects of GoT that make it an appealing way to study real-world face recognition. Characters are seen in a variety of natural viewing contexts, with often substantial changes in appearance, lighting, clothing, age, and viewpoint. Additionally, screen-time is readily accessible from internet databases, allowing for assessment of exposure duration effects. There are many character deaths throughout the series, resulting in lengthy retention intervals between encoding and test for some actors. Finally, there are over 600 actors listed in the show credits, which provides a substantial face corpus from which to prepare stimuli.

From the standpoint of the current study aims, this dataset offers some additional advantages. Each participant completed a standard test of face-recognition, the Cambridge Face Memory Test+ (CFMT+), and provided confidence ratings for each decision. While the original authors examined associations between these variables and accuracy using correlational analysis, we use calibration curves, which are superior for assessing confidence-accuracy calibration (Wixted & Wells, 2017). And, for the first time, we analyze the conjunctive effects of confidence and face recognition ability on accuracy under real-world viewing conditions.

Additionally, whereas eyewitness studies typically use a criminal lineup paradigm, participants in Devue et al (2019) completed an old-new recognition task. As far as we are aware, only one other study has used calibration curves to examine the confidence-accuracy relationship in an old-new face recognition paradigm for a large set of items (> 100 trials) (Tekin & Roediger, 2017). These researchers used a single exposure duration (2-seconds) and a short retention-interval (10 min), and found highest confidence identifications to be about 96% accurate. It is an open question whether this impressive accuracy generalizes to uncontrolled settings with longer retention-intervals and differing levels of exposure.

Finally, the use of another group's dataset carries the benefit of reducing 'researcher degrees of freedom'. If stronger face recognizers continue to make fewer high confidence errors than weaker recognizers in an uncontrolled, naturalistic context then this bolsters claims that there are robust associations between face recognition ability, confidence, and accuracy.

Methods

Participants.

Characteristics of the participants are reported in Devue et al., (2019). Briefly, the results are comprised of 32 participants (20 women and 12 men), aged between 19 and 56 years (*M* = 28.7 years ± 10.5), who completed the task 3-6 months after the end of the sixth season of GoT. All participants watched six seasons of GoT once, and in order as the show aired, with the exception of some who viewed both Seasons 1 and 2 during the same year. While the sample size is low, the large number of trials per participant (n = 168) fits with current recommendations for the logistic mixed effects analysis outlined in the Results section (e.g., McNeish & Stapleton, 2016).

Materials.

Cambridge Face Memory Test + (*CFMT+*). The CFMT+ is a frequently used test that assesses poor to superior face recognition ability (Russell et al., 2009). Participants memorize six male faces in three separate orientations. For each trial, previously viewed faces must be selected from an array of the target face and two foils. The test phase proceeds across 102 trials in five increasingly difficult blocks. Difficulty is manipulated with the use of novel images, visual noise filters, different levels of cropping, and (eventually) the use of a profile view with extra levels of noise. Scores can range from 0 – 102 correct responses, but in practice a score of 34 represents random guessing.

Face Stimuli. Extensive details about the generation of the study materials are provided in Devue et al., (2019), with the materials themselves available on the OSF platform (https://osf.io/wg8vx). The researchers selected 84 actors from GoT from 15 conditions, consisting of the interaction between retention-interval since last viewing (Season 6, 5, 4, 3, 1/2) and three levels of exposure: 'lead characters' [20 – 90 min screen time], 'support characters' [9 – 19 min], and 'bit parts' [<9 min, but role in story for 1 – 3 episodes]. Additionally, 6 characters categorized as 'main heroes' [> 123 min screen time] survived to the end of the sixth season, with the actors serving as training trials for the task. Ninety pictures of unfamiliar faces were collected to serve as foils (i.e., 'new' trials), and "matched the actor set in terms of head orientation, age range, facial expression, attractiveness, presence of make-up, facial hair, or glasses, hairstyle, clothing style, lighting, and picture quality" (Devue et al., 2019). While foils matched the characteristics of the sample of actors as a whole, they were not individually paired to specific actors.

In a similarity manipulation, half of the participants viewed photos of the actors which were similar to their last appearance on the show (*similar*), while the other half viewed photos that were as different as possible (*dissimilar*). These similarity groups were matched on CFMT+ scores, age, and gender. Due to the scarcity of available photos for 'bit part' actors, all participants responded to both *similar* (17 trials) and *dissimilar* (13 trials) pictures for this exposure level, regardless of their assigned similarity condition.

Procedure.

Full details of the procedure are outlined in Devue et al., (2019), so we mention only those pertinent to the present study. Participants completed all tasks on a computer. Following the CFMT+, participants were assigned to a similarity condition, and then started the GoT face recognition task. An easy block consisting of the six 'main heroes' and six foils served to practice the task, and was followed by 168 test trials consisting of 84 actors intermixed with 84 foils. Each trial started with a fixation cross (500 ms), followed by a picture stimulus that remained in the center of the screen until the participant's response or up to 3,000 ms. Participants pressed the 'K' key to indicate they had 'seen' the face before (in GoT or elsewhere), or pressed 'L' to indicate that the face was 'new'. They then provided a confidence rating for this decision using a 5-point scale (1 = *not at all confident*, 5 = *totally confident*).

Results

Data preparation.

Following the lead of the original authors, we discarded 26 trials where participants indicated they recognized an actor from outside of GoT, as well as the training trials (6 'main heroes' + 6 foils per participant). One trial was omitted due to a typo (i.e., score of '2' on accuracy, when only 0 and 1 were possible). We also removed all trials where participants responded in < 300 ms (n = 371; 6.9% of total trials), as this is faster than consistent findings on the time to process face identity, along with the additional time needed to perform a keystroke (e.g., Gosling & Eimer, 2011). In total, this left 4,979 responses from 32 participants. We have uploaded the data file used for the analysis to the OSF platform, along with a cleaned version of the original Devue et al. (2019) file that is more conducive toward coding environments (e.g., R, Python) (https://osf.io/quhsg).

Table 1B.1 shows the breakdown of the frequency of responses into Hits ("Seen"|Actor), Misses ("New"|Actor), Correct Rejections (CR; "New"|Foil), and False Alarms (FA; "Seen"|Foil) by confidence level and a median split of CFMT+ performance, which we categorize as Weaker Face Recognizers (CFMT+ scores of 52-73) and Stronger Face Recognizers (CFMT+ scores of 74-90). Due to low frequencies of responses in confidence categories 1 and 2, we collapsed these levels to form a single confidence level ('1-2').

CFMT+	Confidence	Hit	miss	fa	cr
Weaker Face Recognizers [52,73]	1-2	77	142	81	193
	3	196	257	149	348
	4	174	212	75	384
	5	236	117	28	141
Stronger Face Recognizers [74,90]	1-2	44	96	25	112
	3	104	189	52	290
	4	103	183	28	349
	5	222	131	4	213

Table 1B.1. Frequency of responses of Hits (Seen|Actor), Misses (New|Actor), Correct Rejections (CR; New|Unfamiliar), and False Alarms (FA; Seen|Unfamiliar) categorized by confidence level and CFMT+ Median split.

Tables 1B.2 and 1B.3 show the frequencies of hits, misses, correct rejections, and false alarms across CFMT+ median split for the exposure duration and retention-interval manipulations, respectively. Due to the single-block design, the same foil counts (i.e., false alarms and correct rejections) are present in all levels of these within-subjects manipulations. To obtain an adequate trial count for the retention-interval contrasts (especially at the upper-end of the confidence scale), we recoded this variable into 'Long Delay' (Seasons 1-3; 34 actors), 'Medium Delay' (Seasons 4-5; 32 actors), and 'Short Delay' (Season 6; 18 actors) conditions, based on comparable discriminability within these time periods. The exposure duration contrast is composed of 'leading actors' (longest exposure; 27 actors), 'supporting actors' (medium exposure; 27 actors), and 'bit parts' (shortest exposure; 30 actors).

Finally, Table 1B.4 shows the counts for the between-subjects similarity manipulation. We removed 'bit part' actors who did not match the condition assigned to the participant (e.g., *dissimilar* 'bit part' photos in the *similar* condition). Note that removing the 'bit part' actors causes a slight difference in the total actor counts (i.e., hits + misses) for the similarity manipulation as compared to the total count for the full sample and the other manipulations.

CFMT+	Confidence	Exposure	hit	miss	fa	cr
Weaker Face Recognizers [52,73]	1-2	'Bit Parts'	28	61	81	193
		'Supports'	25	51		
		'Leads'	24	30		
	3	'Bit Parts'	73	138	149	348
		'Supports'	63	76		
		'Leads'	60	43		
	4	'Bit Parts'	26	115	75	384
		'Supports'	75	60		
		'Leads'	73	37		
	5	'Bit Parts'	13	53	28	141
		'Supports'	62	37		
		'Leads'	161	27		
Stronger Face Recognizers [74,90]	1-2	'Bit Parts'	15	41	25	112
		'Supports'	21	31		
		'Leads'	8	24		
	3	'Bit Parts'	34	97	52	290
		'Supports'	38	62		
		'Leads'	32	30		
	4	'Bit Parts'	19	104	28	349
		'Supports'	43	49		
		'Leads'	41	30		
	5	'Bit Parts'	0	73	4	213
		'Supports'	58	33		
		'Leads'	164	25		

Table 1B.2. Frequency of Hits, Misses, Correct Rejections (CR), and False Alarms (FA), categorized by short ('bit parts'), medium ('supports') and long ('leads') exposures, as well as CFMT+ Median split.

CFMT+	Confidence	Delay	hit	miss	fa	cr
Weaker Face Recognizers [52,73]	1-2	Long	33	71	81	193
		Medium	29	44		
		Short	15	27		
	3	Long	77	122	149	348
		Medium	89	91		
		Short	30	44		
	4	Long	55	98	75	384
		Medium	74	70		
		Short	45	44		
	5	Long	72	37	28	141
		Medium	86	56		
		Short	78	24		
Stronger Face Recognizers [74,90]	1-2	Long	18	47	25	112
		Medium	19	32		
		Short	7	17		
	3	Long	43	77	52	290
		Medium	45	79		
		Short	16	33		
	4	Long	36	74	28	349
		Medium	32	78		
		Short	35	31		
	5	Long	68	59	4	213
		Medium	85	44		
		Short	69	28		

Table 1B.3. Frequency of Hits, Misses, Correct Rejections (CR), and False Alarms (FA) categorized by long (Seasons 1-3), medium (Seasons 4-5) and short (Seasons 6) retention-intervals, as well as CFMT+ Median split.

Similarity	CFMT+	Confidence	hit	miss	fa	cr
Similar	Weaker Face Recognizers [52,73]	1-2	28	62	27	92
		3	54	102	58	175
		4	57	87	36	181
		5	96	73	18	122
	Stronger Face Recognizers [74,90]	1-2	23	39	16	54
		3	41	82	21	144
		4	24	85	5	162
		5	62	64	3	122
Dissimilar	Weaker Face Recognizers [52,73]	1-2	38	48	54	101
		3	105	89	91	173
		4	106	61	39	203
		5	136	12	10	19
	Stronger Face Recognizers [74,90]	1-2	16	32	9	58
		3	51	62	31	146
		4	72	45	23	187
		5	160	30	1	91

Table 1B.4. Frequency of Hits, Misses, Correct Rejections (CR), and False Alarms (FA) categorized by whether actors' looked similar to their last appearance on the show ('similar') or as dissimilar as possible ('dissimilar'), as well as CFMT+ Median split. Note that trial counts do not match Table 1B.1 because of the removal of 'bit part' actors who did not match the condition assigned to the participant (e.g., *dissimilar* 'bit part' photos in the *similar* condition).

Is there a strong relationship between confidence and accuracy in a real-world viewing context?

Devue et al., (2019) analyzed the relationship between confidence and overall accuracy using Pearson's correlation coefficients. This analysis found minimal associations between overall accuracy (centered and scaled) and average confidence on accurate trials (*r* = .125), as well as average confidence on inaccurate trials (r = - .096).

One issue with defining the confidence-accuracy relationship in terms of overall accuracy is that research generally shows a stronger correspondence between confidence and accuracy for identifications (i.e., 'seen' responses) than non-identifications (i.e., 'new' responses) (e.g., Brewer & Wells, 2006). Separating these response types may reveal more robust relationships than previously reported. Additionally, correlation analysis addresses a fundamentally different question than is typically of interest to applied memory researchers (Juslin, Olsson, & Winman, 1996). Whereas correlation coefficients measure covariation, or the tendency for one variable to increase/decrease as another variable increases/decreases, applied researchers are generally more interested in the accuracy of responses made with a particular level of confidence.

As a concrete example of this difference, imagine that a participant provides the highest possible confidence rating to every trial. The correlation between confidence and accuracy is zero because, regardless of whether accuracy increases/decreases, confidence remains the same. However, despite there being zero correlation, the participant would be perfectly calibrated if they were correct on every trial. Given that the participant used the highest possible confidence rating, we observed their response to be correct 100% of the time.

An easy way to visualize the probative value of confidence is with a calibration curve (Tekin & Roediger, 2017; see also Mickes, 2015). Along the X-axis are progressively increasing confidence values. On the Y-axis is a proportion representing the number of correct items over the sum total of items at this level of confidence (i.e., correct / (correct + incorrect)). Points are plotted representing Y-*accuracy* at X-*confidence level*. The slope of the lines connecting the points provides additional information. Upward sloping lines signal increasing accuracy with higher levels of confidence, whereas flat lines indicate little difference in predictive power between two confidence ratings.

Figure 1B.1 shows the calibration curves for all identification ('seen') (hits/[fa + hits]) and non-identification ('new') (cr/[cr + misses]) responses in the GoT task, collapsed across participants. Replicating the eyewitness research, there is clearly a strong positive relationship between higher confidence responses and identification accuracy. The highest confidence level ('5') boasts accuracy rates of 93.5% (95% HDI[1], [89.8, 97.0]), as compared to 53.3% (95% HDI, [46.3, 61.3]) at the lowest level ('1-2'). However, as indicated by the flat line in the right panel, there is little association between confidence and accuracy for non-identifications.

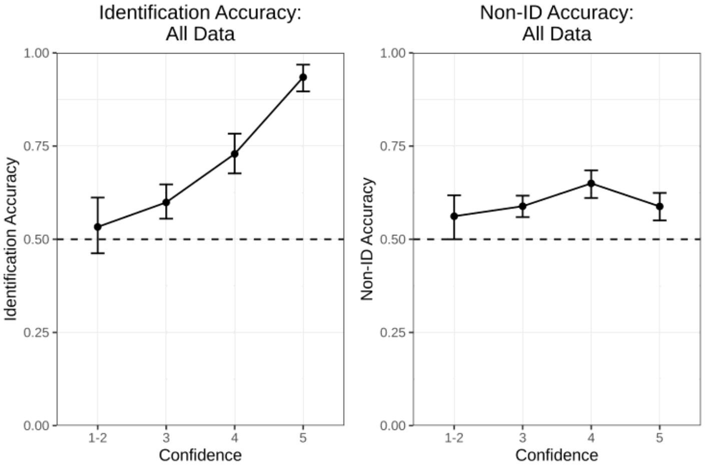


Figure 1B.1. Calibration curves for the full sample of responses. Notably, there is a strong relationship between confidence and accuracy for identifications (left panel), but weaker associations for non-identifications (right panel). The dashed lines at 50% reflect chance accuracy. Error bars reflect 95% HDIs.

[1] Highest Density Intervals (HDI) are presented for consistency with later analyses. These intervals are based on 10,000 bootstrapped resamples and reflect 95% of values where the probability density is greater than points outside these bounds.

Next, we examined the impact of exposure duration (‘leads’ vs. ‘supports’ vs. ‘bit parts’; within-subjects), retention-interval (‘long’ [S1-3] vs. ‘medium’ [S4-5] vs. ‘short’ [S6]; within-subjects), and similarity (‘*similar’* vs. ‘*dissimilar*”; between-subjects) on the predictive value of confidence ratings. We analyzed each of these manipulations separately (i.e., main effects), as there are too few data-points per cell to assess interactions.

Because foils are not matched to specific actors in this single-block design, the same false alarms and correct rejections must be used in (non-)identification accuracy calculations for each condition. However, before computing accuracy scores, we needed to account for the unequal numbers of actor trials across conditions. Without an adjustment, the same hit/false alarm rates (at a given level of confidence) can produce different calibration curves.

For example, imagine that participants respond ‘seen’ to 50% of actor trials and 25% of foil trials with a given level of confidence for both short (18 actors) and medium (32 actors) retention-intervals (i.e., hit rate = 50%, false alarm rate = 25% at this level of confidence). Multiplying out (and assuming no data eliminations), this gives 18 actors * .50 hit rate * 32 participants = 288 hits vs. 32 actors * .50 hit rate * 32 participants = 512 hits for the short and medium conditions, respectively. Naively, these trials would be compared against 84 foils * .25 false alarm rate * 32 participants = 672 false alarms for both groups. Using the formula for identification accuracy [hits / (hits + fa)], we would find accuracy rates of 288 hits / (288 hits + 672 fa) ≈ 43% and 512 hits / (512 hits + 672 fa) ≈ 76%, for the short and medium retention-intervals, respectively. In other words, despite the same use of the confidence scale across conditions, a difference of ~33% emerges due to disparities in the number of actor trials. Moreover, both group’s values are far from the nominal identification accuracy rate expected with a study design implementing equal numbers of actor to foil trials, or .50/ (.50 + .25) ≈ 67%.

To ensure comparability between conditions, we adjusted the frequency of foil trials to match the frequency of actor trials in each condition. Specifically, we multiplied the frequencies of false alarms and correct rejections in a given condition by the ratio of actor trials to the total number of foil trials (adj = $\text{actors}_{\text{condition}}/\text{foils}_{\text{total}}$). Thus, at each level of confidence, adjusted identification accuracy = hits / [hits + adj*fa], and adjusted non-identification accuracy = adj*cr / [adj*cr + misses].

Using the previous example (50% hits, 25% false alarms) with the adjustment produces the correct interpretation of identification accuracy. The frequency of hits remains the same for the short (288 hits) and medium (512 hits) retention-intervals. However, the frequency of false alarms (672 fa) is adjusted in the denominator by the frequency of actor trials in the short (18 actors * 32 participants = 576) and medium (32 actors * 32 participants = 1024) retention-interval conditions over the total number of foil trials (84 trials * 32 participants = 2688 total foils): short ID accuracy = 288 hits /(288 hits + ($576_{\text{short actors}}$ /$2688_{\text{total foils}}$) * 672 fa) ≈ 67% and medium ID accuracy = 512 hits/ (512 hits + ($1024_{\text{medium actors}}$/$2688_{\text{total foils}}$) * 672 fa) ≈ 67%. There is no longer a difference between the groups, and the identification accuracy rates match the nominal rate expected by equal numbers of actor and foil trials.

Due to computation of an adjusted accuracy statistic, binomial confidence intervals are inappropriate for characterizing the uncertainty around the estimated proportion. For this reason, we bootstrapped 95% highest density intervals (HDIs) around each estimate (R functions and example code are available here: https://osf.io/quhsg). This procedure uses a routine that simulates the experimental design by:

1. Calculating the adjusted-accuracy at all levels of confidence for each condition.
2. Randomly sampling participants (with replacement) from the entire dataset (for within-subject only comparisons), or in equivalent number to the size of the between-group comparison (e.g., similarity, CFMT+ score group).
3. Re-calculating the adjusted accuracy statistic from the resampled data.
4. Repeating Steps (2) & (3), 10,000 times.
5. Computing the 95% HDI of the distribution of resampled adjusted-accuracy statistics, using the bayestestR package v. 0.2.5 (Makowski, Ben-Shachar, & Lüdecke, 2019). We use the HDI because it is more agnostic about the symmetry of the resampling distribution than percentile intervals. Specifically, this interval captures modal outcomes, rather than mean outcomes (which are highly influenced by skew). Note that HDI will give the same intervals as a percentile if the resampling distribution is symmetrical.

Figure 1B.2 shows calibration curves for exposure duration (panel a), retention-interval (panel b), and similarity (panel c). The left column displays curves for identifications (hits/[fa*adjustment + hits]), whereas the right column represents non-identifications (cr*adjustment/[cr*adjustment + misses]).

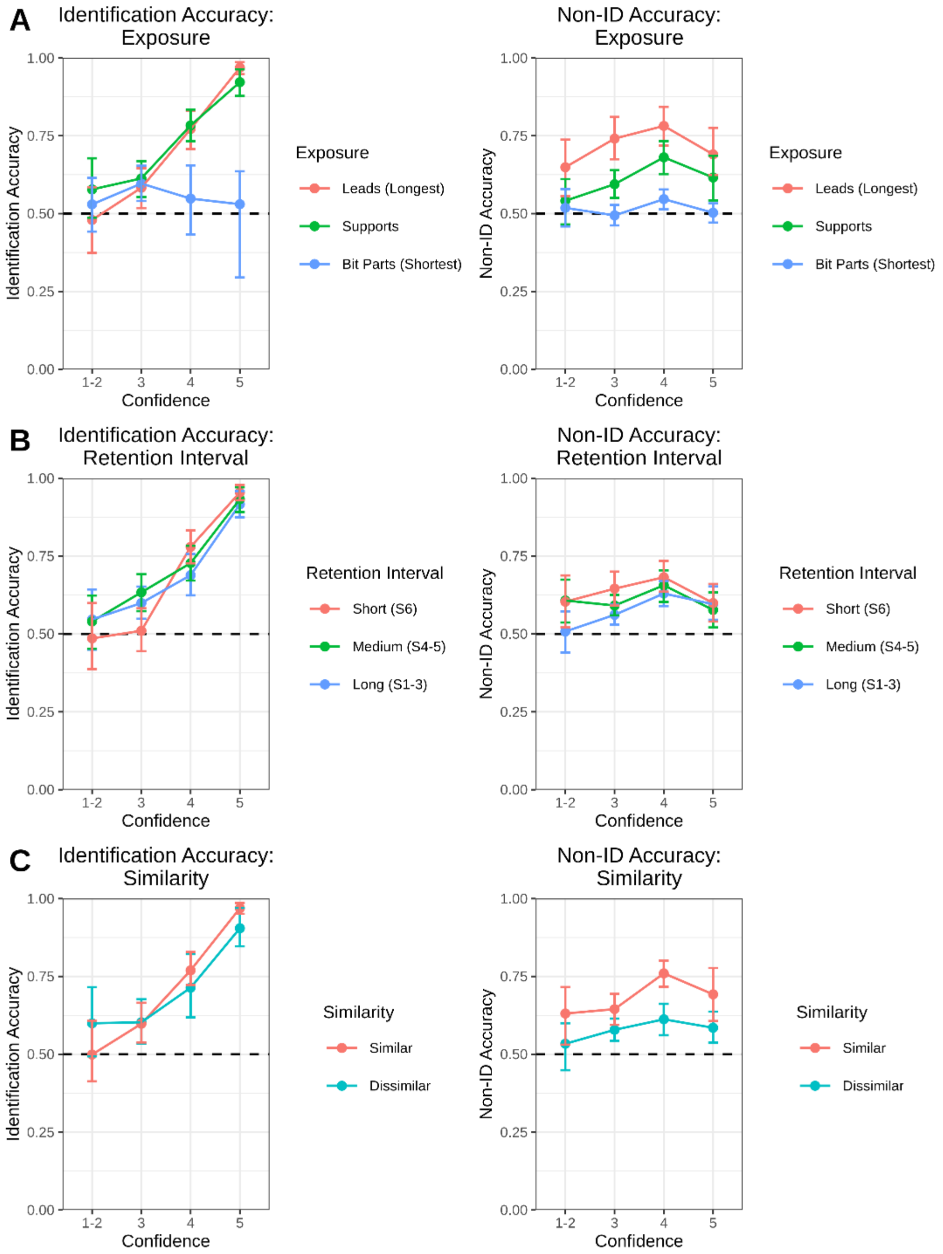


Figure 1B.2. Calibration curves for subsets of Exposure (panel a), Retention-interval (panel b), and similarity (panel c). The dashed line at 50% represents chance accuracy. Highest confidence identifications (left column) are potentially less reliable for shorter exposure durations ('bit parts'), but impressively accurate across all similarity and retention-interval conditions. In contrast, there is little relationship between confidence and accuracy for non-identifications (right column). Comparing the 95% HDIs, for identifications, we see only a significant difference between highest confidence low vs. high exposure.

Beginning with exposure-duration, we found a strong positive relationship between confidence and accuracy for ‘lead’ (longest exposures) and ‘supporting’ (medium exposures) actor identifications. Moreover, highest confidence identifications exceed 90% accuracy for both conditions (‘lead’ = 96.9%, 95% HDI [94.8, 98.6]; ‘support’ = 92.2%, 95% HDI [87.8, 96.3]). However, there is a much weaker confidence-accuracy relationship for ‘bit part’ (shortest exposure) identifications, as demonstrated by the flat line across confidence levels. In fact, the accuracy for highest confidence responses (53.0, 95% HDI [29.5, 63.6]) is nearly identical to that of the lowest confidence responses, ‘1-2’ (53.0, 95% HDI [37.4, 58.5]). For non-identifications, there are non-overlapping HDIs at nearly all levels of confidence between ‘leads’ and ‘bit parts’, suggesting that ‘new’ responses to longer exposures are generally more accurate overall. But, replicating previous research, confidence is a weak predictor of non-identification accuracy across all exposure-durations.

Next, examining Figure 1B.2b-c, the overlapping 95% HDIs suggest there is limited influence of retention-interval (panel b) or similarity (panel c) on accuracy. Highest confidence identifications are impressively accurate, even when considering the longest retention-intervals (91.6%, 95% HDI [87.5, 95.9]), and dissimilar trials (90.5%, 95% HDI [84.7, 97.0]). And, as with the exposure-duration analysis, there is little probative value of confidence for non-identifications. Importantly, these results do not imply that retention-interval and similarity have limited influence on accuracy at a macro-level. In fact, Devue et al. (2019) found reduced discriminability (i.e., lower d’) for longer retention-intervals and dissimilar photos at test. However, it does suggest that when participants use the same level of confidence, the impact of retention-interval and similarity are minimal.

In summary, the results of the calibration curve analyses generally support previous findings. High confidence is strongly indicative of accuracy for identifications (Wixted & Wells, 2017), though this comes with the caveat that short exposures may reduce reliability. Likewise, there are weaker relationships between confidence and accuracy for non-identifications (Brewer & Wells, 2006). Replicating these findings is encouraging, as the uncontrolled encoding context of watching GoT more closely matches our daily experiences than past experimental studies.

However, these results apply to the average individual. We now turn to the question of whether the confidence-accuracy relationship is influenced by individual-differences in face recognition ability.

Does face recognition ability influence the confidence-accuracy relationship in a real-world viewing context?

To begin, we divided participants by a median-split of CFMT+ score into weaker (scores = 52 – 73) and stronger (scores = 74 - 90) face recognition ability groups. Next, we plotted calibration curves for the totality of the data, separated by face recognition group. The positive slopes in Figure 1B.3 (left panel) suggest that confidence is strongly related to identification accuracy for both groups. But, the result that we highlight is at the highest confidence level. As demonstrated by the non-overlapping 95% HDIs, the stronger face recognition group (98.2%, 95% HDI [96.3, 100.0]) is about 9 percentage points more accurate than the weaker face recognition group (89.4%, 95% HDI [84.7, 94.3]). The same figure (right panel) indicates that there is little association between confidence and non-identification accuracy for both face recognition groups.

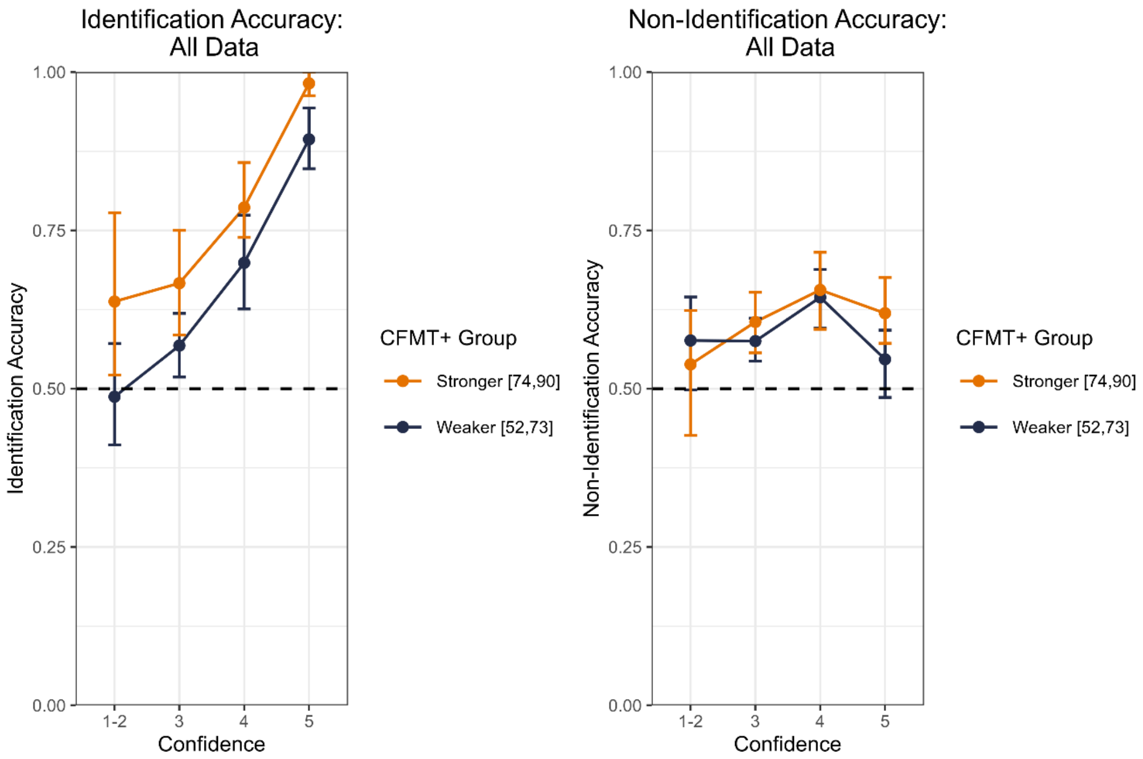


Figure 1B.3. Calibration curves for the full dataset, comparing median splits of CFMT+ score. Stronger face recognizers (i.e., CFMT+ = 74 - 90) are more accurate at the highest-level of confidence for identifications than weaker recognizers (CFMT+ = 52-73) (left panel). There are few differences between groups for non-identifications (right panel). Error bars represent 95% HDIs.

Figure 1B.4 shows the identification accuracy calibration curves for stronger and weaker recognizers across varying levels of exposure (panel a), retention-interval (panel b), and similarity (panel c). Starting with the effects of exposure duration, there is a noticeable gap between stronger and weaker recognizers for highest confidence ‘supporting’ actor identifications. Stronger recognizers commit about 10% fewer highest confidence errors than do weaker recognizers in this exposure condition (97.9%, 95% HDI [95.2, 100.0] vs. 87.3%, 95% HDI [82.1, 92.7], respectively). However, the patterns in the curves are similar across face recognition groups in the other levels of exposure. Though the HDIs do not overlap, highest confidence identifications of ‘lead’ actors are extremely accurate for both stronger and weaker recognizers (99.2%, 95% HDI [98.3, 100.0] vs. 94.6%, 95% HDI [91.4, 97.5], respectively). In contrast, the flat lines for ‘bit part’ identifications signify that the probative value of confidence is limited for shorter exposure durations, regardless of face recognition ability. As reflected by the missing point on the graph, the ‘bit part’ findings should be treated with caution. Table 1B.2 indicates that stronger recognizers never used the highest confidence rating for identifications in this exposure condition.

Next, we found that retention-interval exhibits a graded pattern of differences between stronger and weaker face recognizers. Specifically, the gap between the two groups at the highest confidence level appears to grow with longer retention intervals. Whereas stronger recognizers are about 6 percentage points more accurate than weaker recognizers for short delays (Season 6) (98.7%, 95% HDI [97.3, 100.0] vs. 92.7%, 95% HDI [89.4, 96.3], respectively), the difference increases to 11 points for long delays (Seasons 1-3) (97.7%, 95% HDI [95.4, 100.0] vs. 86.4%, 95% HDI [81.0, 92.1], respectively). As identification accuracy is essentially identical across retention-interval conditions for stronger recognizers, these differences are seemingly fueled by weaker recognizers committing increasing numbers of highest confidence errors over longer delays.

Concluding the identification accuracy results, both groups achieved high levels of accuracy for highest confidence identifications of *similar* faces, though stronger recognizers were about 5 percentage points more accurate than weak recognizers (99.5%, 95% HDI [98.6, 100.0] vs. 94.0%, 95% HDI [92.0, 96.1]). The gap between the two face recognition groups is numerically greater in the *dissimilar* condition, but the overlapping HDIs hint that these differences may not be particularly robust (96.3%, 95% HDI [92.2, 1.00] vs. 87.1%, 95% HDI [79.6, 95.7]).

Finally, Figure 1B.5 displays the non-identification calibration curves for stronger and weaker recognizers across the conditions of exposure duration (panel a), retention-interval (panel b), and similarity (panel c). The curves are essentially flat across all conditions of the manipulations for both groups, which implies that there is limited correspondence between confidence and non-identification accuracy, regardless of face recognition ability.

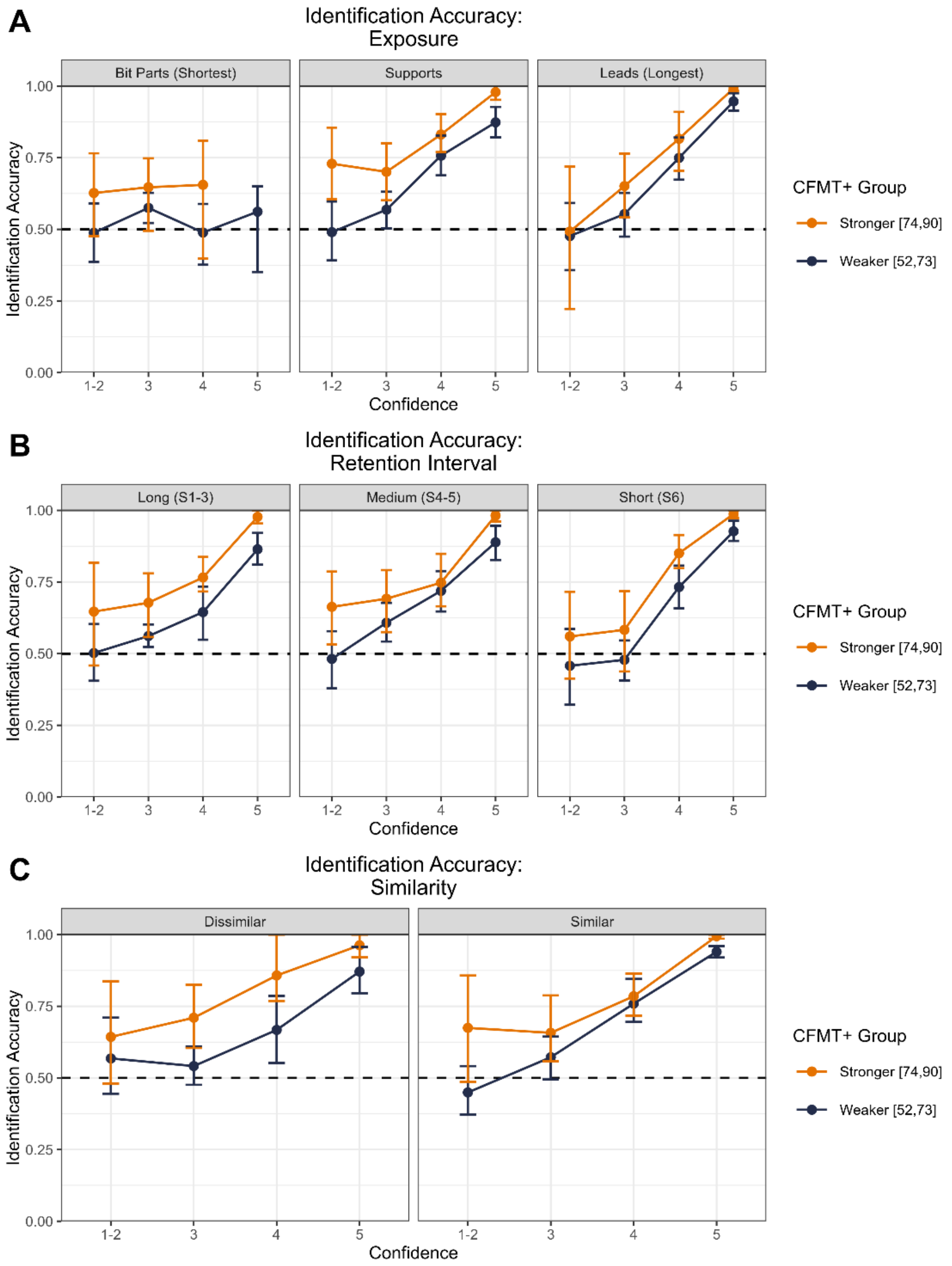


Figure 1B.4. Calibration curves for identification accuracy across manipulations, broken down by CFMT+ Median split. In almost all cases, there is a strong correspondence between confidence and accuracy. However, stronger face recognizers (CFMT+ scores = 74 – 90) commit fewer highest confidence errors than weaker recognizers (CFMT + scores = 52 - 73). The exception is for 'bit part' exposures (panel a, 1st column), where confidence is not predictive of accuracy for either face recognition group. Error bars represent 95% HDIs.

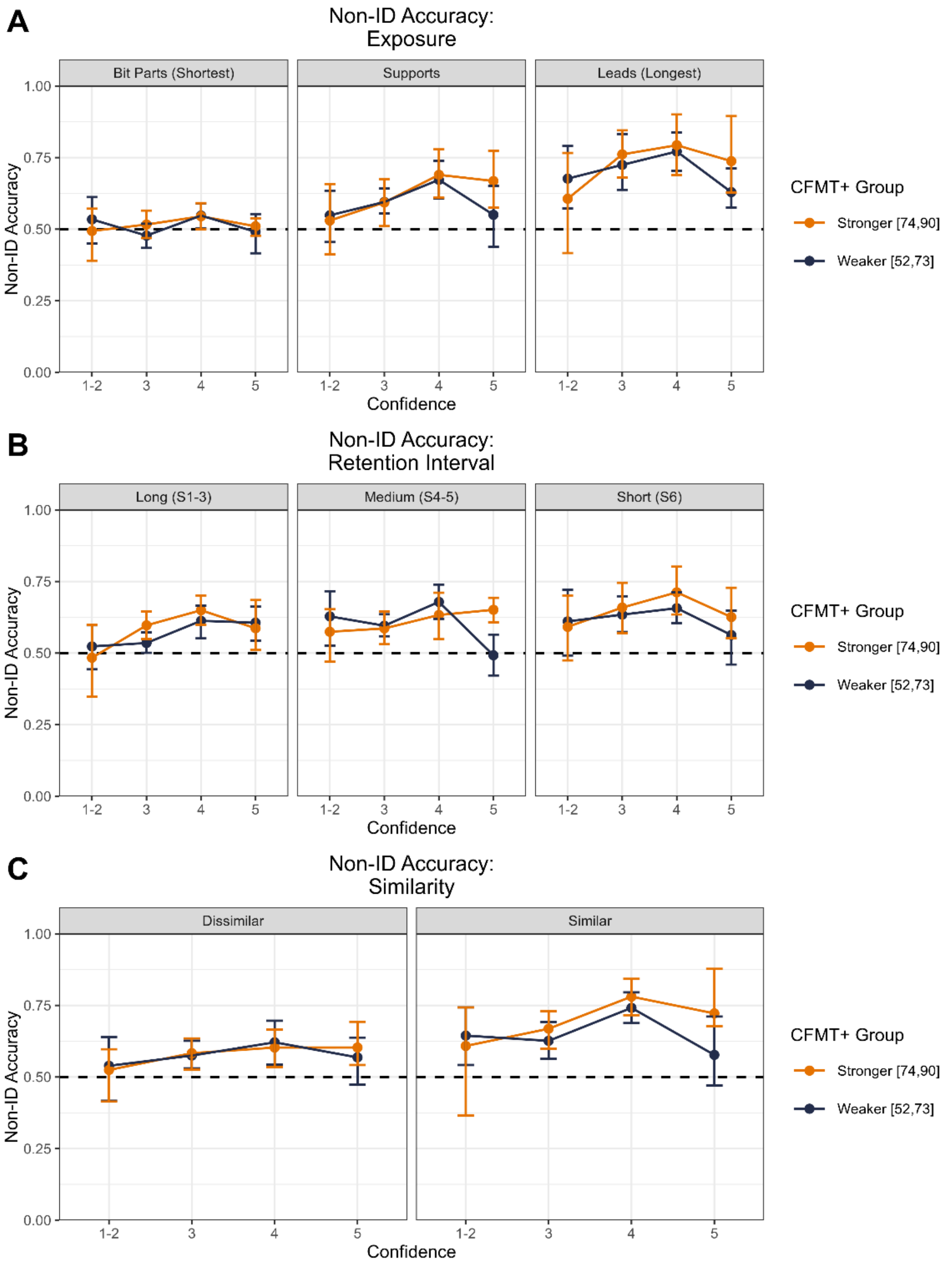


Figure 1B.5. Calibration curves for non-identification accuracy, broken down by CFMT+ Median split. There is little relationship between confidence and accuracy across all manipulations. Additionally, there are few differences between the stronger (CFMT+ scores = 74 – 90) and weaker (CFMT+ scores = 52 – 73) face recognition groups. Error bars represent 95% HDIs.

Does CFMT+ predict identification accuracy above and beyond confidence?

Following Grabman et al. (2019), we fitted logistic mixed effects models to identification responses, consisting of the main effects of Confidence and CFMT+, and their interaction. This method has the advantage of analyzing the effects of CFMT+ as a continuous variable, as opposed to collapsing participants into quantiles (e.g., median split). Moreover, including random intercepts for each participant partials out systematic variance resulting from unmeasured factors that are associated with accuracy (e.g., motivation, fatigue). Combined with visual inspection of the plot of the model output, these analyses clarify whether including CFMT+ significantly improves on predictions of identification accuracy generated from confidence alone.

Due to the single block design, we could not model interactions between the within-subjects conditions (i.e., exposure duration, retention interval), confidence, and CFMT+. Instead, we separately modeled subsets of total false alarms and hits from each condition (e.g., modeling 'lead' actor identifications separately from 'supporting' actor identifications). We gauged the relative contributions of CFMT+ and Confidence to the model by computing p-values from likelihood ratio tests (LRTs), provided by the *afex* package v. 0.24-1 (Singmann, Bolker, Westfall, & Aust, 2018).

There are no consensus standards for assessing the absolute model fit for logistic mixed effects models. Thus, we employed a combination of three methods to evaluate fit. 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 examined residual plots based on deviations between simulated and observed values to check for signs of model misspecification (i.e., ensuring errors are uniformly distributed for each predicted value). And third, we calculated marginal pseudo-R^2 ($R^2_{GLMM(m)}$) for fixed-effects, using the *MuMIn* package (Barton, 2018, version 1.42.1). This statistic compares variance accounted for by fixed effects in the model to remaining error, while partialing out the variance accounted for by the random effect structure (i.e., participant intercept).

We fitted mixed logistic regression models to identifications from the full dataset, as well as each individual subset (exposure, retention-interval) using the *lme4* package v. 1.1-21 (Bates, Maechler, Bolker, & Walker, 2014). Models consisted of the main effects and interaction between Confidence (as an ordered factor) and CFMT score (centered and scaled), as well as a random intercept for each participant. In Wilkinson-Rodgers notation (1973) all models took the form: Accuracy ~ Confidence + CFMT + Confidence:CFMT + (1|Participant). However, a slightly different model was needed for the analysis of the similarity manipulation because the between-subject nature of this manipulation required the inclusion of main effects and interactions of the similar vs. dissimilar contrast. None of these terms contributed significantly to the model (all ps > .144), so for brevity we discuss effects in the context of the terms common to all models.

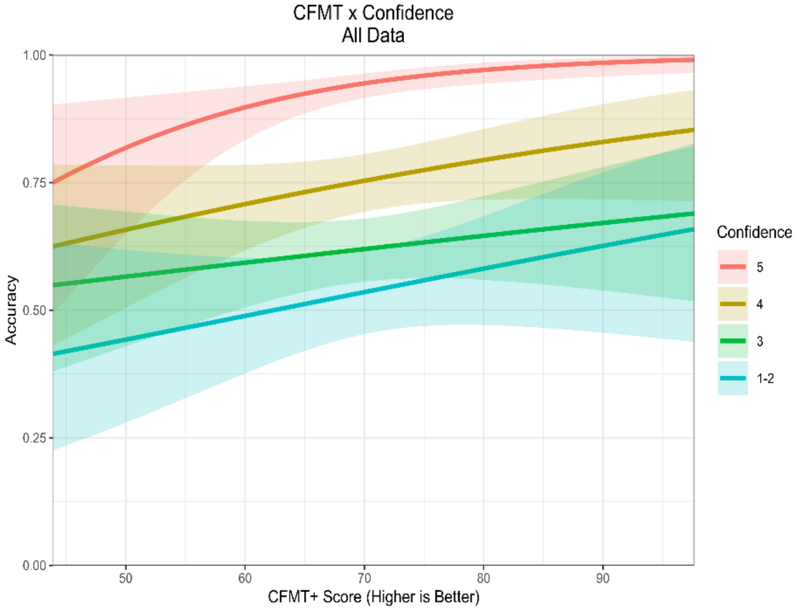


Figure 1B.6. Plot of the logistic mixed effects model for the full dataset, with predictors of CFMT+ score and confidence. Lines represent model estimates, with error shading representing the 95% confidence interval. Notably, high confidence errors (red line) are more frequent when participants are worse face recognizers.

Table 1B.5 shows the chi-square and significance values for each of the three terms (CFMT, Confidence, CFMT:Confidence) across models. In all cases, Kolmogorov-Smirnov tests suggested that the models adequately fit the data (all ps > .135), with residual plots showing no signs of major misspecification. Starting with the model of the full dataset, Figure 1B.6 suggests there is a strong main effect of confidence, such that higher confidence values are associated with greater predicted accuracy. Critically, replicating previous research, the model predicts that weaker face recognizers will make more frequent high confidence errors (as represented by the red line) than stronger recognizers.

		CFMT **(DF = 1)**		**Confidence** **(DF = 3)**		**CFMT:Confidence** **(DF = 3)**		**$R^2_{GLMM(m)}$**
Model		X^2	*Sig*	X^2	*Sig*	X^2	*Sig*	
All Data		6.91	.009**	197.64	< .001***	6.51	.089	.221
Delay	*Long*	6.43	.011*	96.80	< .001***	6.26	.100	.184
	Medium	4.66	.031*	133.48	< .001***	5.67	.128	.230
	Short	7.33	.007**	195.55	< .001***	6.12	.106	.323
Exposure	*Bit Parts*	0.24	.627	2.86	.414	3.27	.351	.013
	Supports	7.06	.008**	99.85	< .001***	5.50	.139	.189
	Leads	2.99	.084	326.96	< .001***	7.61	.055	.408
Similarity		7.68	.006**	64.26	< .001***	4.46	.216	.279

* < .05; ** < .01; *** < .001

Table 1B.5. Output for logistic mixed effects models examining the main effects of CFMT+, Confidence, and their interaction. Models are fitted to the full dataset, as well as subsetted conditions of exposure, retention-interval, and similarity. Significance values are computed by Likelihood ratio tests. CFMT+ contributes to all, except in the cases of the shortest (‘bit parts’) and longest (‘leads’) exposures.

Figure 1B.7 reveals that this interpretation persists across most of the manipulation analyses. Examining Table 1B.5, CFMT+ makes a significant contribution to nearly all of the models. In general, the models predict that weaker recognizers will commit more high confidence errors than stronger recognizers. Two notable exceptions come from the analysis of exposure duration. Consistent with the calibration curves, both confidence and face recognition ability are not particularly predictive of accuracy for ‘bit part’ identifications. Additionally, while confidence is strongly predictive of accuracy for ‘lead actors’, face recognition ability does not contribute significantly to this model.

In summary, the logistic mixed-effects modeling analyses largely reach the same conclusions as the calibration plots: face recognition ability influences the rate of high confidence errors across the dataset as a whole, and across the manipulations of retention-interval and similarity. Concurrently, there are minimal differences in the predictive value of confidence between stronger and weaker recognizers at the longest exposure durations (where confidence is similarly predictive across CFMT+), and the shortest exposure durations (where confidence is not predictive of accuracy across CFMT+).

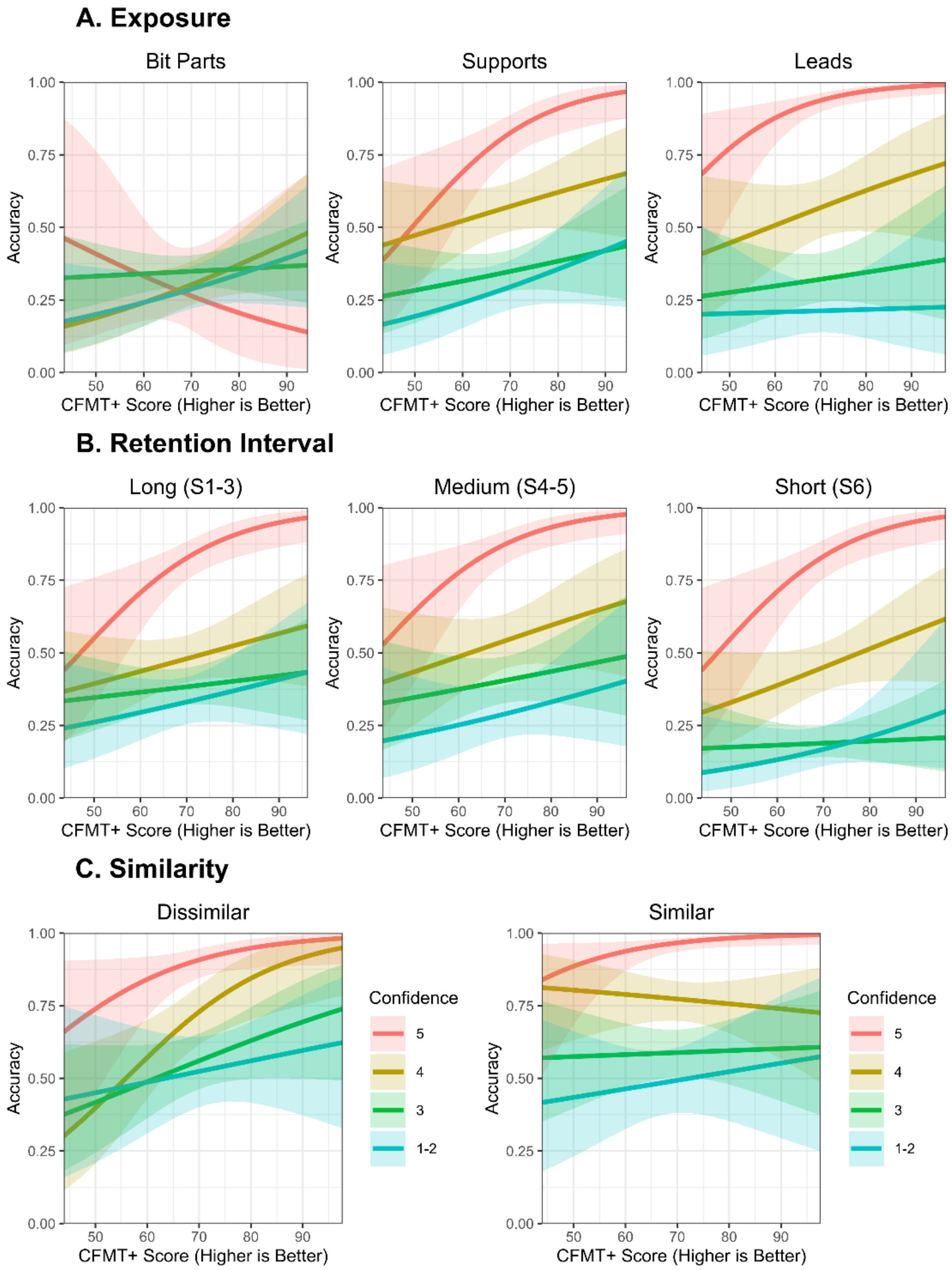


Figure 1B.7. Plots of the logistic mixed effects model Exposure (panel a), Delay (panel b), and Similarity (panel b) subsets, with predictors of CFMT+ score and confidence. Lines represent model estimates, with error shading representing the 95% confidence interval. Notably, high confidence errors (red line) are more pronounced when participants are worse face recognizers.

Study 1B Discussion

The ability to evaluate whether we have encountered a person before carries real-world implications. The consequences of misplaced confidence range from the relatively mild (e.g., mistaking one actor for another on a television show) to the devastatingly severe (e.g., convicting an innocent person in a police lineup), yet knowledge about the probative value of confidence is rooted in experiments that inadequately capture the complexity of real-world face recognition. And, excepting results from one previous experiment (Grabman et al., 2019), little is known about how individual differences impact the confidence-accuracy relationship.

Participants in Devue et al (2019) grappled with many of the challenges posed by everyday face recognition. They watched the television show *Game of Thrones* for personal entertainment, incidentally encoding the faces of hundreds of actors. Retention-intervals were much longer than in previous studies (up to 6 years) and characters appeared in varied contexts, often with substantial changes in appearance.

Despite the substantial difficulties imposed by this real-world, uncontrolled viewing context, our reanalysis of the data finds that confidence ratings are generally predictive of identification accuracy (e.g., the strong positive slopes for identification in Figure 1B.1). Highest confidence identifications were remarkably accurate for retention-intervals of >3 years (91.6%), and for photos that were as dissimilar as possible from the actors' last appearance in GoT (90.5%).

In contrast, we found weak correspondence between confidence and accuracy for 'bit part' identifications. Sauer, Palmer, and Brewer (2019) recently posited there may be boundary conditions where the confidence-accuracy relationship does not hold -- especially in real-world settings. The aspects of the 'bit part' trials that are dampening this relationship will need further clarification. As Devue et al (2019) note, beyond screen-time there are other differences embedded in the exposure-duration manipulation (e.g., semantic information about the characters, access to varying viewpoints).

Taken together, these findings add compelling evidence to claims of a robust association between confidence and accuracy for identifications in face recognition tasks (Tekin & Roediger, 2017; Wixted & Wells, 2017). However, replicating Grabman et al. (2019), high confidence identifications by stronger face recognizers (upper median of CFMT+) were more reliable (98.2% overall) than those made by weaker recognizers (89.4%). The gaps between these two groups were largest when identifying photos at delays >3 years (97.7% vs. 86.4%), for medium exposures (97.9% vs. 87.3%), and (potentially) for dissimilar appearance (96.3% vs. 87.1%). It is surprising that the effects of face recognition ability on the confidence-accuracy relationship are so clear, given that the sample did not include individuals at the extreme tails of CFMT+ performance (i.e., scores closer to chance or perfect). We expect that differences would increase when contrasting wider ranges of ability.

Importantly, face recognition ability is likely one among many individual-differences that influence the confidence-accuracy relationship for face recognition tasks. Kantner and Dobbins (2019) estimated the relative contributions of accuracy and individual-differences (broadly defined) to the confidence ratings participants assigned to recognition trials of paintings and words. The results showed both factors accounting for near-equal variance in confidence for identifications, while individual-differences explained 13-20x more variance in confidence for non-identifications. Martschuk, Sporer, & Sauerland, (2019) found that older eyewitnesses tended to make high confidence errors more frequently than younger participants. Additionally, a 'confidence trait' may influence how people use confidence scales across many unrelated tasks (e.g., Stankov, Kleitman, & Jackson, 2015). The Wixted & Wells (2017) review convincingly argues for determining the probative value of individual levels of confidence. We believe the next step is to determine the probative value of confidence for *responses by individuals*. Ideally these studies will use a combination of experimental and real-world designs, with an eye for capturing the complexities of face recognition.

In summary, we find that confidence ratings are generally predictive of accuracy in a difficult, uncontrolled real-world test of face recognition. Importantly, we replicate that face recognition ability influences the reliability of high confidence ratings, despite using other researchers' data, a different memory paradigm (old-new vs. lineup), and a more ecologically valid context.

Study 1B References

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36–40. https://doi.org/10.1016/j.paid.2013.12.011

Barton, J. J. S., & Corrow, S. L. (2016). The problem of being bad at faces. *Neuropsychologia*, *89*, 119–124. https://doi.org/10.1016/j.neuropsychologia.2016.06.008

Barton, K. (2018). *MuMIn: Multi-model inference*. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-21. https://CRAN.R-project.org/package=lme4

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277–291. https://doi.org/10.1037/a0029635

Brewer, N., & Burke, A. (2002). Effects of Testimonial Inconsistencies and Eyewitness Confidence on Mock-Juror Judgments. *Law and Human Behavior*, *26*(3), 353–364. https://doi.org/10.1023/A:1015380522722

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11–30. https://doi.org/10.1037/1076-898X.12.1.11

Devue, C., Wride, A., & Grimshaw, G. M. (2019). New insights on real-world human face recognition. *Journal of Experimental Psychology: General*, Vol. 148, pp. 994–1007. https://doi.org/10.1037/xge0000493

Duchaine, B. C., & 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*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

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

Gosling, A., & Eimer, M. (2011). An event-related brain potential study of explicit face recognition. *Neuropsychologia*, *49*(9), 2736–2745. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2011.05.025

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*(2), 233–243. https://doi.org/10.1016/j.jarmac.2019.02.002

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

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know? *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2018.1319

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*, Vol. 22, pp. 1304–1316. https://doi.org/10.1037/0278-7393.22.5.1304

Kantner, J., & Dobbins, I. G. (2019). Partitioning the sources of recognition confidence: The role of individual differences. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-019-01586-w

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*. https://doi.org/10.1016/j.cognition.2017.12.005

Mansour, J. K., Beaudry, J. L., & Lindsay, R. C. L. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials. *Behavior Research Methods*, *49*(6), 2235–2254. https://doi.org/10.3758/s13428-017-0855-0

Martschuk, N., Sporer, S. L., & Sauerland, M. (2019). Confidence of Older Eyewitnesses: Is It Diagnostic of Identification Accuracy? *Open Psychology*, *1*(1), 132–151. https://doi.org/10.1515/psych-2018-0010

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*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

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

Morgan, C. A., Hazlett, G., Baranoski, M., Doran, A., Southwick, S., & Loftus, E. (2007). Accuracy of Eyewitness Identification is significantly associated with performance on a standardized test of face recognition. *International Journal of Law and Psychiatry*, *30*(3), 213–223. https://doi.org/https://doi.org/10.1016/j.ijlp.2007.03.005

Nguyen, T. B., & Pezdek, K. (2017). Memory for disguised same- and cross-race faces: The eyes have it. *Visual Cognition*, *25*(7–8), 762–769. https://doi.org/10.1080/13506285.2017.1329762

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55–71. https://doi.org/10.1037/a0031602

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, *0*(0). https://doi.org/10.1111/bjop.12368

Read, J. D., Lindsay, D. S., & Nicholls, T. (1998). The relation between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? *Eyewitness Memory: Theoretical and Applied Perspectives.*, pp. 107–130. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Righi, G., Peissig, J. J., & Tarr, M. J. (2012). Recognizing disguised faces. *Visual Cognition*, *20*(2), 143–169. https://doi.org/10.1080/13506285.2012.654624

Russell, R., Duchaine, B. C., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin and Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in Using Eyewitness Confidence to Diagnose the Accuracy of an Individual Identification Decision. *Psychology, Public Policy, and Law*, *25*(3), 147–165. https://doi.org/10.1037/law0000203

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 12887–12892. https://doi.org/10.1073/pnas.1421881112

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *afex: Analysis of factorial experiments*. R package version 0.24-1. https://CRAN.R-project.org/package=afex

Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the Trait of Confidence. In *Measures of Personality and Social Psychological Constructs*. https://doi.org/10.1016/B978-0-12-386915-9.00007-3

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*(1), 1–13. https://doi.org/10.1186/s41235-017-0086-z

Terry, R. L. (1994). Effects of facial transformations on accuracy of recognition. *The Journal of Social Psychology*, *134*(4), 483–492.

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*(3), 530–548. https://doi.org/10.1037/a0019972

Wilmer, J. B., Germine, L. T., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., … Duchaine, B. C. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*(11), 5238–5241. https://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*(2), 192–203. https://doi.org/https://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*(1), 10–65. https://doi.org/10.1177/1529100616686966

Young, A. W., & Burton, A. M. (2017). Recognizing Faces. *Current Directions in Psychological Science*, *26*(3), 212–217. https://doi.org/10.1177/0963721416688114

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*(2), 137–142. https://doi.org/https://doi.org/10.1016/j.cub.2009.11.067

Part II: Interpreting Eyewitness Confidence Statements

Prior knowledge influences interpretations of eyewitness confidence statements: "The witness picked the suspect, they must be 100% sure." *(Grabman & Dodson, 2019*)

Imagine you are a police officer. After a few months investigating a string of recent robberies, you finally have a suspect. Following the recommended 'double-blind' procedure, another officer unaware of the identity of the suspect interviews an eyewitness to one of the crimes. The colleague administers a lineup consisting of the suspect and five additional individuals known to have not committed the crime. The eyewitness points to the suspect, and says, "I'm very sure it's him". How certain is the eyewitness? Would your answer change if you did not know which lineup member was the suspect?

These are increasingly important questions in the criminal justice system. Unlike most laboratory studies, where confidence is collected on a numeric scale (e.g. 0 – 100% certain), police are advised by the United States Department of Justice to "ask the witness to state, in his or her own words, how confident he or she is in the identification" (Yates, 2017). Evidence suggests that people generally report confidence verbally, rather than numerically (Brun & Teigen, 1988; Budescu & Karelitz, 2003; Erev & Cohen, 1990; Wallsten, Budescu, Zwick, & Kemp, 1993). For example, 20 out of 23 mock witnesses in a lineup identification task provided verbal (rather than numeric) expressions of certainty when given instructions closely mirroring Department of Justice guidelines (i.e. "in your own words, please explain how certain you are in your response") (Dodson & Dobolyi, 2015). Using identical prompts, the same authors recently found that 275 out of 342 participants expressed confidence verbally for every response to twelve mock lineups (Dobolyi & Dodson, 2018).

The critical question is whether law enforcement professionals accurately interpret the intended meaning of verbal expressions of confidence. Unfortunately, much research outside the eyewitness domain indicates that the answer is no (e.g., Budescu, Por, Broomell, & Smithson, 2014; Budescu, Broomell, & Por, 2009; Budescu & Wallsten, 1985; Gurmankin, Baron, & Armstrong, 2004; Mullet & Rivet, 1991; Reagan, Mosteller, & Youtz, 1989). Further, expertise does not eliminate interpersonal variability in the interpretation of verbal expressions of certainty, even when verbal statements are drawn from a content area where raters have considerable familiarity (Beyth-Marom, 1982; Bryant & Norman, 1980; Nakao & Axelrod, 1983; Wallsten, Fillenbaum, & Cox, 1986). In one illustration of this difficulty, Nakao & Axelrod (1983) found that physicians showed just as much variability as non-physicians in their assessment of the intended probability value of 17 out of 22 verbal modifiers (e.g.," Invariably") that are commonly used to express frequencies of medical events.

Few studies have examined whether eyewitness expressions of confidence are similarly likely to be misinterpreted (Cash & Lane, 2017; Dodson & Dobolyi, 2015, 2017). There is urgent need for examining this issue, given that an eyewitness's confidence guides many criminal justice decisions. For example, research indicates that the level of an eyewitness's confidence is likely the most important influence on jury decision-making (e.g., Bradfield & Wells, 2000). Moreover, in a synthesis of over 30 years of research, Wixted and Wells (2017) note a strong positive relationship between eyewitness accuracy and confidence at the time of an initial identification, assuming that investigators follow "pristine eyewitness identification procedures" (e.g., one suspect per lineup; see Wixted & Wells, 2017 for a full review). However, one general assumption underlying this relationship is that the witness's confidence is properly understood by evaluators, which may not always be the case.

One major finding is that confidence interpretations become less consistent as contextual information is introduced (e.g., Brun & Teigen, 1988). This likely occurs because interpreters have varied conceptions of the utility and prior likelihood of the events the confidence phrases are modifying (Beyth-Marom, 1982). Recent work by Dodson and Dobolyi (2015) suggests that perceptions of certainty are influenced by how witnesses justify their selections from lineup. In this study, participants viewed twelve mock eyewitnesses' identifications of one of the six members of the lineup, and the eyewitness's statement of confidence about his/her identification (e.g., "I am very certain. I remember his hair"). The critical manipulation was that some participants saw the confidence statement only ("I am very certain") whereas other participants saw the confidence statement with either a featural justification – one that referred to a visible feature about the suspect (e.g., "I am very certain. I remember his hair.") or an unobservable justification – one that referred to a quality about the suspect that is unobservable to anyone but the eyewitness (e.g., "I am very certain. He looks like a friend of mine."). When asked to translate the verbal confidence statements into a numeric value of certainty, participants rated confidence only statements similarly to those where the eyewitness provided an unobservable justification. However, featural statements were interpreted as meaning significantly lower values than both other conditions, especially when the eyewitness used language associated with high certainty (see Cash & Lane, 2017 for a replication and extension of this effect).

An open question is whether evaluators' judgements of confidence are also impacted by contextual factors unstated by (or even unknown to) the witness, such as whether the suspect confessed to the crime. These contextual influences are (to our knowledge) unexplored in evaluations of confidence statements, but do appear in other criminal justice research. For instance, Kassin, Dror and Kukucka (2013) reviewed a variety of judgements in different forensic procedures that are affected by contextual knowledge. Polygraph examiners are more likely to interpret an interviewee's polygraph chart as deceptive when they are told that the interviewee later confessed to the crime than when told that someone else confessed (e.g., Elaad & Ginton., 1994). Expertise likely does not mitigate the influence of external knowledge as multiple studies show the effects in both experienced polygraph examiners (Elaad and Ginton, 1994), and fingerprint examiners (e.g., Dror & Charlton, 2006; see also Dror, Charlton & Péron, 2006). One final key point that Kassin et al (2013) emphasize is that contextual knowledge is more likely to influence judgments when the evidence is ambiguous rather than clear-cut.

The "post-identification feedback" literature suggests that investigators' contextual knowledge can directly alter the witness's actual level of confidence (see Steblay, Wells, & Bradfield Douglass, 2014 for a recent meta-analysis). By their nature, lineups contain an individual that the police have identified as their suspect. Wells and Bradfield (1998) demonstrated that investigators can inflate eyewitnesses' confidence reports by validating their selection from a lineup (e.g. "Good, you identified the suspect"), or (to a lesser extent) deflate certainty by invalidating the selection (e.g. "Actually the suspect was [somebody else]"). Of particular note, after receiving confirmatory feedback, roughly half of the witnesses who chose an incorrect face from the lineup reported high levels of confidence (6 or 7 on a 7-point scale) (Wells & Bradfield, 1998). This could lead observers to erroneously place high credibility on these witnesses' testimony (Steblay et al., 2014; Wells & Bradfield, 1998). Concerns about investigator influence on witness's lineup decisions and confidence reports lead these authors (and others) to emphasize double-blind lineup administration procedures (e.g., Dror, Kukucka, Kassin, & Zapf, 2017; Kovera & Evelo, 2017; Steblay et al., 2014),

In addition to influencing eyewitness reports, prior knowledge of the police suspect could also affect evaluators' perceptions of eyewitness certainty. In fact, in explaining the post-identification feedback effect, Wells and Bradfield (1999) speculate that witnesses adjust confidence reports based on how a hypothetical outside observer would perceive the statement. A confidence expression, such as “I’m pretty sure it’s him,” may be interpreted as conveying a higher level of confidence when the statement refers to an individual who corresponds to the police’s suspect than when it does not. This is important to document, because ultimately evaluators of these statements must make decisions about how to proceed in legal settings (e.g., police officers, judges, jurors).

The purpose of this study is to investigate whether prior knowledge of the police suspect impacts perceptions of eyewitness confidence. The results bear on current recommendations of best-practices for eliciting eyewitness confidence statements. If contextual information leads evaluators to misinterpret witnesses’ intended level of certainty, current standards for “pristine lineup conditions” (including double-blind administration) (Wixted & Wells, 2017) do little to mitigate misinterpretations, because most legal decisions (e.g., arrests, evidence review, indictment) are made with knowledge of the police suspect. These mistakes could lead evaluators to put undue emphasis on the testimony of witnesses who identify the police suspect, potentially increasing false convictions. Further, Wells, Yang, and Smalarz (2015) note that picking someone other than the suspect (i.e. “fillers”) with high confidence should be considered strong exculpatory evidence, meaning that downplaying the importance of disconfirmations could decrease exoneration rates.

Finally, we investigate whether knowledge of the police suspect moderates other contextual effects, such as the way eyewitnesses justify their selections from a lineup. Dobolyi & Dodson (2018) recently found that participants who picked a face from a lineup provided a featural justification for nearly half of their verbal confidence statements. Given the ubiquity of these expressions, manipulating an eyewitness’s justification for a level of confidence may provide a more complete picture of the factors that influence interpretations of eyewitness confidence statements.

Experiment A

Experiment A manipulates prior knowledge of the suspect within-subjects, and justification condition between-subjects. In some conditions, participants will know which lineup member is the police's suspect whereas in other conditions they will not have such knowledge. Given previous research, we expect that identical eyewitness statements of confidence will be evaluated as meaning higher or lower levels of confidence when the eyewitness has chosen someone from a lineup that either matches or mismatches, respectively, the police's suspect. Further, we hypothesize that this knowledge will moderate effects of eyewitness justification. Specifically, we predict that (a) *identifying the suspect will diminish the featural justification effect*, as this evidence supports that the feature the witness chose is diagnostic, and (b) *filler identifications will enhance (or have minimal influence on)* the featural justification effect, because this bolsters perceptions of low featural discriminability.

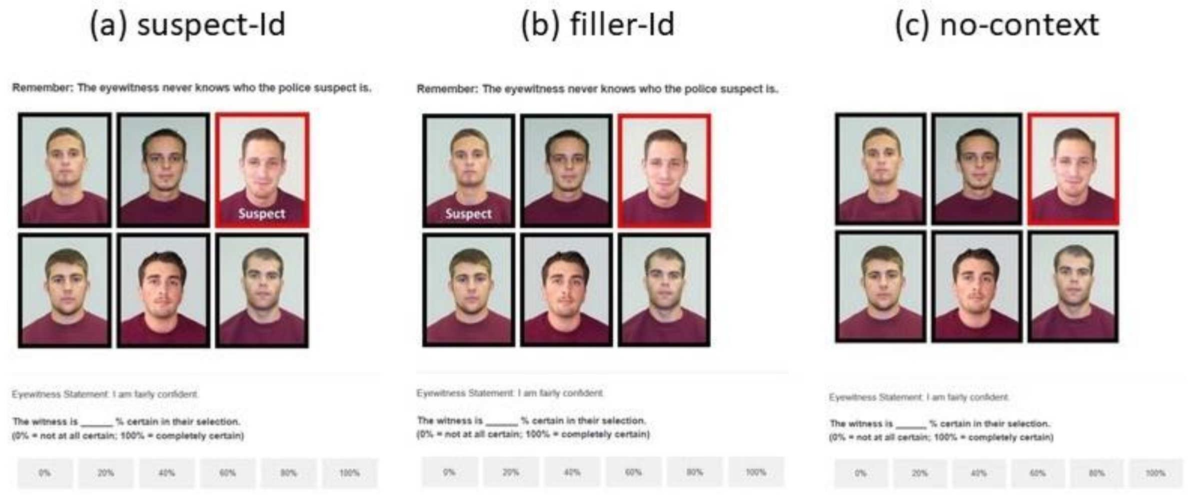
Method

Participants

We analyzed the data of 181 participants, located in the United States, who completed the experiment over the internet using Amazon's Mechanical Turk (mean age = 37.07, SD = 10.92, range = 20 - 74, 58.01% female, 84.53% White/Caucasian), and were randomly assigned to one of two justification conditions: (a) Confidence only (n= 93) and (b) Featural Justification (n= 88). The sample size was sufficient to detect moderate-sized effects with greater than 95% power at an alpha level of .05, according to G*POWER (Faul, Erdfelder, Lang, & Buchner, 2007).

Materials

All participants interpreted the verbal confidence of mock eyewitnesses responding to six “fair” lineups (see Dobolyi & Dodson, 2013 for methodology of lineup creation). Fair lineups meet the conditions that no face is comparatively noticeable, and that individuals naïve to the target stimulus are equally likely to select each face (e.g. Gronlund, Carlson, Dailey, & Goodsell, 2009; Malpass, Tredoux, & McQuiston-Surrett, 2007). As shown in Figure 2.1, lineups consisted of the head and shoulders of six white males arranged in a 2x3 grid, exhibiting a neutral expression, and wearing a solid maroon colored t-shirt. In each case, one face was randomly selected to serve as the eyewitness’s identified suspect and outlined with a red border. In the within-subjects prior knowledge manipulation, the word “suspect” in white text represented the police suspect and was superimposed over either the eyewitness’s selection (suspect-ID), another randomly selected face (filler-ID), or omitted (i.e. no police suspect, no-context).



(a) suspect-Id

Remember: The eyewitness never knows who the police suspect is.

Eyewitness Statement: I am fairly confident.

The witness is ______ % certain in their selection.
(0% = not at all certain; 100% = completely certain)

(b) filler-Id

Remember: The eyewitness never knows who the police suspect is.

Eyewitness Statement: I am fairly confident.

The witness is ______ % certain in their selection.
(0% = not at all certain; 100% = completely certain)

(c) no-context

Eyewitness Statement: I am fairly confident.

The witness is ______ % certain in their selection.
(0% = not at all certain; 100% = completely certain)

Figure 2.1. In Experiment A, participants evaluated the eyewitness’s intended numeric confidence expression under 2 justification conditions and when: (a) the eyewitness chose the same person as the police (suspect-Id), (b) the eyewitness chose someone other than the police suspect (filler-Id), and (c) when given no information about the suspect (no-context).

Accompanying each lineup was the eyewitness's written verbal expression of confidence in their selection. Briefly, expressions of confidence were transcribed verbatim from participant responses to a previous eyewitness study, and reflected either "moderate" (e.g. "I am fairly confident") or "high" (e.g. "I am positive") levels of certainty (see Dodson and Dobolyi, 2015 for a full summary on expression generation). Additionally, statements in the "featural" condition included justifications reflecting visible features of the suspect (e.g. "I remember his nose") (see Table 2.1 for confidence statements and justification conditions).

Confidence Level	**Context Condition**	**Expression -- Verbal**	**Expression – Numeric (Experiment C Only)**	**Featural Justification**
Moderate	suspect-ID	I am pretty certain.	I am 60% certain.	I remember his chin.
	no-context	I am fairly confident.	I am 60% confident.	I recall his hair.
	filler-ID	I am pretty sure.	I am 60% sure.	I remember his ears.
High	suspect-ID	I am very certain.	I am 100% certain.	His mouth is memorable.
	no-context	I am very confident.	I am 100% confident.	I remember the shape of his head.
	filler-ID	I am positive.	I am 100% positive.	I recall his eyes.

Table 2.1. Confidence expressions used in Experiments A – C. Depending on condition, participants in Experiment C saw only Verbal expressions, or Numeric expressions. In all experiments, half of participants saw confidence statements with the addition of a featural justification.

Procedure

Figure 2.1 (panels a – c) shows examples of the task. At study onset, participants were randomly assigned to a single justification condition and viewed the same sequence of six lineups. We counterbalanced the order of all possible pairings of the prior knowledge (suspect-ID, filler-ID, no-context) and statement confidence (moderate, high) conditions across participants.

Given the goal of approximating how police officers interpret eyewitness statements, we started by broadly informing participants about the process police use to generate criminal lineups. We told them that sometimes, when police create a lineup of faces, they include one person that they think could be the criminal. The remaining faces in the lineup are innocent individuals that the police know did not commit the crime. We emphasized that the eyewitness is never told whether his/her choice matches the police's suspect.

For suspect-known lineups, participants pretended to be the police officer who created the lineup. For no-context lineups, participants pretended to be a police officer who did not have a particular suspect in mind. In all cases, the participant's task was to translate the written statements into a numerical value of certainty, by filling in the sentence, "the witness is ___% certain in their selection," using a scale ranging from 0% (not at all certain) – 100% (completely certain) in 20% increments.

To ensure participants understood all the instructions, and to preview the task, we first showed examples of suspect-ID and no-context lineups composed of six colorful smiley faces. In each case the eyewitness statement read, "I know it's him". Participants were instructed to pretend that the example eyewitness was completely certain about their decision. Finally, upon completing the task, we checked for comprehension by asking participants, "when the eyewitness looked at the lineup, did he/she know who the police suspect was?" and "other than the police suspect, is it possible that anyone else in the lineup committed the crime?" We did not analyze the data of 175 individuals who responded < 80% on either one of the smiley lineups (N = 127), and/or answered "yes" to either of the final manipulation checks (N = 93). Data can be found on the OSF (https://osf.io/8chmz/).

Experiment A Results and Discussion

We analyzed participants' estimates of the intended meaning of an eyewitness's verbal expression of confidence with a 2 (Justification: confidence only, featural; between) x 3 (Prior Knowledge: suspect-ID, filler-ID, no-context; within) x 2 (Confidence Level: high vs. moderate; within) mixed ANOVA. Replicating past results, a main effect of Justification condition, $F(1, 179) = 9.76$, $MS_e = 985.18$, $p = .002$, $\eta^2_p = .052$, reflects participants perceiving eyewitnesses as less confident when the eyewitness's confidence-statement was accompanied by a featural justification ($M = 64.58$, $SD = 14.62$) than when the eyewitness provided a confidence statement only ($M = 70.53$, $SD = 10.84$). As hypothesized, a main effect of Prior Knowledge, $F(2, 358) = 45.24$, $MS_e = 364.43$, $p < .001$, $\eta^2_p = .202$, is fueled by higher perceived confidence for statements about an identification that matched the police's suspect ($M = 75.14$, SD = 14.93) than in the no-context condition (M = 65.75, SD = 15.53; t (180) = 7.14, $p < .001$, Cohen's d = .53, 95% CI [.37, .68]), which in turn showed higher perceived confidence than the filler-ID condition ($M = 62.04$, $SD = 20.38$; $t(180) = 2.85$, $p = .005$, Cohen's $d = .21$, 95% CI [.06, .36]). And, as expected, there was a main effect of Confidence, F (1, 179) = 489.59, $MS_e = 254.61$, $p < .001$, $\eta^2_p = .732$, with increased perceived confidence for statements expressing high confidence ($M = 78.38$, $SD = 15.24$) than moderate confidence ($M = 56.91$, $SD = 14.04$).

These main effects should be interpreted in the context of a significant two-way interaction between Prior Knowledge and Confidence Level, *F* (2, 358) = 5.29, MS$_e$ = 203.38, *p* = .005, η^2_p = .029. As Figure 2.2 shows, the same moderate confidence statement is interpreted as meaning a higher numeric value when the statement refers to a suspect-ID lineup (*M* = 65.52, *SD* = 17.78) than a no-context lineup (*M* = 55.91, *SD* = 18.46), which in turn is rated higher than the filler-ID lineup (*M* = 49.28, *SD* = 20.71), *t*(180) = 5.61, *p* < .001, Cohen's *d* = .42, 95% CI [.26, .57] and *t*(180) = 4.17, *p* < .001, Cohen's *d* = .31, 95% CI [.16, .46]. For high confidence statements, we also found that participants perceived greater confidence for suspect-ID lineups (*M* = 84.75, *SD* = 17.18) than no-context lineups (*M* = 75.58, *SD* = 17.18), *t*(180) = 6.33, p < .001, Cohen's *d* = .47, 95% CI [.31, .62]. However, in contrast to the results for moderate statements, ratings for filler-ID lineups (*M* = 74.81, *SD* = 27.28) did not significantly differ from no-context lineups, *t*(180) = .41, *p* = .684, Cohen's *d* = .03, 95% CI [-.12, .18].

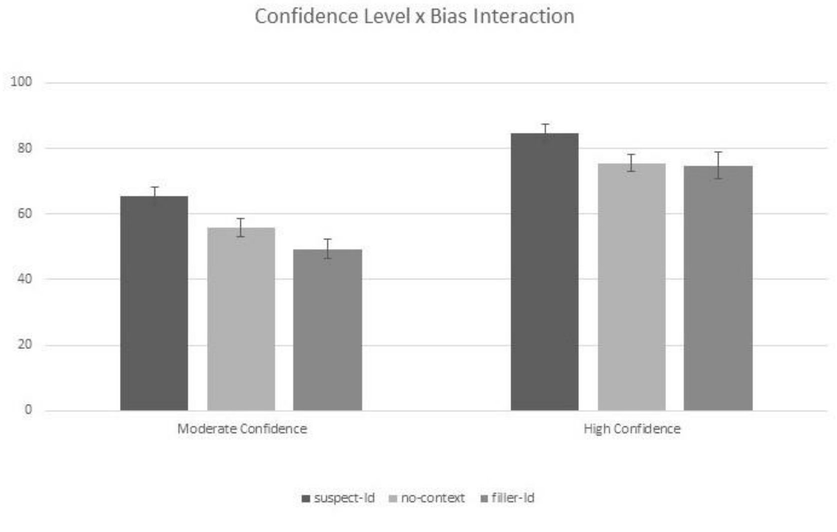


Figure 2.2. In Experiment A, we found a significant two-way interaction between the Confidence and Context conditions. As seen on the graph, the effect of the Context manipulation is stronger for moderate confidence than high confidence statements. Error bars indicate 95% confidence intervals of the mean.

Unexpectedly, we did not observe any further interactions (all *ps* $\geq .070$, $\eta^2_p \leq .015$). Contrary to our hypotheses, the Prior Knowledge manipulation did not significantly impact the effects of featural justification on perceived confidence.

Taken together, the results of Experiment A support the conclusion that knowledge of the police suspect influences perceptions of eyewitness confidence. Despite acknowledging that the witness did not know who the police suspect was, meaning that this information should not influence certainty expressions, participants perceived identical confidence statements as expressing lower (higher) values when the statement was about a member of the lineup that (mis-)matched the police suspect, than when the suspect was unknown. Logically this is an error.

However, a potential limitation of Experiment A is that Prior Knowledge is manipulated within-subjects. Participants could have altered their responses to no-context, suspect-ID, and filler-ID lineups, based on their expectation that there should be a difference between the three conditions. In Experiment B, we address this concern by manipulating knowledge of the police suspect in a between-subjects manner.

Experiment B

Experiment B directly replicates the method of Experiment A, though this time prior knowledge is manipulated between-subjects. If task demand is the sole factor accounting for discrepancies between the suspect-known and no-context lineups, then in this experiment we would expect minimal perceived differences in confidence statement ratings between each of the conditions.

Method

Participants

All 476 participants who passed study manipulation checks were located in the United States and completed the experiment over the internet using Amazon mechanical Turk (mean age = 36.36 years, SD = 11.59, range = 18 - 75, 58.61% female, 81.93% White/Caucasian). Participants were randomly assigned to one of six conditions, each comprised of 78 – 81 participants, representing the 3 x 2 intersection between the Prior Knowledge (suspect-ID, filler-ID, no-context) and Justification (Statement Only, Featural) manipulations.

Procedure

All participants viewed the same order of two lineups, randomly chosen from the pool of six lineups in Experiment A. We arbitrarily selected a high confidence and moderate confidence eyewitness statement to accompany each lineup, counterbalancing the order of presentation across participants. One face in each lineup was bordered in red, representing the eyewitness's selection, with both lineups corresponding to the same Prior Knowledge condition (either suspect-ID, filler-ID, or no-context). Participants in the suspect-ID and filler-ID conditions read the same instructions as in Experiment A, without the paragraphs describing that sometimes police officers do not know the police suspect, and pretended to be the officer who created the lineup. Those in the no-context condition received no information about the lineup generation process, and simply imagined themselves to be police officers. All other aspects of the design and procedure are identical to Experiment A. We excluded 259 individuals who either assigned values < 80% for the smiley lineup (N = 208), and/or failed one or more of the manipulation checks (N = 87). Data for this experiment is on the OSF (https://osf.io/8chmz/)

Experiment B Results and Discussion

We examined perceived confidence using a 2 x 3 x 2 mixed ANOVA with between-subjects factors of Justification (statement only, featural) and Prior Knowledge (suspect-ID, filler-ID, no-context), and a within-subjects factor of Confidence Level (moderate, high). In line with Experiment A, we observed a main effect of Justification, *F*(1,470) = 14.66, MS_e = 694.04, *p* < .001, η^2_p = .030, with lower levels of perceived confidence when participants encountered confidence statements with an accompanying featural justification (*M* = 61.46, *SD*= 20.37) than a confidence statement only (*M* = 68.06, *SD* = 19.49). Unsurprisingly, we found an effect of Confidence Level, *F* (1, 470) = 485.11, MS_e = 206.01, *p* < .001, η^2_p = .508, such that participants interpreted high confidence statements (*M* = 75.00, *SD* = 22.65) as meaning a higher value than moderate confidence statements (*M* = 54.50, *SD* = 21.70).

Interestingly, we observed a main effect of Prior Knowledge, *F* (2, 470) = 23.34, MS_e = 694.04, *p* < .001, η^2_p = .090. As in Experiment A, participants perceived the identical confidence statements as meaning a lower numeric value in the filler-ID condition (*M* = 56.52, *SD* = 21.20) than in either the no-context or suspect-ID conditions, *t* (294.82) = 5.41, *p* < .001, Cohen's *d* = .61, 95% CI [.38, .83] and *t*(309.56) = 5.84, *p* < .001, Cohen's *d* = .66, 95% CI [.43, .88], respectively. However, perceived confidence was comparable in the suspect-ID (*M* = 69.68, *SD* = 18.80) and no-context conditions (*M* = 68.00, *SD* 16.32), *t* (316) = .85, *p* = .394, Cohen's *d* = .10, 95% CI [-.12, .32].

We did not observe a three-way interaction, nor any significant two-way interactions with the Justification condition (all F's < 1.53, all *ps* ≥ .218, all $\eta^2_p \leq .006$). Additionally, in contrast to Experiment A, we did not find a significant interaction between Confidence Level and Prior Knowledge, *F* (2, 470) = 1.09, MS_e = 206.01, *p* = .338, η^2_p = .005.

Overall, we generally replicated Experiment A, even when using a more conservative Between-subjects design. Participants in the filler-ID condition, based on their own knowledge of the police suspect, perceived the witness as having a lower level of confidence, relative to the no-context condition, even though the same confidence statement was used in both conditions. While we did not replicate the finding of amplification of perceived confidence in the suspect-ID condition, the effects are in the same direction. These results rule out the critique that the effects of context in Experiment A were solely caused by task demands induced by its within-subjects design. Additionally, arguably, the within-subjects design is a more ecologically valid test as police officers often administer and interpret responses to multiple lineups over the course of their careers.

Given that participants in both Experiments A and B showed susceptibility to contextual effects when they interpreted eyewitness statements, in Experiment C we test whether clarifying the witness's confidence level using numerical information can protect against these errors.

Experiment C

One outstanding question in the judgement literature is the degree to which contextual effects are influenced by expressing confidence numerically instead of verbally. There are reasons to think that these effects will be weaker when witnesses clarify their level of confidence using numeric information. Research shows that context effects are most pernicious when evidence is ambiguous (Kassin et al., 2013). Multiple studies recommend that numerical probabilities are preferable to verbal probabilities because they are less vulnerable to flexible interpretations (e.g. Budescu et al., 2009, 2014; Nakao & Axelrod, 1983).

When interpreting eyewitness confidence, the ambiguity of verbal statements may lead investigators to place more weight on other accessible information, such as whether the witness selected the police suspect. In contrast, expressing eyewitness confidence numerically may serve to protect investigators from unintentionally misinterpreting the eyewitness's intended level of confidence, because there is less need to search for additional context behind the statement. For example, there should be little chance of misunderstanding the eyewitness's level of confidence when s/he expresses 100% confidence in the identification.

However, there is an alternative explanation. Eyewitnesses are viewed as inaccurate when they select someone from the lineup that is not the police's suspect, which in turn causes participants to discount and lower their perception of the eyewitness's level of certainty. This account is supported by recent evidence in studies of the justification bias. Cash and Lane (2017) found that participants perceived confidence statements with featural justifications as denoting both lower eyewitness certainty and *accuracy* than when confidence statements were presented alone. Moreover, Dodson & Dobolyi (2017) observed that even when participants were shown numeric expressions of confidence they still rated eyewitnesses as less accurate when the confidence statement was accompanied with a featural justification than when it appeared by itself.

Experiment C is the first test of whether expressing confidence numerically (rather than verbally) reduces the effects of prior knowledge on interpretations of eyewitness confidence statements. Moreover, as a conceptual replication of Dodson & Dobolyi (2017), we test whether these benefits extend to statements with featural justifications. If discordance is reduced in either case, then there may be enormous practical benefits to asking eyewitnesses for numeric expressions of confidence.

Method

Participants

In this experiment, 403 participants were randomly assigned to view numeric or verbal statements of certainty, with one of the two Justification conditions from Experiment A (n = 101 for all between-group combinations, excepting numeric x featural where n = 100). These participants, located in the United States, completed the experiment over the internet using Amazon's Mechanical Turk (mean age = 37.04, SD = 11.34, range = 18 - 72, 60.79% female, 84.12% Caucasian), and passed all manipulation checks. We calculated that there was > 95% power to detect moderate effects at an alpha level of .05.

Materials and Procedure

Materials and procedures are similar to Experiment A, except for two changes. First, half of the participants were presented with numeric expressions of eyewitness confidence, rather than verbal as before (see Table 2.1 for numeric confidence statements and justifications). The statements remained nearly identical; however instead of using an adverb of degree (e.g. "very") or an adjective (e.g. "positive") to indicate certainty, mock eyewitnesses reported 60% confidence in cases of moderate certainty, and 100% when highly certain. The other half of participants completed the experiment using the same confidence statements from Experiment A. To avoid suspicion, statements in the example lineups corresponded to the participant's numeric/verbal random group assignment.

Second, we altered the response scale to avoid mirroring the numeric statements of certainty. Participants indicated perceived confidence by using a slider in a visual analog scale, ranging from a low value of "Not at all certain" to "Completely Certain". The scale was subdivided into 6 equal sections to analogize the results to Experiments A and B. We did not include any visual demarcation points (e.g. tick marks) to avoid the potential numeric information of counting up the elements. All lineups, Prior Knowledge conditions, and counterbalancing remain identical to Experiment A. A total of 297 individuals did not meet criteria for analysis -- either assigning values < 80% for at least one smiley lineup (N = 149) and/or responding with "yes" to at least one manipulation check (N = 199). Data for this experiment is available on the OSF (https://osf.io/8chmz/).

Experiment C Results and Discussion

We used a mixed ANOVA to analyze the 2 x 2 x 3 x 2 factors of Format (verbal vs. numeric; between), Justification (statement only, featural; between), Prior Knowledge (suspect-ID, filler-ID, no-context; within), and Confidence Level (moderate, high; within). Although we did not find a four-way interaction, $F(2,798) = .43$, $\mathrm{MS_e} = .46$, $p = .651$, $\eta^2_p = .001$, there are several significant interactions and main effects.

Beginning with the main effects, we replicated previous findings for the Justification and Confidence Level manipulations, $F(1,399) = 15.67$, $\mathrm{MS_e} = 2.73$, $p < .001$, $\eta^2_p = .038$ and $F(1, 399) = 1171.09$, $\mathrm{MS_e} = 1.10$, $p < .001$, $\eta^2_p = .746$, respectively. Participants perceived more certainty for confidence only expressions ($M = 4.54$, $SD = .60$) than those with an accompanying featural justification ($M = 4.27$, $SD = .76$), and assigned increased values to high confidence statements ($M = 5.14$, SD .88) compared to moderate confidence statements ($M = 3.68$, $SD = .75$). A main effect of Prior Knowledge replicated Experiment A, $F(1.58, 629.21) = 110.96$, $\mathrm{MS_e} = 1.25$, $p < .001$, $\eta^2_p = .218$), with participants perceiving greater confidence for suspect-ID lineups ($M = 4.77$, $SD = .77$) than no-context lineups ($M = 4.42$, $SD = .77$), $t(402) = 8.82$, p < .001, Cohen's $d = .44$, 95% CI [.34, .54], which in turn showed increased perceived confidence compared to filler-ID lineups (*M* = 4.03, *SD* = 1.12), *t*(402) = 8.40, *p* < .001, Cohen's *d* = .42, 95% CI [.32, .52]. Finally, we found a main effect for Format, with numeric expressions (*M* = 4.51, *SD* = .66) showing higher perceived confidence than verbal expressions (*M* = 4.30, *SD* = .71), *F*(1,399) = 9.37, MS_e = 2.73, *p* = .002, η^2_p = .023.

The primary aim of this experiment was to investigate whether numeric, as compared to verbal, expressions of confidence reduce the effects of prior knowledge on judging an eyewitness's intended confidence. We found a significant three-way interaction between Confidence Level x Statement Format x Prior Knowledge, *F*(2, 798) = 13.46, MS_e = .46, *p* < .001, η^2_p = .033), which we discuss below. Replicating Dodson & Dobolyi (2017), we did not observe significant three- or two-way interactions between the Format and Justification manipulations, suggesting that participants were just as susceptible to the featural justification effect in the numeric condition as in the verbal condition.

Confidence Level x Statement Format x Prior Knowledge

Figure 2.3 displays the 3-way interaction between Confidence Level, Statement Format, and Prior Knowledge. The 3-way interaction is due to a stronger interaction between Format and Prior Knowledge for moderate confidence level statements, *F*(2,798) = 10.88, MS_e = .46, *p* < .001 η^2_p = .026, compared to high confidence level statements, F(2,798) = 4.37, MS_e = .46, *p* = .013, η^2_p = .011).

Beginning with moderate confidence statements, we found an effect of Prior Knowledge in both Format conditions. As seen in the top half of Figure 2.3, participants perceived both verbal and numeric expressions of confidence as meaning a higher value when the expression referred to a suspect-ID lineup (numeric *M* = 3.98, SD = .88; verbal *M* = 4.14, *SD* = 1.08) than a no-context lineup (numeric *M* = 3.75, *SD* = .76, *t*(200) = 3.35, p = .001, Cohen's *d* = .24, 95% CI [.10, .38]; verbal *M* = 3.69, *SD* = 1.09, *t*(201) = 5.37, *p* < .001, Cohen's *d* = .38, 95% CI [.23, .52]), which in turn were rated significantly higher than expressions of confidence that referred to a filler-ID lineup (numeric *M* = 3.40, *SD* = .98, *t*(200) = 5.18, *p* <.001 , Cohen's *d* = .36, 95% CI [.22, .51]; verbal *M* = 3.12, *SD* = 1.14, *t*(201) = 6.89, *p* < .001, Cohen's *d* = .48, 95% CI [.34, .63]). To determine if there is a meaningful attenuation of Prior Knowledge effects in the numeric compared to the verbal conditions, we computed difference scores between the no-context

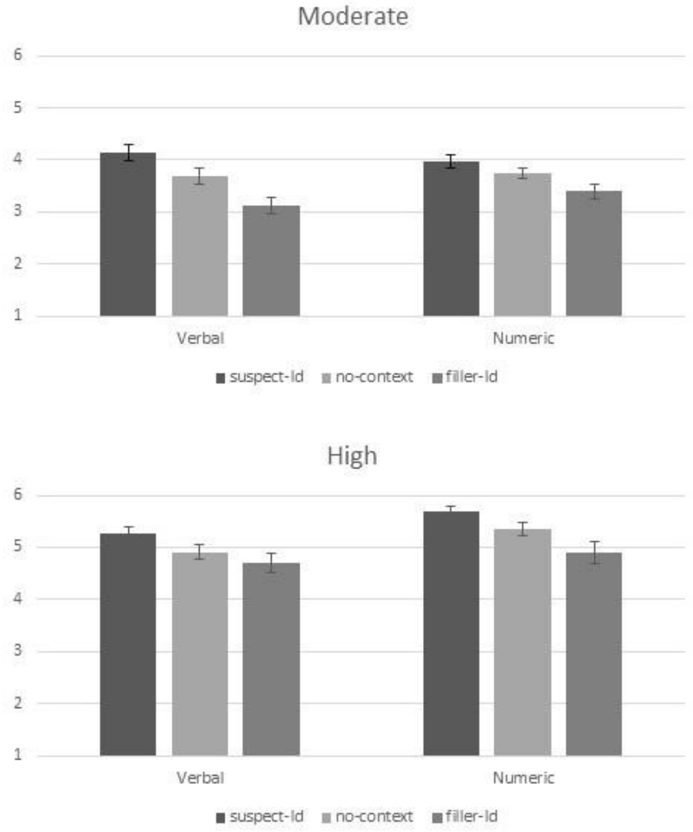


Figure 2.3. In Experiment C, participants evaluated the eyewitness's intended confidence when expressed either verbally or numerically. Higher scores indicate greater perceived confidence. The figure displays the interaction between intended Confidence of the expression, Format of the expression, and the Context condition. Error bars indicate 95% confidence interval of the mean

condition and the suspect-known conditions and compared these scores across statement Format. The comparison in perceived confidence between the suspect-ID and no-context conditions showed a smaller difference in the numeric format (*M* = .23, *SD* = .97) than in the verbal format conditions (*M* = .46, *SD* = 1.21), *t*(383.92) = 2.08, *p* = .038, Cohen's *d* = .21, 95% CI [.01, .40]. Similarly, the numeric condition (*M* = .35, *SD* = .95) produced a smaller difference in perceived confidence between the filler-ID and no-context conditions, relative to the verbal condition (*M* = .57, *SD* = 1.18), *t*(385.34) = 2.08, *p* = .039, Cohen's *d* = .21, 95% CI [.01, .40]. Overall, supporting our hypothesis, there are smaller effects of Prior Knowledge in the numeric condition than in the verbal condition when participants evaluate moderate levels of confidence.

The bottom half of Figure 2.3 shows the effects of Prior Knowledge for high confidence statements in both Format conditions. For both verbal and numeric statements of confidence, participants perceived the statements as meaning a higher value when they referred to suspect-ID lineups (numeric *M* = 5.69, *SD* = .73; verbal *M* = 5.26, *SD* = .95) than to no-context lineups (numeric *M* = 5.35, *SD* = 1.00, *t*(200) = 4.99, p < .001, Cohen's *d* = .35, 95% CI [.21, .49]; verbal *M* = 4.91, *SD* = 1.02, *t*(201) = 5.39, *p* < .001 , Cohen's *d* = .38, 95% CI [.24, .52]), which exceeded ratings for filler-ID lineups (numeric *M* = 4.90, *SD* = 1.53, *t*(200) = 5.04, *p* < .001, Cohen's *d* = .36, 95% CI [.21, .50]; verbal *M* = 4.71, *SD* = 1.37, *t*(201) = 2.22, *p* = .027, Cohen's *d* = .16, 95% CI [.02, .29]). As we did for the moderate confidence statements, we computed difference scores between the no-context condition and the suspect-ID and filler-ID conditions in order to examine our hypothesis that the numeric format will lessen the effect of Prior Knowledge as compared to the verbal format. There was little effect of numeric vs. verbal format on differences in perceived confidence in the comparisons between (a) the no-context and suspect-ID conditions (*M* = .34, *SD* = .96 vs. *M* = .35, *SD* = .91, respectively), *t*(401) = .088, *p*

= .930, Cohen's *d* = .01, 95% CI [-.19, .20), and (b) the no-context and filler-ID conditions (*M* = .45, *SD* = 1.27 vs. *M* = .20, *SD* = 1.30, respectively), *t*(401) = - 1.95, *p* = .052, Cohen's *d* = .19, 95% CI [.00, .39]). In sum, when participants evaluate high confidence statements, there are similar effects of Prior Knowledge on perceived confidence in both the numeric and verbal conditions.

Prior Knowledge x Justification

Finally, there was a significant two-way interaction between the Context and Justification manipulations, *F*(1.58, 629.21) = 5.00, MS_e = 1.25, *p* = .012, η^2_p = .012. As seen in Figure 2.4, we found near-equal featural justification effects for both suspect-ID (statement only *M* = 4.95, *SD* = .59 vs. featural *M* = 4.58, *SD* = .88) and no-context lineups (statement only *M* = 4.60, *SD* = .66 vs. featural *M* = 4.25, *SD* = .84), *t*(379.66) = 4.65, *p* < .001, Cohen's *d* = .46, 95% CI [.26, .66] and *t*(349.92) = 4.87, *p* < .001, Cohen's *d* = .48, 95% CI [.29, .68], respectively. However, deviating slightly from Experiments A and B, we found no significant difference between the statement only and featural conditions for filler-ID lineups, *t*(401) = .78, *p* = .436, Cohen's *d* = .08, 95% CI [-.12, .27].

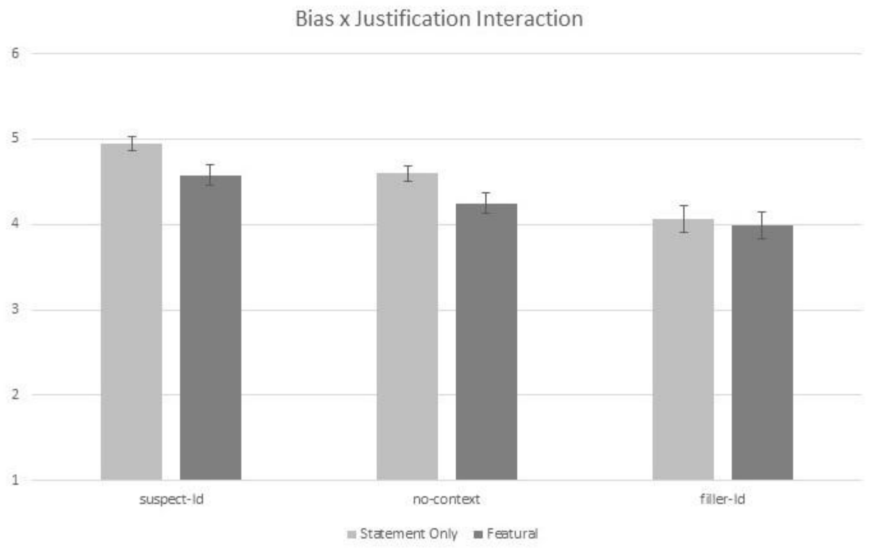


Figure 2.4. In Experiment C, we found a significant interaction between the Context and Justification manipulations. For both suspect-Id and no-context lineups, participants judged the identical expression of confidence as meaning a lower value when it was accompanied by featural information than when the confidence statement was presented alone. By contrast, there are no differences between the statement only and featural conditions for filler-Id lineups. Error bars indicate 95% confidence intervals of the mean.

Part II Discussion

For a moment, again imagine yourself as one of the investigators in the opening of this section. After noting the eyewitness's confidence about who they selected from the lineup, you take the short walk to your superior's office. "I see that the witness selected Face 1 from the lineup. How confident did he seem?" she asks. The results of three studies suggest that prior knowledge of the suspect will influence perceptions about the witness's certainty.

Across manipulations, Experiments A and C show that participants perceived the identical statement of confidence as meaning a higher level of certainty when the eyewitness's selection from the lineup matched the police suspect, relative to a no-context condition. Similarly, in all experiments (excepting "High Confidence" statements in Experiment A), participants perceived the same confidence statement as meaning a lower value when the eyewitness's selection differed from the police suspect, again relative to a no-context condition. These effects reflect judgment errors. Since participants acknowledged that the eyewitness is unaware of which lineup member is the police's suspect it is impossible for the eyewitness's level of certainty to be influenced by whether or not their choice happens to coincide with the police's suspect. Finally, results suggest that prior knowledge may add to pre-existing contextual effects, such as how the witness justifies their selection from a police lineup. Though, there is a notable exception in Experiment C, wherein the featural justification effect was eliminated for filler-ID lineups.

Is it possible to attenuate the effects of prior knowledge? We investigated whether numeric expressions of confidence would be less susceptible than verbal expressions to contextual influences because of their greater interpretive clarity. Supporting this hypothesis, Experiment C showed that participants were less influenced by knowledge of the police suspect when the witness expressed moderate confidence statements numerically rather than verbally, demonstrating some protective effect of a numeric format. However, the differences observed between the format conditions is small (about .1 SD units), and contextual knowledge still impacted perceptions of certainty for moderate statements. Moreover, when eyewitnesses imparted high confidence, there was little benefit to expressing confidence numerically, as Experiment C showed comparable effects of confirmation and disconfirmation on statement evaluations for both confidence formats. This is discouraging, because high confidence statements are most likely to be used to guide legal decision making. As a whole, these results suggest that there may be some benefit to asking witnesses to clarify their confidence using a numerical indicator, though this alone is unlikely to fully cancel out contextual effects.

We view the findings of these experiments with a caveat. We did not observe a significant difference between suspect-ID and no-context lineups when the Context condition was manipulated between participants in Experiment B, though participants continued to perceive less confidence for filler-ID lineups. We theorize that the consistent differences between the filler-Id and no-context conditions is due to a ‘presumption of calibration’ (Tenney, Spellman, & MacCoun, 2008), wherein participants view the eyewitness as well-calibrated (or justified) in their lineup decisions (especially if highly confident), unless given reason to think otherwise. Selection of a lineup filler is a strong signal that the eyewitness is poorly-calibrated, thus resulting in lower perceived confidence across all experiments, relative to the no-context control condition. However, task demand may explain why there are discordant findings for the suspect-Id and no-context lineup comparisons across experimental designs. In the between-subject condition, there is nothing to suggest that the witness’s selection for no-context lineups differs from the police suspect, meaning that participants simply treated these lineups as if they are the same as the suspect-Id lineups. In contrast, when participants viewed all Context conditions in Experiments A and C (i.e., within-subjects design), they may have inferred that there should be a clear ordering of witness accuracy, resulting in bolstered confidence ratings for the suspect-ID condition as compared to the no-context condition.

We would like to emphasize that we believe the within-subjects design is the more valid test of real-world practice, given that many legal roles will consider evidence in the light of their experience with previous cases. However, one might argue that the effects of knowing the suspect are relatively modest, with Cohen’s *d*s for statistically significant findings ranging from .16 to .61. On the other hand, there are reasons to think that contextual effects will have an even larger influence in the real-world as compared to what we observed in these experiments.

Our participants viewed lineups with an arbitrarily assigned police suspect. This differs from real-world practice where police officers construct lineups with *their own* suspect in mind. Research on the confirmation bias suggests that the tendency to interpret evidence as consistent with our beliefs will be stronger in the latter case where there is (or should be) evidence supporting a particular person as the suspect as compared to the former case where the suspect was simply labeled as the police's suspect (Bastardi, Uhlmann, & Ross, 2011; Klayman & Ha, 1987; Kunda, 1990; Nickerson, 1998). A recent survey of forensic science examiners found that many professionals show a bias blind spot, acknowledging that others may be affected by confirmation bias, but seeing themselves as impervious to its effects, or able to combat them by setting aside prior beliefs and expectations (Kukucka, Kassin, Zapf, & Dror, 2017). Experience in the field did little to protect against this assumption, while training on cognitive biases helped minimally. We predict that contextual effects will be even more pronounced when participants hold strong prior beliefs about suspect guilt before evaluating eyewitness statements.

Additionally, this study placed participants in the role of a police officer, yet we believe that the observed effects could generalize to decisions in other legal roles. For example, in deciding whether to admit eyewitness testimony, judges often refer to criteria set forth by the United States Supreme Court in *Neil v. Biggers* (1972) and expanded in *Manson v. Brathwaite* (1977) (see Wells & Quinlivan, 2009 for a review). These include whether the witness (a) had ample opportunity to view the individual, (b) paid sufficient attention, (c) provided a detailed description, (d) made the identification promptly after the event, and –crucially— (e) professed certainty about the identification. In assessing a case, judges are generally aware of the suspect, so they may be more likely to ascribe reliability to a witness statement when the identification matches the suspect than when it does not. This would result in more evidence implicating the suspect making it to trial than is warranted by the witness's certainty. Future studies should address whether experienced legal decision makers are similarly prone to these contextual effects.

Finally, our results suggest that current recommendations for assessing eyewitness confidence are incomplete. Present procedures, such as double-blind lineup administration, prevent investigators from influencing eyewitness reports (and should continue to be used!). However, they do not mitigate the effects of prior knowledge on interpretations of eyewitness statements. To reiterate, it is probable that legal decision makers are aware of the suspect in advance of hearing the witness's confidence statement. Are there any ways to ameliorate contextual effects?

One potential method comes from the post-identification feedback literature, wherein witnesses are asked to think about their certainty level after making an identification, but *before* receiving dis-/confirmatory feedback (e.g. "Good job, you identified the suspect") from an investigator (Neuschatz et al., 2007; Quinlivan et al., 2009; Wells & Bradfield, 1998, 1999). This 'confidence prophylactic' is shown to dampen the impact of feedback in the short-term (Neuschatz et al., 2007; Quinlivan et al., 2009; Wells & Bradfield, 1998, 1999), though is less effective in preventing long-term effects (Neuschatz et al., 2007; Quinlivan et al., 2009). Nevertheless, it may be beneficial to have non-blinded evaluators reflect on the confidence statement of the eyewitness, before showing them the witness's identification. Unlike witnesses, who may need to recall their certainty weeks or months after the fact for a trial, evaluators usually make decisions on a shorter time-scale, meaning that timing the intervention is less crucial. Undoubtedly, this will result in some difficulties related to other current recommended practices (e.g. video recording the lineup sessions), but making use of video editing and other technologies could mitigate these issues. Research should explore if this and other methods can limit the effects of prior knowledge on interpretations of eyewitness statements.

In summary, three experiments demonstrated that participants altered perceptions of verbal confidence ratings when they knew if a witness's identification matched the police suspect. Clarifying the witness's statement by using a numeric judgement of certainty may provide some protection against the biasing influences of prior knowledge when evaluators are interpreting moderate levels of confidence, but not when they interpret high levels of confidence. Finally, future efforts should be undertaken to explore contextual effects, and examine potential methods to limit biasing interpretations based on prior knowledge of the police suspect. We are certain that researchers will rise to this challenge, but leave it up to the reader to interpret the exact level of 'certainty'.

Part II References

Bastardi, A., Uhlmann, E. L., & Ross, L. (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological science*, *22*(6), 731-732.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of forecasting*, *1*(3), 257-269.

Bradfield, A. L., & Wells, G. L. (2000). The perceived validity of eyewitness identification testimony: A test of the five Biggers criteria. *Law and Human Behavior*, *24*(5), 581-594.

Bryant, G. D., & Norman, G. R. (1980). Expressions of probability: words and numbers. *The New England Journal of Medicine*, *302*(7), 411-411.

Budescu, D. V., Broomell, S., & Por, H. H. (2009). Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. *Psychological science*, *20*(3), 299-308.

Budescu, D. V., & Karelitz, T. M. (2003, July). Inter-Personal Communication of Precise and Imprecise Subjective Probabilities. In *ISIPTA* (pp. 91-105).

Budescu, D. V., Por, H. H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, *4*(6), 508-512.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, *36*(3), 391-405.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: ambiguous, context-dependent, or both?. *Organizational Behavior and Human Decision Processes*, *41*(3), 390-404.

Cash, D. K., & Lane, S. M. (2017). Context influences interpretation of eyewitness confidence statements. *Law and human behavior*, *41*(2), 180-190.

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*(4), 345.

Dobolyi, D.G., & Dodson, C.S. (2018). Actual vs. perceived eyewitness accuracy and confidence and the featural justification effect. *Journal of Experimental Psychology: Applied, 24*(4), 543 – 563.

Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and human behavior*, *39*(3), 266-280.

Dodson, C. S., & Dobolyi, D. G. (2017). Judging guilt and accuracy: highly confident eyewitnesses are discounted when they provide featural justifications. *Psychology, Crime & Law*, *23*(5), 487-508.

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification*, *56*(4), 600-616.

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic science international*, *156*(1), 74-78.

Dror, I. E., Kukucka, J., Kassin, S. M., & Zapf, P. A. (2018). No one is immune to contextual bias – not even forensic pathologists. *Journal of Applied Research in Memory and Cognition*.

Elaad, E., Ginton, A., & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making*, *7*(4), 279-292.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational behavior and human decision processes*, *45*(1), 1-18.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *15*(2), 140-152.

Gurmankin, A. D., Baron, J., & Armstrong, K. (2004). Intended message versus message received in hypothetical physician risk communications: exploring the gap. *Risk Analysis*, *24*(5), 1337-1347.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*(1), 42-52.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, *94*(2), 211-228.

Kovera, M. B., & Evelo, A. J. (2017). The case for double-blind lineup administration. *Psychology, Public Policy, and Law, 23*(4), 421-437.

Kukucka, J.K., Kassin, S.M., Zapf, P.A., & Dror, I.E. (2017). Cognitive bias and blindness: A global survey of forensic science examiners. *Journal of Applied Research in Memory and Cognition*, *6*(4), 452-459.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480-498.

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In *The handbook of eyewitness psychology, Vol II: Memory for people* (pp. 155-178). Lawrence Erlbaum Mahwah, NJ.

Manson v. Brathwaite. (1977). 432 U.S. 98.

Mullet, E., & Rivet, I. (1991). Comprehension of verbal probability expressions in children and adolescents. *Language & Communication*, *11*(3), 217-225.

Nakao, M. A., & Axelrod, S. (1983). Numbers are better than words: Verbal specifications of frequency have no place in medicine. *The American Journal of Medicine*, *74*(6), 1061-1065.

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

Neuschatz, J. S., Lawson, D. S., Fairless, A. H., Powers, R. A., Neuschatz, J. S., Goodsell, C. A., & Toglia, M. P. (2007). The mitigating effects of suspicion on post-identification feedback and on retrospective eyewitness memory. *Law and Human Behavior*, *31*(3), 231-247.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175-220.

Quinlivan, D. S., Neuschatz, J. S., Jimenez, A., Cling, A. D., Douglass, A. B., & Goodsell, C. A. (2009). Do prophylactics prevent inflation? Post-identification feedback and the effectiveness of procedures to protect against confidence-inflation in earwitnesses. *Law and Human Behavior*, *33*(2), 111-121.

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of applied psychology*, *74*(3), 433.

Steblay, N. K., Wells, G. L., & Douglass, A. B. (2014). The eyewitness post identification feedback effect 15 years later: Theoretical and policy implications. *Psychology, Public Policy, and Law*, *20*(1), 1-18.

Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, *44*(5), 1368-1375.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society*, *31*(2), 135-138.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, *25*(5), 571-587.

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*(3), 360-376.

Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the postidentification-feedback effect be moderated?. *Psychological Science*, *10*(2), 138-144.

Wells, G. L., & Quinlivan, D. S. (2009). Suggestive eyewitness identification procedures and the Supreme Court's reliability test in light of eyewitness science: 30 years later. *Law and Human Behavior*, *33*(1), 1-24.

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

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*(1), 10-65.

Yates, S.Q. (2017, Jan 1). Memorandum for heads of department law enforcement components all department prosecutors. *Subject: Eyewitness identification: Procedures for conducting photo arrays*. https://www.justice.gov/archives/opa/press-release/file/923201/download.