The Role of Feature and Familiarity Justifications on the Interpretation of Eyewitness Confidence

David Gyula Dobolyi

Charlottesville, VA 22902

Bachelor of Arts, University of Maryland, 2007

Master of Arts, University of Virginia, 2012

A Dissertation Proposal Presented to the

Graduate Faculty of the University Of Virginia

In Candidacy for the Degree of Doctor of Philosophy

Department of Psychology

University of Virginia

August, 2015

Abstract

I conducted four experiments that advance our knowledge of the interpretation of eyewitness confidence. These experiments focused on four key questions: (1) are eyewitness justifications—when combined with confidence and decision time—meaningful postdictors of identification accuracy?; (2) how accurately can observers interpret the intended meaning of an eyewitness's confidence statement given a particular type of justification?; (3) what is the consequence of a particular kind of justification on an observer's behavior?; and (4) are differences in perceived confidence across different types of justifications a result of expertise with faces or do these findings represent a more general memory phenomenon?

These experiments yielded several key findings: 1) familiarity justifications were significantly more likely to occur when not choosing within a lineup than any other justification type; 2) when choosing a face from a lineup, familiarity justifications were associated with lower accuracy and a poorer confidence/accuracy relationship at higher levels of confidence than other justification types; 3) differences in perceived confidence were minimized when justification type varied within-subjects, although multiple observable features were perceived as more confident than a single observable feature; 4) highly confident unobservable justifications were rated as stronger evidence than both observable featural justifications and confidence alone; and 5) the featural justification effect was not specific to faces, but also occurred for novel objects (e.g., greebles) and other crime-relevant stimuli (e.g., cars and weapons), suggesting that expertise alone does not drive the *featural justification effect* (i.e., the finding that observable justifications are perceived as less confident than unobservable justifications and confidence alone; Dodson & Dobolyi, 2015)—rather it may rely on judgments of the perceived memorability of features made "on the fly."

Keywords: eyewitness confidence, eyewitness identification, eyewitness memory, confidence justifications, confidence-accuracy, decision time, unobservable inflation effect

Acknowledgments

I would like to thank everyone who has helped me over the course of the past six years. First and foremost, I wish to extend my gratitude to my advisors, Dr. Chad Dodson and Dr. Michael Kubovy. If it were not for Chad, I may not have attended graduate school: his kind response to an email I sent from New Zealand expressing an interest pursuing psychology led to everything that has since followed, including this dissertation. Moreover, his unwavering support and guidance over these many years has been a cornerstone of my graduate experience.

Dr. Michael Kubovy introduced me to my second line of research involving "big data," which has since led to post-doctoral employment. Michael's career advice throughout the years has been critical to my professional development, and I am exceptionally grateful. Additionally, both Michael and Dr. Eric Turkheimer were responsible for fostering my passion for statistics and data analysis from the beginning by way of the graduate statistics curriculum. I learned a great deal from them not only as a student, but also as a teaching assistant.

I also wish to thank Dr. Daniel T. Willingham for his continued personal and professional advice. Whenever I had any kind of question, he was always willing and able to answer it. I owe the same gratitude to Dr. Wesley Weimer, who kindly volunteered to serve as my out of area representative after having provided me with advice even prior to joining the graduate school, and also to Dr. Jonathan Haidt, who continues to offer invaluable career advice and support whenever needed.

A wide range of individuals have helped me learn a great deal about conducting research: in no particular order, Dr. Steven M. Boker, John R. Nesselroade, Dr. Brian A. Nosek, Dr. Timo von Oertzen, and Dr. Dennis T. Proffitt. I also owe a great deal of thanks for my many research assistants, and in particular Stephen Wisner, who worked tirelessly even into the early morning

hours. I also wish to thank the department staff for all of their help, including but not limited to Rebecca G. Anderson, Morgan Davis, Donna L. Hearn, Tabitha Lillard, Tammy Seal, Debra Snow, Vivienne S. Spauls, and Stacy W. Sties.

Finally, I wish to extend my gratitude to several friends and fellow graduate students for their support, including Diana Dinescu, Matthew Domiteaux, Claire La Fleur, Laura Getz, Meret Hofer, Murteza Husain, Alan Kush, Nauder Namaky, Rachel Narr, Benjamin Scott, Adi Shaked, and Spencer Watkins, to name only a few. And last but certainly not least, I am grateful to my parents and in particular my sister, Dr. Kinga Dobolyi, for her continued love and support.

Table of Contents

The Role of Feature and Familiarity Justifications on the Interpretation of Eyewitness Confidence

Eyewitness identifications have major legal implications because these identifications influence both police investigations and jury decision-making (e.g., Semmler, Brewer & Douglass, 2012). Existing research has shown repeatedly that eyewitness accuracy is related to confidence, and strongly so in the case of positive identifications (i.e., lineups from which a suspect was chosen; e.g., Brewer & Wells, 2006; Sauerland & Sporer, 2009).

There is, however, one fundamental gap in the research literature and this is the issue of interpretation of verbal expressions of eyewitness confidence by other parties. Nothing is known about how accurately others interpret expressions of confidence generated by eyewitnesses. Consider the example of an eyewitness who makes an identification from a lineup and states, "I'm pretty sure it's him." Although the eyewitness may intend this statement to mean that he is 80% confident, others may interpret it as meaning that he is 50% confident. A fundamental question, then, is when eyewitnesses give a verbal expression of certainty about their identification does this expression mean the same thing to them as it does to police, jurors, and others who receive and interpret this expression?

Although there is no research on understanding the intended meaning of eyewitness expressions of confidence, there has been over 40 years of research in other domains on how well others can interpret verbal expressions of probability and confidence. These studies have examined contexts including but not limited to: climatology and global warming (e.g., Budescu, Por, & Broomell, 2012), disease transmission during the SARS epidemic (e.g., Young & Oppenheimer, 2009), and risk related to medication side effects (e.g., Young & Oppenheimer, 2006). The common account identified by all of these studies is that people have difficulty

understanding verbal expressions of probability (Budescu et al., 2012). In the case of the climatology studies, the impact of this disagreement was certainly not trivial: in one study, only 7.7% of participants interpreted the phrase "likely" in the way intended by the Intergovernment Panel on Climate Change (Budescu, Broomell, & Por, 2009). This disagreement is also not limited to naïve audiences, as inter-individual differences in interpretation also influences experts within a field (e.g., Beyth-Marom, 1982; Wallsten, Fillenbaum, & Cox, 1986).

Given the difficulty of interpreting verbal expressions of confidence, one might assume that individuals would avoid using verbal expressions and instead would express their confidence numerically (e.g., "the outcome is 80% certain"). This assumption would be wrong: research across a variety of paradigms has shown that individuals prefer using words (e.g., "the outcome is pretty certain") rather than numbers to express uncertainty (e.g., Brun & Teigen, 1988; Budescu, Karelitz, & Wallsten, 2003; Erev & Cohen, 1990). For example, a survey by Wallsten, Budescu, Zwick, and Kemp (1993) found that 65% of respondents preferred communicating probability using verbal expressions of confidence (although ironically, 70% preferred receiving numeric statements). However, it is noteworthy that even numeric expressions of certainty are open to flexible interpretation by others (e.g., Flugstad & Windschitl, 2003; Windschitl, Martin, & Flugstad, 2002), and the relationship between confidence and accuracy is nearly identical regardless of whether verbal or numeric scales are used (e.g., Weber, Brewer & Margitich, 2008). Thus, it should be clear that research is needed in areas where misinterpretation is associated with serious consequences, such as within the eyewitness context.

Currently, police are advised to ask eyewitnesses to state how certain they are in their own words about a lineup identification (Technical Working Group on Eyewitness Evidence, 2003). However, until recently, no published research has examined how others understand

these expressions of eyewitness confidence. Our preliminary investigation explored how individuals express and interpret confidence in an eyewitness paradigm, revealing three main findings (Dodson & Dobolyi, 2015). First, we found that providing an additional justification of confidence (e.g., "I'm very sure it's him. I remember his chin"), in contrast to a statement of confidence alone (e.g., "I'm very sure it's him"), increases misunderstanding in others. Second, we observed that these justification-induced misunderstandings occur only when expressions of confidence refer to a specific, observable facial feature (e.g., "I remember his nose") but not to unobservable qualities (e.g., "He is really familiar"). Third and perhaps most notably, we found the extent of misinterpretation to be greatest when eyewitnesses were most confident in their responses (e.g., "I am certain"). We refer to the culmination of these findings as the *featural justification effect*, which suggests—somewhat counter intuitively—that expressions of confidence referring to observable, featural responses give participants specific additional information that is open to interpretation and judgment, leading to greater misunderstanding of the intended meaning of confidence statements (e.g., although an eyewitness said "I remember his nose," is that nose truly memorable?). In contrast, for unobservable responses, there is no analogue for judging the likely accuracy of the response and thus less misinterpretation (e.g., if an eyewitness says "I just remember him," your options are to either accept that statement as fact or not—the statement itself is not open to objective reevaluation).

For my dissertation I conducted four experiments that advance our knowledge of the interpretation of eyewitness confidence. These experiments focused on four key questions: (1) are eyewitness justifications—when combined with confidence and decision time—meaningful postdictors of identification accuracy?; (2) how accurately can observers interpret the intended meaning of an eyewitness's confidence statement given a particular type of justification?; (3)

what is the consequence of a particular kind of justification on an observer's behavior?; and (4) are differences in perceived confidence across different types of justifications a result of expertise with faces or do these findings represent a more general memory phenomenon?

Experiment 1

Previous research has shown that faster lineup identifications are more likely to be accurate than slower identifications, and that confidence and accuracy are more strongly associated when positive identifications are made (e.g., Brewer & Wells, 2006; Dunning & Stern, 1994; Sauerland, Sagana, & Sporer, 2012; Weber & Brewer, 2004). While this research has been crucial from a legal perspective, it has overlooked an important aspect of eyewitness identifications: namely justifications of confidence, and whether or not these justifications are associated with other aspects of the identification such as decision time and accuracy. For example, is it the case that an eyewitness who refers to an observable feature (e.g., "I remember his nose") as a basis for a response is more likely to make a faster, more accurate identification than one who refers to familiarity (e.g., "He's familiar")?

This is possible, given that Reinitz and colleagues have observed that: 1) people are both more accurate and more confident when they remember an event based on featural versus familiarity information; and 2) conversely, when controlling for confidence, people are more likely to be overconfident when their memory for a face is based on featural information rather than familiarity (Reinitz, Peria, Seguin, & Loftus, 2011; Reinitz, Seguin, Peria, & Loftus 2012). However, it is unclear if their findings extend to the eyewitness setting, since their paradigm differed from typical eyewitness identification tasks.

We previously observed that eyewitnesses used three types of justifications for lineup identifications (Dodson & Dobolyi, 2015): 1) familiarity statements, such as "He's familiar,"

which do not refer to specific details about the lineup faces; 2) observable feature statements, such as "I remember his chin," which do mention specific, observable facial features; and 3) unobservable feature statements, such as "He looks like a friend of mine," which refer to specific features that are not directly observable by a third party. However, Dodson and Dobolyi (2015) did not investigate if particular types of identifications—for example, those made quickly and thus more accurately—are more likely to be associated with a particular type of justification.

Previous research has shown that confidence and decision times are predictive of accuracy, particularly when multiple postdictors are combined (Brewer & Wells, 2006; Sauerland & Sporer, 2009; Weber, Brewer, Wells, Semmler, & Keast, 2004). However, these findings are strongest when participants identify a face within a lineup (i.e., when choosing; Sauer, Brewer, Zweck, & Weber, 2010; Sporer, Penrod, Read, & Cutler, 1995). For example, Sauerland and Sporer (2009) observed that fast (6s or less) and highly confident (90 – 100%) individuals showed an impressive 97% accuracy rate when they selected someone from a lineup; by contrast, slow and unconfident participants were only 32% accurate. Dunning and Perretta (2002) identified a cutoff point they refer to as the *10- to 12-second rule*, according to which faster responses produce accuracy rates of 90% or higher whereas slower responses produce accuracy rates of 50% or lower. To explain this finding, they refer to the work of Dunning and Stern (1994), who found that accurate identifications are associated with automatic recognition or "pop-out" that occurs quickly as opposed to an effortful process of elimination among the lineup faces that takes substantially longer. By contrast, Brewer, Weber, Clark, and Wells (2008) argue that the *10- to 12-second rule* must be considered in terms of other boundary conditions such as retention interval, lineup size, and facial distinctiveness; nevertheless, they also found a similar overall pattern: higher accuracy when identifications occurred more quickly.

For non-identifications (i.e., "not present" responses), however, confidence and decision times are much less powerful for predicting accuracy (e.g., Brewer & Wells, 2006; Dunning & Stern, 1994; Sauer et al., 2010). Nevertheless, under various boundary conditions postdictors can also be predictive when participants reject faces within a lineup. For example, by identifying non-choosers who were highly confident, responded quickly, and convinced that the target was "absent," Sauerland et al. (2012) noted a non-identification accuracy of 87.5%; by contrast, non-choosers who met none of these criteria and were "insecure" about their decisions were only 37.5% accurate.

Experiment 1 uses a standard eyewitness recognition task to determine if faster, more accurate responses are more likely to involve feature- versus familiarity-based recognition. Participants will view black and white faces and then after a delay their memory for these faces will be tested using a series of lineups in which a previously seen face may or may not be present. In addition to identifying a face or making a "not present" response, participants will provide written expressions of certainty consisting of 1) verbal expressions of confidence and 2) justifications of confidence. By coding these eyewitness expressions based on their content and combining that information with identification accuracy and decision time, Experiment 1 will determine if the accuracy of a lineup identification can be predicted based on (a) how quickly eyewitnesses respond, (b) how confident they are, and (c) how their confidence is justified.

Method

Participants. Participants were 275[1] white individuals between the ages of 18 and 40 (M = 27.82, SD = 5.24, range = 18.0 – 39.3; 57.81% female) who were recruited via Amazon

[1] The initial sample consisted of 384 participants who completed the task in full. Of these, 34 were removed for failing to complete the smiley instructions check correctly (i.e., by not providing a numeric confidence of 100) and one more was removed for providing one or more

Mechanical Turk in exchange for payment. An a-priori power analysis conducted using G*Power (Faul, Erdfelder, Lang & Buchner, 2007) with $\alpha = .05$ showed that we would have over 99% power to detect medium-sized effects (Cohen's $f = .25$; Cohen, 1988) in the context of a repeated measures ANOVA with 240 participants. All participants gave consent and completed a brief demographic questionnaire.

Design. The experiment was entirely within-subjects, consisting of a 2 (Lineup Race: Same Race vs. Cross Race) by 2 (Target Presence: Target Present vs. Target Absent) design.

Materials. The experiment was conducted using a custom browser-based framework built using PHP, jQuery/JavaScript, MySQL, and HTML. The entire experiment is available online at http://dodsonlab.com/studies/faces_rate_mult_2014/. Stimuli for the experiment consisted of the six black and six white lineups used in Dobolyi and Dodson (2013). These stimuli also included both a casual and a formal photo of each lineup target so that the faces shown during encoding would not be identical to those shown at test.

Procedure. The procedure was essentially identical to Dobolyi and Dodson (2013). During encoding, participants viewed a randomized series of six black and six white faces with casual facial expressions (e.g., smiling) and varied street clothing in a "head and shoulders" shot. The 12 faces repeated as a block four times with the stipulation that the same face not appear at the end of one block and at the beginning of the next (i.e., the same face never appeared back to back). Each face was shown for three seconds with a one second interstimulus interval. An additional four faces (two black and two white) appeared at the beginning of encoding as

---

blank confidence statements, leaving 349. Another 67 (or 19.20%) were then removed for providing one or more verbal expressions that included numbers (e.g., "I am 80% certain that I have not seen these faces before"), yielding 282. Finally, seven more participants were removed during the coding of verbal expressions: of these, five misunderstood task instructions, one made disparaging remarks about the task, and one simply said "yes" for every lineup.

primacy fillers and another set of four appeared once at the end as recency fillers; none of the fillers appeared again during the task.

After a five-minute delay involving a word search, participants were instructed to pretend they were eyewitnesses to a crime. Participants were told that they would see a series of lineups with six faces per lineup and that the photos in these lineups would not be identical to the ones they saw earlier (i.e., in place of casual attire and expressions, lineups members all wore an identical maroon t-shirt and exhibited neutral facial expressions). Participants were also informed that some lineups would contain a previously witnessed face, whereas others would not. As shown in Figure 1, the participant's task was either to identify one of the faces in the lineup or to make a "Not Present" response by highlighting their selection with a mouse click.

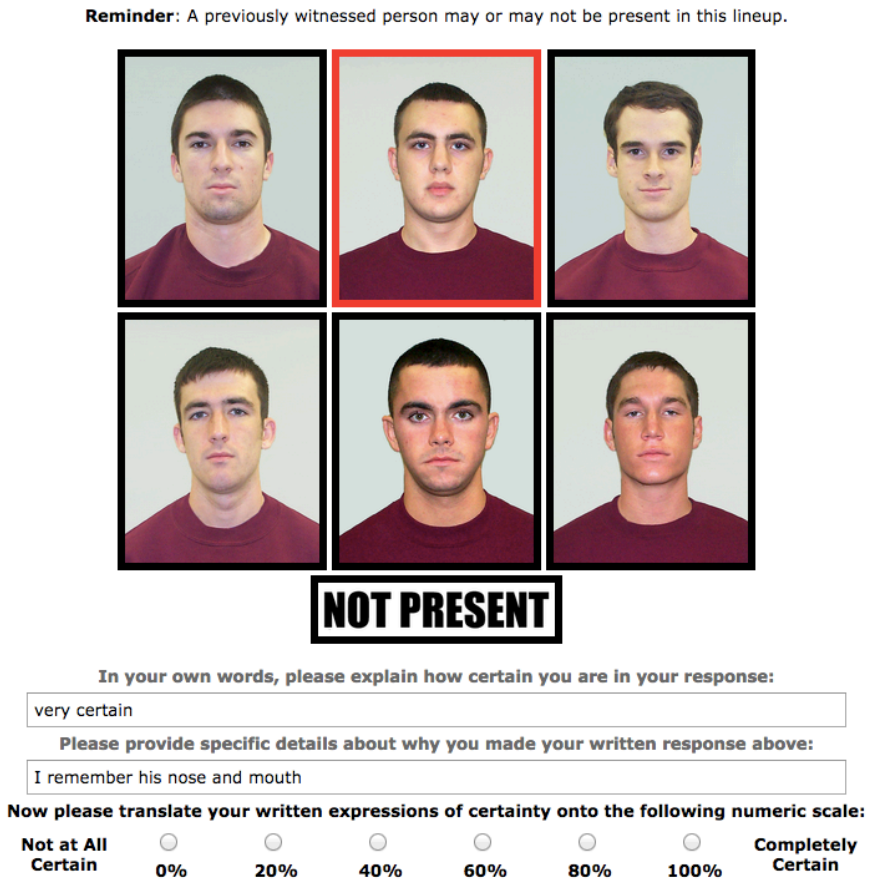


Figure 1. An example of the lineup recognition task from Experiment 1. The participant has identified the top-center face as having previously been seen by highlighting it in red.

Participants were given further instructions on how to provide expressions of certainty for their identifications. To help introduce the interface, participants submitted a series of responses using text entry boxes that exactly matched those shown underneath lineups during the actual task (see Figure 1). The first box included the following instructions: "In your own words, please explain how certain you are in your response"—language that exactly corresponds to National Institute of Justice guidelines about asking eyewitnesses about an identification (p. 39, Technical Working Group on Eyewitness Evidence, 2003). They were then asked to explain this confidence statement by providing "specific details about why you made your written response above." Finally, they translated their written expressions of certainty onto a six-point numeric scale ranging from 0% (Not at All Certain) to 100% (Completely Certain).

Before moving on to the actual task, participants were shown a yellow smiley face and then immediately completed a practice lineup involving six palette-swapped smiley faces. Because the yellow smiley face was always present, this lineup served as an instructions check: only participants who correctly identified the yellow smiley face with complete certainty (i.e., by providing a numeric confidence rating of 100) were included in the final dataset.

The 12 critical lineups consisted of six black and six white lineups, one for each target face shown during encoding. Within each lineup race, half of the lineups included a previously witnessed face (i.e., Target Present) and half did not (i.e., Target Absent). Lineup order was randomized such that no more than three lineups in a row were Target Present or Target Absent, and no more than three lineups in a row were black or white.

After completing the 12 lineups, participants were asked to fill out a short demographics survey including questions about age, sex, and race. Finally, participants were thanked for their involvement and debriefed.

Results

Replication of Prior Results. The 275 participants' data produced an initial sample of 3300 lineup observations (i.e., because each participant assessed 12 lineups). Previous work by Dodson & Dobolyi (under review) has shown a relationship between confidence, decision time, and choosing type on accuracy using a similar eyewitness paradigm. To replicate this effect, I fit an identical linear mixed model using *lme4* (Bates et al., 2014; version 1.1-7) in *R* (R Core Team, 2015) to the one used in that study, except for one difference: the present experiment only used a single confidence scale (i.e., a six-point scale from 0 to 100) whereas theirs used nine different confidence scale types.[2]

Prior to running the model, I cleaned the data using the same method described in Dodson & Dobolyi (under review). First, I log transformed decision time and removed 51 observations from the data (1.55%) that were more than three median absolute deviations from the sample median, leaving 3249 observations for the model dataset. I then fit a generalized linear mixed model of binomial accuracy using a participant intercept for the random effects and the full interaction of the following fixed effects: Lineup Response (Chooser vs. Non-Chooser), Lineup Race (Same Race vs. Cross Race), Confidence, and Decision Time. The latter two

[2] Similar to Dodson and Dobolyi (under review), I use regression rather than receiver operating characteristic (ROC) to assess the data because I am interested in the interactions of continuous predictors (e.g., confidence and decision time). Because a partial area under the curve (pAUC) comparison must be conducted categorically, the method is not suited for answering key questions in the present design (e.g., how does justification type interact with decision time continuously across the full range of responses?).

continuous predictors (i.e., Confidence and Decision Time) were centered and scaled prior to running the model.

A likelihood ratio test of the model conducted using the *afex* package (Singmann, Bolker, & Westfall, 2015; version 0.13-145) showed nearly identical effects to those observed in Dodson & Dobolyi (under review). A full summary of significant terms is provided in Table 1 below, but I focus on the highest order effects to which lower order effects are marginal: 1) a three-way interaction between Lineup Response, Confidence, and Decision Time, $\chi^2(1) = 7.67$, $p < .01$, and 2) a two-way interaction between Lineup Response and Lineup Race, $\chi^2(1) = 4.26$, $p = .04$.

Effect	χ^2	**df**	**p**
DecisionTime	2.39	1	.12
Confidence	258.03	1	<.0001
LineupRace	10.14	1	<.01
LineupResponse	93.02	1	<.0001
DecisionTime:Confidence	4.22	1	.04
DecisionTime:LineupRace	1.55	1	.21
Confidence:LineupRace	1.36	1	.24
DecisionTime:LineupResponse	5.90	1	.02
Confidence:LineupResponse	135.39	1	<.0001
LineupRace:LineupResponse	4.26	1	.04
DecisionTime:Confidence:LineupRace	2.85	1	.09
DecisionTime:Confidence:LineupResponse	7.67	1	<.01
DecisionTime:LineupRace:LineupResponse	0.01	1	.93
Confidence:LineupRace:LineupResponse	1.52	1	.22
DecisionTime:Confidence:LineupRace:LineupResponse	1.07	1	.30

Table 1. The likelihood ratio table for the four-way interaction of the fixed effects predicting identification accuracy from Lineup Response, Lineup Race, Confidence, and Decision Time. In Wilkinson-Rogers notation, a colon (i.e., ":") indicates an interaction (Wilkinson & Rogers, 1973).

Figure 2 below shows the significant three-way interaction; model estimates were computed using the *effects* package (Fox, 2003; version 3.0-4). Consistent with Dodson and Dobolyi (under review), chooser lineups (i.e., ones in which a face was selected) showed a strong relationship between confidence and accuracy, particularly for decisions that were made quickly (e.g., notice that the red line in the left panel for 100% confident choosers is nearly at 100% accuracy for identifications made in less than five seconds); by contrast, for non-choosers (i.e., "not present" responses) neither decision time nor accuracy strongly predicted accuracy, as shown by the tightly grouped and parallel confidence lines across the full range of decision time. Note that accuracy tends to be higher for non-chooser lineups because half of the lineups were target-absent within the design. Moreover, chance when choosing is lower than chance for non-choosing, since five out of the six lineups are foils and thus open to mistaken identification responses.

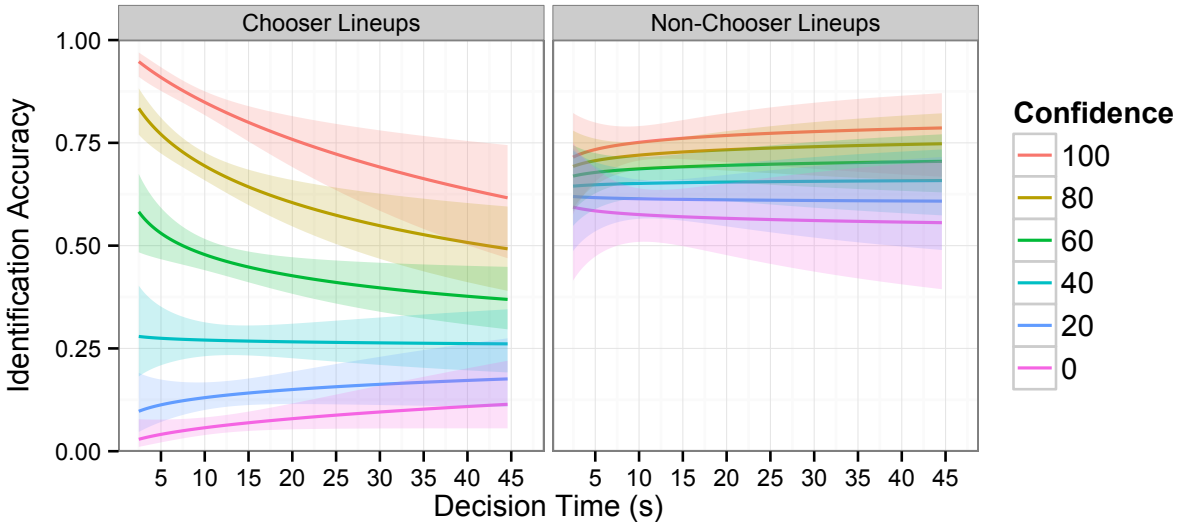


Figure 2. The three-way interaction between lineup response, confidence, and decision time on identification accuracy. Similar to Dodson and Dobolyi (under review), I find a strong

confidence/accuracy relationship for choosers (left panel), but not for non-choosers (right panel). Error shading represents a 95% confidence interval.

Figure 3 below shows the significant two-way interaction between Lineup Response and Lineup Race. Based on a follow-up contrast conducted using the *phia* package (De Rosario-Martinez, 2015; version 0.2-0), accuracy is higher for same-race chooser lineups than for cross-race chooser lineups, $\chi^2(1, N = 256) = 12.45$, $p < .001$. By contrast, there is no difference in accuracy for non-choosers based on lineup race, $\chi^2(1, N = 256) = 0.66$, $p = .42$. This effect was also previously found in Dodson and Dobolyi (under review).

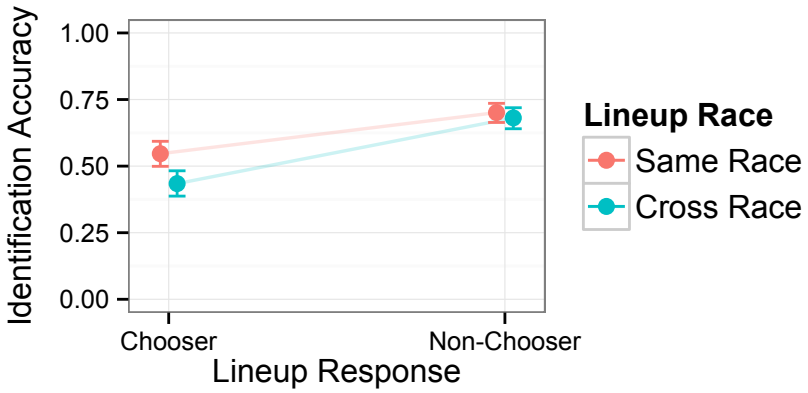


Figure 3. The two-way interaction between lineup response and lineup race on accuracy. For chooser lineups (left), accuracy is higher for same-race lineups than cross-race lineups. By contrast, for non-chooser lineups (right), there is no difference based on lineup race. Note that lines connecting points across lineup response are provided only to visually aid interpretation of the interaction. Error bars represent a 95% confidence interval.

Use of Justification Types. While the effects shown in the previous section replicate prior work (i.e., by Dodson & Dobolyi, under review), the primary purpose of Experiment 1 is to

evaluate how different justifications of confidence influence the relationship between accuracy and lineup race, lineup response, confidence, and decision time. To accomplish this, the first step was coding the participants' verbal expressions into five pre-determined categories guided by the findings of Dodson and Dobolyi (2015): 1) familiarity, which explicitly require the use of the word "familiar" (e.g., "He is very familiar"); 2) observable feature, which included a count of features mentioned (e.g., "I remember his nose and mouth" [two observable features]); 3) unobservable feature, which also included a count (e.g., "He looks like a friend of mine" [single unobservable feature]); 4) mixed, meaning more than one of the preceding categories was used; and 5) unknown, which was used when the statements did not fit into one of the pre-defined categories (see the following paragraph for examples of unknown statements). The statements were randomly divided into four sets, and each set was given to two of eight research assistants for categorization (see Appendix B for the instructions that were given to the research assistants, additional details on the coding process, and sample statements made by participants). Inter-rater agreement on the categorization of the 3300 original statements was 84.61% (i.e., 2792 statements were categorized in the same way by both research assistants; across the four rater pairs, overall agreement was relatively consistent: 85.29%, 87.56%, 86.96%, and 78.62%). Statements that were not agreed upon were removed from the analysis.

Of the 2792 statements that showed agreement, a total of 1334 (47.78%) were categorized as unknown. The majority of these unknown statements (74.14%) were made when participants chose the "not present" response when evaluating a lineup (e.g., "I don't recognize any of them," "these faces were not shown"). For the smaller subset of positive identifications, unknown statements included examples like "I think I remember him but am not sure" or "I definitely remember this guy." These statements were also excluded from the following analyses, leaving

a total of 1621 statements that were both agreed upon by raters and fell within the categories of interest: familiarity, observable feature, unobservable feature, and mixed.[3]

Table 2 below breaks down the 1621 statements across factors related to the associated lineup: first by Lineup Response (Chooser vs. Non-Chooser), then by Lineup Race (Same Race vs. Cross Race), and finally by Justification Type (Familiarity [F], Observable Feature [O], Unobservable Feature [U], and Mixed [M]). For justification types involving a feature, a number is included next to the category code in reference to how many features were mentioned (e.g., O2 refers to a statement that mentioned two observable features). The table also includes descriptives: mean confidence ratings, mean decision times, mean accuracy, and counts.

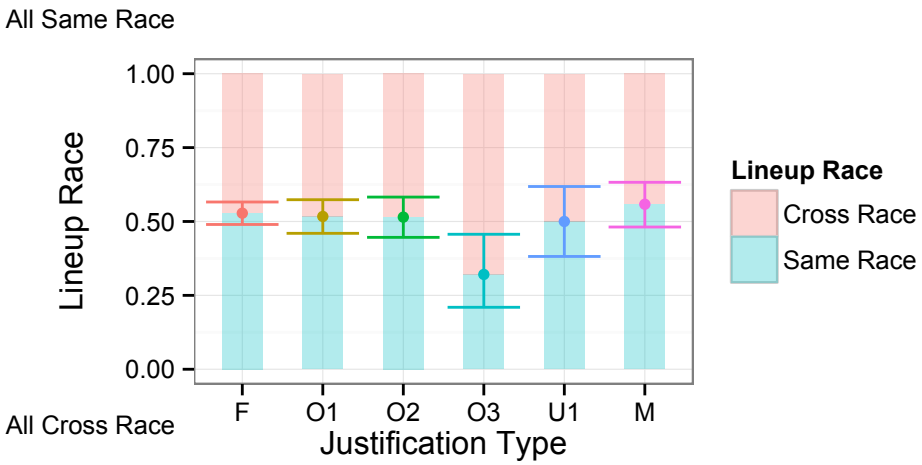
[3] Mixed justifications included 163 statements (e.g., "The eyebrows look familiar").

Lineup Response	Lineup Race	Justification Type	Mean Confidence	Mean Decision Time (ms)	Mean Accuracy	n
Chooser	Same Race	F	40.42	15932	.33	96
Chooser	Same Race	M	53.25	15727	.48	77
Chooser	Same Race	O1	72.05	13135	.60	146
Chooser	Same Race	O2	75.00	13913	.66	100
Chooser	Same Race	O3	83.53	16818	.53	17
Chooser	Same Race	O4	NA	NA	NA	NA
Chooser	Same Race	O5	100.00	5762	.75	4
Chooser	Same Race	U1	81.25	14375	.75	32
Chooser	Same Race	U2	73.33	36986	.33	3
Chooser	Cross Race	F	45.64	13809	.25	117
Chooser	Cross Race	M	54.06	16810	.42	64
Chooser	Cross Race	O1	77.01	13182	.56	134
Chooser	Cross Race	O2	72.13	14398	.56	94
Chooser	Cross Race	O3	77.58	12009	.58	33
Chooser	Cross Race	O4	80.00	8599	.67	9
Chooser	Cross Race	O5	93.33	8871	1.00	3
Chooser	Cross Race	U1	80.67	11687	.73	30
Chooser	Cross Race	U2	NA	NA	NA	NA
Non-Chooser	Same Race	F	61.35	12205	.69	252
Non-Chooser	Same Race	M	57.14	16780	.86	14
Non-Chooser	Same Race	O1	63.33	15169	.50	6
Non-Chooser	Same Race	O2	76.00	23218	.80	5
Non-Chooser	Same Race	O3	NA	NA	NA	NA
Non-Chooser	Same Race	O4	NA	NA	NA	NA
Non-Chooser	Same Race	O5	NA	NA	NA	NA
Non-Chooser	Same Race	U1	80.00	7126	1.00	1
Non-Chooser	Same Race	U2	NA	NA	NA	NA
Non-Chooser	Cross Race	F	54.64	13396	.64	194
Non-Chooser	Cross Race	M	65.00	21447	.50	8
Non-Chooser	Cross Race	O1	87.50	16459	.88	8
Non-Chooser	Cross Race	O2	56.00	12081	1.00	5
Non-Chooser	Cross Race	O3	66.67	21546	.33	3
Non-Chooser	Cross Race	O4	NA	NA	NA	NA
Non-Chooser	Cross Race	O5	NA	NA	NA	NA
Non-Chooser	Cross Race	U1	86.67	76717	1.00	3
Non-Chooser	Cross Race	U2	NA	NA	NA	NA

Table 2. Mean confidence, decision time, and accuracy for chooser and non-chooser responses that are accompanied by justifications that have been categorized as either familiarity (F), observable feature (O), unobservable feature (U), or mixed (M). Categories involving features

include a number to indicate the number of features mentioned (e.g., O3 refers to three observable features).

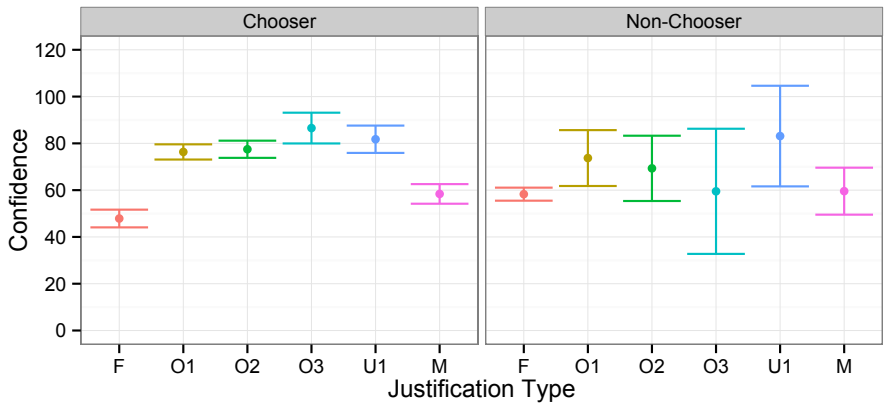
Three key patterns are apparent in Table 2. The first is the lack of a strong effect of lineup race. For example, within chooser lineup responses, there is very little difference among the counts across the different justification types for same- vs. cross-race lineups. To investigate this, I fit a generalized linear mixed model predicting lineup race (binary) from the fixed effect of Justification Type;[4] the random effect consisted of a participant intercept.[5] A likelihood ratio test showed the main effect was non-significant, $\chi^2(5) = 9.83$, $p = .08$. However, as shown in Figure 4 below, three observable justifications were more strongly associated with cross-race lineups, although clearly this is a only a trend, mainly because O3 statements were rare overall.



[4] Categories O4, O5, and U1 were excluded from all models within this section because there were very few observations, as shown in Table 2. Also, for this model in particular, including a fixed effect interaction of Justification Type with Lineup Decision led to non-convergence due to rank deficiency (in Table 2, note that there are no O3, non-chooser, same race responses).
[5] For all models reported within this section, random effects were settled upon after trying various combinations of main effects and interactions of all factors not included within the fixed effects and then selecting the model with the lowest AIC.

Figure 4. The association between lineup race and justification type. As discussed in footnote 4, categories O4, O5, and U1 were excluded from this regression because there were very few observations, as shown in Table 2. Point estimates reflect the probability of making a same-race response for each justification type, including a 95% confidence interval. The stacked bars provide a visual indicator of proportion of same- versus cross-race responses.

The second pattern involves the degree of confidence participants assigned to responses that are accompanied by the different types of justifications. This was assessed using a linear mixed model predicting confidence from the fixed effect interaction of Justification Type and Lineup Response; the random effect consisted of a participant intercept and a slope for Accuracy.[6] A follow-up likelihood ratio test showed a significant main effect of Justification Type, $\chi^2(5) = 82.89$, $p < .0001$, which was qualified by a significant interaction between Justification Type and Lineup Response, $\chi^2(5) = 16.89$, $p < .01$. I focus on the higher order interaction to which the main effect is marginal.



[6] I also fit several cumulative link mixed models treating Confidence as an ordered factor (mainly after noting the 95% CI for U1 extending past 100 confidence), but results were similar to the LMMs and thus I report the latter, which are simpler to interpret and plot.

Figure 5. Confidence in chooser and non-chooser responses that are based on the different kinds of justifications. As discussed in footnote 4, categories O4, O5, and U1 were excluded from this analysis because there were very few observations, as shown in Table 2. Error bars indicate a 95% confidence interval (see footnote 6 regarding the 95% CI for U1 non-choosers, which is based on very few observations [3, as shown in Table 2]).

As shown in Figure 5 above, when choosing, participants were comparably confident for observable versus unobservable feature justifications (i.e., O1, O2, O3 vs. U1), $\chi^2(1, N = 256) = 0.27$, $p = .60$. However, these featural justifications were associated with an average confidence rating that was 27.40 points higher than the average confidence rating assigned to familiarity and mixed responses (i.e., O1, O2, O3, U1 vs. F, M), $\chi^2(1, N = 256) = 228.07$, $p < .0001$. Overall, chooser familiarity justifications were associated with the lowest confidence, even compared to the second lowest category, i.e., mixed statements (i.e., F vs. M), $\chi^2(1, N = 256) = 16.73$, $p < .0001$.[7]

For non-choosers, there were minimal differences between the confidence ratings associated with observable versus unobservable feature justifications (i.e., O1, O2, O3 vs. U1), $\chi^2(1, N = 256) = 1.69$, $p = .19$. And, non-chooser featural justifications were associated with higher confidence than familiarity and mixed responses (i.e., O1, O2, O3, U1 vs. F, M), albeit to a lesser degree of 12.49 points on the confidence scale, $\chi^2(1, N = 256) = 4.87$, $p = .03$. However, unlike for choosers, the mean confidence ratings assigned to familiarity and mixed justifications were comparable for non-chooser responses (i.e., F vs. M), $\chi^2(1, N = 256) = 0.06$, $p = .81$.

[7] An additional contrast also showed that overall confidence for chooser versus non-chooser responses was comparable, such that there was no significant difference, $\chi^2(1) = 1.33$, $p = .25$.

The third key pattern in Table 2 is that familiarity-based justifications are more frequent for non-chooser than chooser responses. In other words, individuals are much more likely to refer to the absence of familiarity (e.g., "none of them are familiar") to justify a confidence rating for a response of "not present" than to refer to the presence of familiarity (e.g., "he is familiar") to justify an lineup identification. To evaluate this, I fit a generalized linear mixed model of binary Chooser Type predicted by Justification Type; the random effect consisted of a participant intercept and slopes for Accuracy, Decision Time, and Lineup Race. A likelihood ratio test confirmed the significant main effect, $\chi^2(5) = 667.67$, $p < .0001$. As shown in Figure 6 below, familiarity statements were substantially more likely to occur in the context of a non-chooser response than a chooser response as compared to all other justification types (i.e., F vs. O1, O2, O3, U1, M), $\chi^2(1, N = 256) = 146.62$, $p < .0001$. However, mixed statements were also slightly more likely to occur in the context of a non-chooser response than observable and unobservable justifications (i.e., M vs. O1, O2, O3, U1), the latter of which were essentially at ceiling (i.e., observable and unobservable statements were almost always associated with chooser responses and almost never with non-chooser responses), $\chi^2(1, N = 256) = 7.92$, $p < .01$.

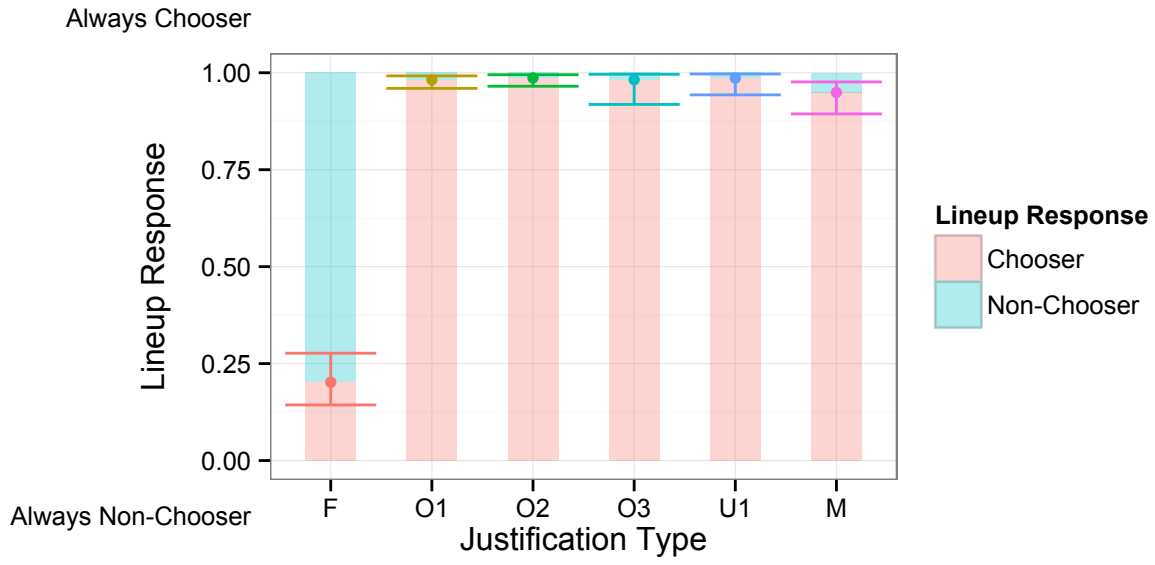


Figure 6. The probability of making a chooser response based on justification type. As discussed in footnote 4, categories O4, O5, and U1 were excluded from this analysis because there were very few observations, as shown in Table 2. Point estimates reflect the probability of making a chooser response for a given justification type, including a 95% confidence interval. The stacked bars provide a visual indicator of proportional choosing rates.

Based on the raw data shown in Table 2, 89.38% of all non-chooser justifications were based on familiarity; by contrast, within chooser responses, familiarity justifications occurred only 22.21% of the time. This disproportionate use of the familiarity justification within non-chooser responses necessitates focusing the analysis of accuracy on chooser responses, since non-chooser responses are so dominated by familiarity statements. Additionally, focusing on chooser responses is warranted given that the confidence/accuracy relationship is stronger for chooser than non-chooser responses (e.g., Brewer & Wells, 2006; Sauerland et al., 2012) and because chooser responses have far greater relevance from an applied perspective (i.e., for criminal proceedings).

Identification Accuracy. Based on the previous analyses, the analysis of identification accuracy included only those justification-types that were well represented (see Table 2 for a breakdown). Specifically, I excluded responses that were based on either the O4, O5, and U2 justifications because of their infrequency of use. Lastly, because mixed justifications represent combinations from several of the other categories, these were not directly comparable to homogenous justifications types and were thus also excluded from the analysis of accuracy.[8] Of the 1621 statements analyzed in the previous section, this left 1276. In addition, as discussed in

[8] A future study will investigate the interactions of different justification types in a controlled manner, as discussed in the discussion section of Experiment 2.

the previous section, the analysis of accuracy by justification type focuses on chooser responses, which represent 799 of these 1276 statements.

The analysis of accuracy also investigated decision time. Responses that took longer than one minute were removed from the analysis, and these consisted of 8 responses of the 799 (0.01%), leaving 791 chooser responses in the final sample. Again, decision times were also log transformed prior to model fitting.

I conducted a logistic regression on accuracy using a linear mixed model with the following fixed effects: the full interaction of Justification Type (F, O1, O2, O3, U1), Lineup Race (Same Race vs. Cross Race),[9] Confidence, and (log-transformed) Decision Time. The random effects included a random intercept within participant, the variance of which was normally distributed.[10] Continuous predictors (i.e., Confidence and Decision Time) were centered and scaled prior to running the model.

I used a multi-model selection approach (Burnham & Anderson, 2002) to find the best possible model of accuracy. Table 3 below shows a subset of the models tested including their AICs, with the best model highlighted in bold.[11]

---

[9] Lineup Race was included as a fixed effect rather than a random effect because I was interested in how it might interact with other fixed effects. Within each lineup race, we had a nested set of stimuli (e.g., within white lineups, there were six exemplars). Analysis of a model that included stimulus as a random effect term nested within lineup race (but no fixed effect for lineup race) showed no strong variance based on lineup race (see Appendix A for a plot of this random effect).

[10] I also fit a model with a slope of Lineup Race within participant, but this model had a higher AIC by four points (899.32 vs. 895.32), so I retained the simpler model. More complex random effect structures (e.g., a slope for JustificationType) could not be fit due to the variable number of observations within participant (e.g., a given participant may never make a familiarity response or never choose across all 12 lineups).

[11] Because I tested many models, it is impractical to provide an exhaustive list. Rather than trying every combination, I focused on a guided selection using Wald ANOVAs to find the optimal model (i.e., the best compromise between complexity and parsimony, based on AIC).

Model Formula	df	AIC
JustificationType * DecisionTime * Confidence * LineupRace + (1 + LineupRace | Participant)	43	899.92
JustificationType * DecisionTime * Confidence * LineupRace + (1 | Participant)	41	895.92
(JustificationType + DecisionTime + Confidence + LineupRace)^3 + (1 | Participant)	37	892.66
(JustificationType + DecisionTime + Confidence + LineupRace)^2 + (1 | Participant)	24	881.01
JustificationType + DecisionTime + Confidence + LineupRace + (1 | Participant)	9	876.82
(JustificationType + DecisionTime + Confidence + LineupRace)^2 - LineupRace:JustificationType - LineupRace:Confidence + (1 | Participant)	**19**	**872.41**
(JustificationType + DecisionTime + Confidence + LineupRace)^2 - LineupRace:JustificationType - LineupRace:Confidence + (1 + LineupRace | Participant)	21	876.28
(JustificationType + DecisionTime + Confidence + LineupRace)^2 - LineupRace:JustificationType - LineupRace:Confidence - LineupRace:DecisionTime + (1 | Participant)	18	873.08

Table 3. Model formulae, degrees of freedom, and AICs for the model comparison conducted in Experiment 1 to find the best predictors of accuracy. In Wilkinson-Rogers notation, an asterisk (i.e., *) represents an interaction that includes all marginal terms (i.e., sub-interactions, notated with a colon [:]); raising summed factors to a power equates to the inclusion of all interactions and marginal sub-interactions through that power (e.g., the fourth line in the table represents all two-way interactions of the four factors in parentheses, along with all main effects). Mixed effects terms are always included as the rightmost term within the model (e.g., [1 + LineupRace | Participant] indicates an intercept term [1] and a slope for LineupRace within each Participant).

As shown in Table 3 above, the best model based on AIC included all main effects and several two-way interactions as fixed effects and a participant intercept as the random effect. I used a likelihood ratio test (LRT) to assess the significance of this model's terms. Results of this LRT are summarized in Table 4 below.

The approach involved narrowing down by different interaction degrees and adding and removing factors as needed while being mindful of the marginality principle.

Effect	χ^2	df	p
LineupRace	5.84	1	.02
DecisionTime	4.82	1	.03
Confidence	59.16	1	<.0001
JustificationType	9.92	4	.04
LineupRace:DecisionTime	2.68	1	.10
DecisionTime:Confidence	5.61	1	.02
DecisionTime:JustificationType	*9.26*	*4*	*.06*
Confidence:JustificationType	*9.42*	*4*	*.05*

Table 4. The results of a likelihood ratio test on the best model from the analysis of accuracy in Experiment 1. Note that the final two model terms highlighted in italics, i.e., the interaction of Decision Time and Justification Type and the interaction of Confidence and Justification Type are only marginally significant, with *p* values of .0549 and .0515, respectively.

Results from this model are consistent with the analysis summarized in the earlier replication section: the same terms show significance (e.g., the interaction of Decision Time and Confidence and the main effect of Lineup Race). Focusing on the highest order terms, the following are significant based on the LRT: the interaction of Decision Time and Confidence, $\chi^2(1) = 5.61$, $p = .02$, the main effect of Lineup Race, $\chi^2(1) = 5.84$, $p = .02$, and the main effect of Justification Type, $\chi^2(4) = 9.92$, $p = .04$. The interaction of Confidence and Justification Type and Decision Time and Justification Type were only marginally significant, as described in the table caption, although the former is discussed within the text while the latter is presented in Appendix A.[12]

[12] A plot and description of the second marginally significant interaction between Justification Type and Decision Time was originally included within the text as well, but it was moved to Appendix A mainly due to the fact that meaningful differences were difficult to discern.

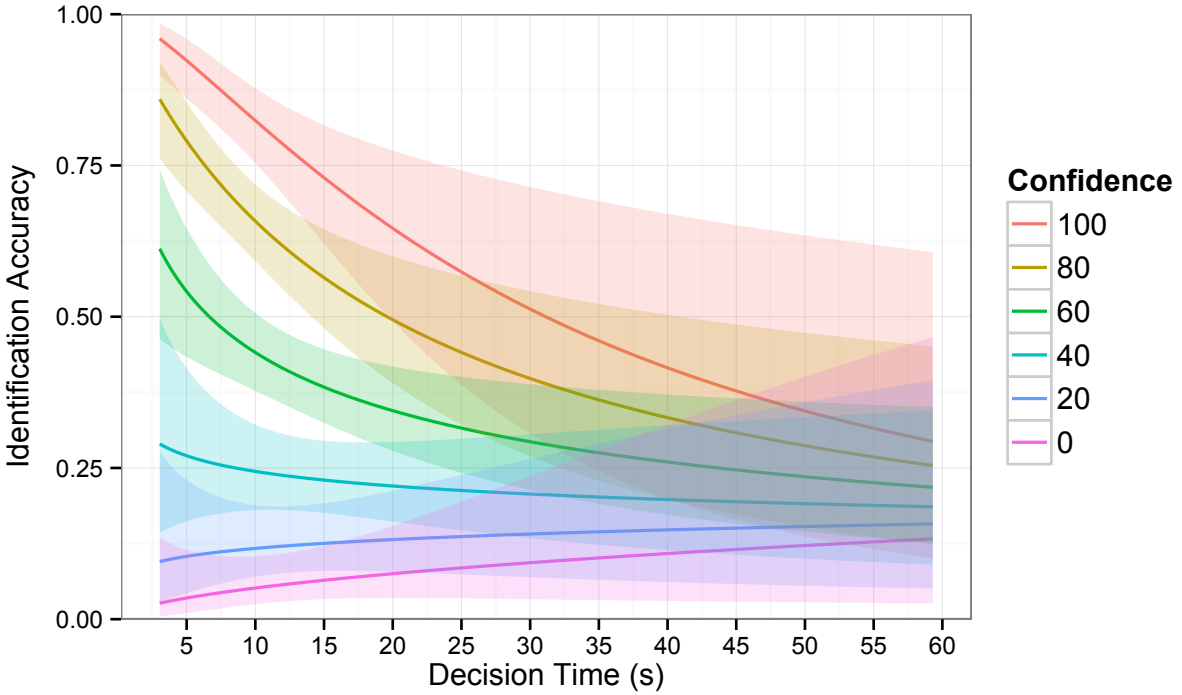


Figure 7. The interaction between decision time and confidence in the best model of chooser identification accuracy in Experiment 1 that included justification type coding. Consistent with expectations, higher confidence is associated with higher accuracy, particularly for the responses made most quickly. Error shading represents a 95% confidence interval.

Figure 7 above shows the interaction between Decision Time and Confidence. Exactly as expected and consistent with the analysis in the previous section involving the larger set of data (i.e., these data are a sub-sample of those analyzed in the replication section), confidence is associated with decision time such that higher confidence leads to higher identification accuracy, particularly when responses are made quickly. Similarly, as shown in Figure 8 below, the main effect of Lineup Race was consistent with prior findings: accuracy was higher for same-race lineups than for cross-race lineups.

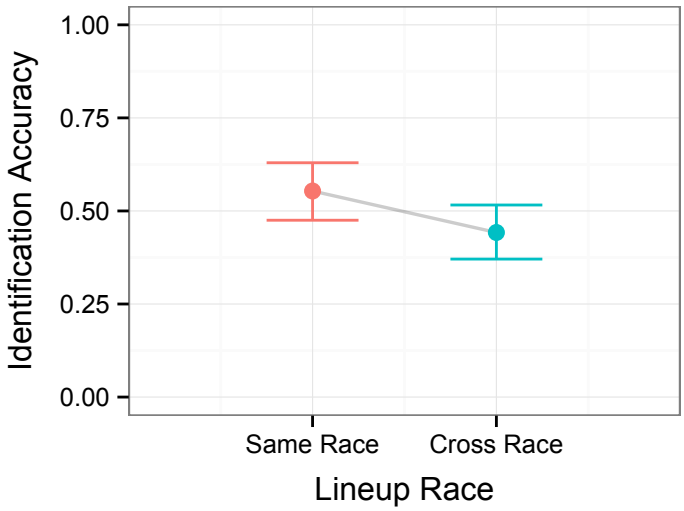


Figure 8. The main effect of lineup race in the best model of chooser identification accuracy in Experiment 1 that included justification type. Consistent with expectations, accuracy is higher for same-race lineups than for cross-race lineups. Error bars represent a 95% confidence interval.

Of greater interest is the effect of Justification Type on identification accuracy. As shown in Figure 9 below, familiarity-based identifications were associated with the lowest identification accuracy overall, whereas identifications based on either observable features (particularly those that mentioned two features) or unobservable features were associated with higher identification accuracy, as verified by a follow-up contrast, $\chi^2(1, N = 241) = 6.57$, $p = .01$. An additional follow-up contrast showed no significant difference in identification accuracy between observable features (i.e., across feature counts) or unobservable features with a single feature, $\chi^2(1, N = 241) = 0.50$, $p = .48$. By contrast, observable feature justifications involving a single feature were associated with higher identification accuracy than familiarity-based justifications, $\chi^2(1, N = 241) = 3.86$, $p < .05$, and unobservable feature statements with a single

feature were also associated with higher identification accuracy than familiarity, $\chi^2(1, N = 241) = 4.30$, $p = .04$. Overall, familiarity statements were associated with lower identification accuracy than feature-based statements regardless of whether the latter were observable or unobservable.

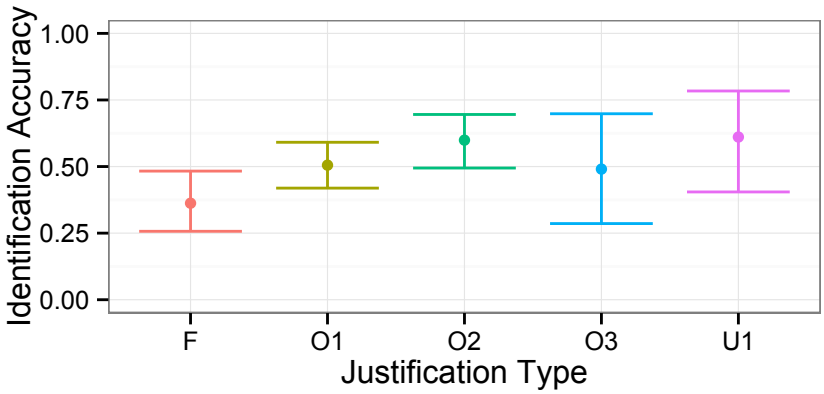


Figure 9. The main effect of justification type on chooser identification accuracy in Experiment 1. F refers to familiarity, O refers to observable feature, and U refers to unobservable feature statements; numbers next to O and U refer to the number of features mentioned. Comparatively, accuracy for familiarity statements is lower than all other types of justifications, but particularly lower than an observable feature statement that mentions two features. Error bars indicate a 95% confidence interval.

The main effect of Justification Type must be considered relative to the marginal interaction between Justification Type and Confidence on identification accuracy. Regarding the former, Figure 10 below shows the trajectory of different justification types across the confidence scale. As is somewhat apparent from the figure, the greatest separation occurs for familiarity responses at higher levels of confidence (i.e., 60 and above). For all justification types besides familiarity, identification accuracy increases consistently with confidence; familiarity justifications, however, show a weaker confidence/accuracy relationship at higher

levels of confidence: specifically, familiarity statements made with high confidence are more prone to overconfidence.

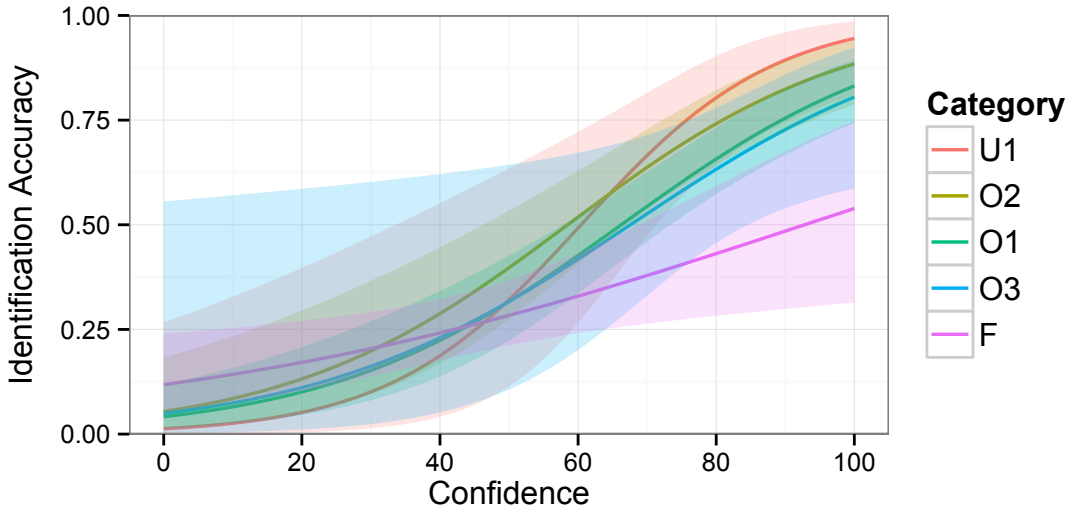


Figure 10. The marginally significant interaction between confidence and justification type in the best model of chooser identification accuracy in Experiment 1. F refers to familiarity, O refers to observable feature, and U refers to unobservable feature statements; numbers next to O and U refer to the number of features mentioned. Familiarity responses trend toward lower identification accuracy at higher confidence levels (i.e., 60 and up) in contrast to feature-based responses, which show a stronger confidence/accuracy relationship across the confidence scale. Error bars represent a 95% confidence interval.

Discussion

Prior studies have shown a stronger relationship between confidence, decision time, and identification accuracy for choosers than for non-choosers (e.g., Brewer & Wells, 2006; Dodson & Dobolyi, under review; Sauerland et al., 2009), and I replicated this pattern in Experiment 1. Beyond replication, the primary goals of this experiment were 1) to assess whether or not

different justification types are associated with different types of responses (e.g., are people more likely to mention familiarity when saying "not present"?) and 2) to determine if different justification types are more predictive of accuracy. Regarding the former issue, Experiment 1 showed that individuals are much more likely to use a familiarity response than any of the other pre-defined justification types when not choosing from a lineup. By contrast, for choosers, there was greater variability among the different types of justification statements including familiarity, observable features, and unobservable features. Given that the bulk of the non-chooser responses involved familiarity and the confidence/accuracy relationship has consistently been shown to be weaker for non-chooser responses (e.g., Brewer & Wells, 2006), this suggests that familiarity statements are also associated with a weaker confidence accuracy relationship, at least when individuals say "not present."

For chooser lineups, I fit a model that predicted accuracy based on the type of justification participants mentioned in combination with other postdictors of identification accuracy that are known to be informative, such as confidence and decision time (e.g., Brewer & Wells, 2006; Sauerland & Sporer, 2009; Weber et al., 2004). This analysis identified a clear pattern: when choosing involves familiarity, identifications are less likely to be accurate than when choosing involves either an observable or unobservable feature. This finding is consistent with Reinitz et al. (2011, 2012), who also found higher identification accuracy for memories involving a feature than for familiarity. However, Experiment 1 also showed that familiarity-based responses produce a greater dissociation between confidence and accuracy at higher levels of confidence (i.e., 60 and up) compared to feature-based responses. For familiarity-based responses, accuracy was reduced at higher levels of confidence; by contrast, for other justification types, the confidence/accuracy relationship was stronger across the entire

confidence scale. This latter finding conflicts with Reinitz et al. (2012), who found a greater degree of overconfidence for feature-based memories than for familiarity-based memories when controlling for the overall level of confidence in a non-eyewitness facial identification task.

Dodson and Dobolyi (2015) showed that others are more likely to perceive the identical confidence statement that is justified with reference to an observable feature as connoting a lesser degree of confidence, as compared to statements that are either justified on the basis of an unobservable feature or contain no justification (i.e., confidence statement only). Therefore, it will be interesting to see what happens when I show the lineups and justification statements generated in Experiment 1 to a new set of participants in Experiment 2. This second experiment is crucial for understanding the relationship between intended confidence, as measured in Experiment 1, and perceived confidence, which will be measured in Experiment 2, and whether or not this relationship varies across the different justification types.

For example, how accurately can individuals perceive the intended numeric confidence of statements that are accompanied by either a familiarity or observable or unobservable justification? This is an important question considering that Experiment 1 showed that: 1) familiarity-based responses are less accurate than feature-based responses when individuals choose a face from a lineup and 2) that the confidence/accuracy is weaker for familiarity-based statements at higher levels of confidence.

Experiment 2

When an eyewitness identifies someone from a lineup and states, “I’m pretty sure it’s him,” how do we know that police, jurors and others will interpret this expression of confidence in the way that was intended by the eyewitness? Perhaps an eyewitness means he is only 50% sure but others might think he is 80% sure, or vice versa. The objective of this experiment, then,

is to examine how accurately people can interpret eyewitness expressions of certainty about a lineup identification.

The problem of potentially misinterpreting eyewitness expressions of confidence is greater with verbal (e.g., "I'm fairly sure it's him") than with numeric expressions (e.g., "I'm 80% sure it's him") because verbal expressions tend to be less clear than numeric expressions (e.g., Beyth-Marom, 1982; Budescu & Wallsten, 1985; Gurmankin, Baron, & Armstrong, 2004). There are no published data about the frequency with which eyewitness statements of confidence are expressed primarily numerically (e.g., "I'm 80% sure it's him"), primarily verbally (e.g., "I'm pretty certain it's him") or as a mixture of both. However, much research using a variety of different paradigms shows that, when given a choice, individuals generally prefer to express their confidence with words (e.g., "fairly certain") rather than with numbers (e.g., "75% sure"; e.g., Brun & Teigen, 1988; Budescu et al., 2003; Erev & Cohen, 1990). For example, Wallsten et al. (1993) observed that 65% of their participants expressed uncertainty verbally rather than numerically. So, this apparent preference for verbal expressions of confidence suggests that misinterpretations of eyewitness confidence could be a frequent occurrence.

Dodson and Dobolyi (2015) showed that individuals are more likely underestimate high confidence statements about a lineup identification, e.g., "I'm positive it's him," when this confidence statement is accompanied by a justification that refers to a visible feature about the accused (e.g., "I remember his chin") than when it is accompanied by either an unobservable justification or no justification (i.e., confidence statement only). But, there are two critical unknowns. First, does this featural justification effect generalize to situations when highly confident eyewitnesses refer to multiple features to justify their level of confidence in a lineup identification? Second, across a range of confidence-levels, how does the particular content of

an eyewitness's justification influence how accurately others understand the intended meaning of the confidence-statement?

Experiment 2 uses a task nearly identical to Experiment 1 in Dodson and Dobolyi (2015). Participants will encounter a series of 10 accurate chooser lineups from the sample generated in Experiment 1. Half of these lineups will be same-race and the other half cross-race; within each race, each of the five lineups will represent one of the following justification types: 1) confidence only, 2) familiarity, 3) observable feature, 4) multiple observable features, and 5) unobservable feature. Confidence only phrases were generated using statements from the other four categories, but with the justification removed: for example, the unobservable statement "Pretty certain. I remember this picture because he looks like one of my friends" appeared as the confidence only statement "Pretty certain." The purpose of confidence only is to serve as a control condition to measure perceived confidence in the absence of any additional justification.

Participants will translate these justifications onto the same numeric scale as in Experiment 1, thereby producing a measure of perceived confidence (as opposed to intended confidence). Thus, I will then be able to compare perceived confidence to intended confidence from the prior experiment and see how it varies across different justification types.

Method

Participants. Participants were 300 white individuals between the ages of 18 and 40 (M = 28.85, SD = 5.25, range = 18.0 – 39.7; 50.33% female) who were collected via Amazon Mechanical Turk in exchange for payment. An a-priori power analysis conducted using G*Power (Faul et al., 2007) using $\alpha = .05$ showed that we would have over 99% power to detect medium-sized effects (Cohen's $f = .25$; Cohen, 1988) in the context of a repeated measures

ANOVA with 300 participants. All participants gave consent and completed a brief demographic questionnaire.

Design. The experiment was entirely within-subjects, consisting of a 5 (Justification Type: Confidence Only, Familiarity, Observable Feature, Multiple Observable Features, and Unobservable Feature) by 2 (Lineup Race: Same Race vs. Cross Race) design.

Materials. The experiment was conducted using a custom browser-based framework built using PHP, jQuery/JavaScript, MySQL, and HTML. The entire experiment is available online at http://www.dodsonlab.com/studies/faces_e2/ and via a debug link that allows the full task to be seen in a single page: http://dodsonlab.com/studies/faces_e2/?debug.

Stimuli for the experiment consisted of five black and five white lineups from the preceding study that were also used in Dodson and Dobolyi (2015) and Dobolyi and Dodson (2013).[13] Moreover, lineup decisions and justification statements were taken from the sample produced by participants in Experiment 1, producing a total set of stimuli that consisted of 267 lineups with associated confidence expressions and justification statements. The stimuli were selected under several constraints with the goal of producing a variety of exemplars within each of the five categories (e.g., Familiarity, Observable Feature, etc.). Specifically, the criteria for selecting stimuli from Experiment 1 included: 1) inter-rater agreement in justification type coding; 2) the eyewitness in Experiment 1 had an intended confidence level of 60, 80, or 100 for the confidence statement; 3) the eyewitness had an identification decision time of 60s or less; and 4) the eyewitness correctly identified the target within the lineup (i.e., a correct chooser decision). The Confidence Only condition consisted of the same lineups and statements that were used in the other conditions, but the justifications were omitted. For example, the

[13] One black and one white lineup were dropped from Experiment 2 because each participant needed to see 10 lineups to represent all levels of the 5x2 within-subjects design.

Familiarity phrase "Sort of certain. He looks familiar" appeared in the Confidence Only condition as "Sort of certain."

Procedure. The procedure was essentially identical to Dodson and Dobolyi (2015). Participants were instructed to pretend they were police officers and told that they would view a series of eyewitness lineups. They were informed that lineups might involve either the identification of a suspect or a "Not Present" response (in fact, all lineups shown involved choosing a face within the lineup, but the "Not Present" response still appeared on the lineup screen); moreover, they were told that each lineup would include a written expression of certainty for the eyewitness's decision. The task was "to translate the written expressions onto a numeric confidence scale" using a six-point confidence scale that ranged from 0 (Not at All Certain) to 100 (Completely Certain) in increments of 20, as shown in Figure 11 below.

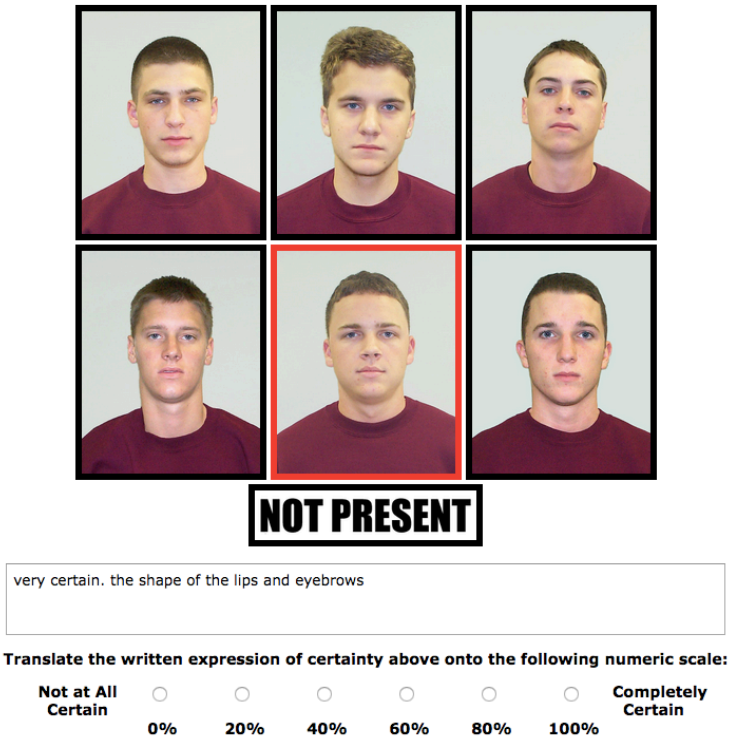


Figure 11. An example of the expression of certainty translation task from Experiment 2. The confidence expression of "very certain" is justified with two observable features: "the shape of the lips and eyebrows."

Participants viewed a practice lineup prior to the 10 critical lineups. The practice lineup consisted of six smiley faces of different colors, with a yellow smiley face highlighted as the eyewitness's selection. The eyewitness's expression of certainty for the practice lineup was always the following phrase: "I am completely certain this was the yellow smiley face I saw earlier. The color was yellow." Only participants who translated this expression of certainty to 100 were included in the final dataset.

Critical lineups were shown exactly as they were rated in Experiment 1: within the 2x3 array of faces, the faces were presented identically to how the Experiment 1 participant encountered them (e.g., in the example shown in Figure 11 above, the identified face appeared in the bottom center, and the foils appeared in those exact locations within the array). The order of the 10 critical lineups was randomized such that no more than three black or white lineups appeared in succession. Moreover, the experiment balanced how often each of the 267 lineup stimuli was shown across all participants.

After completing all 10 lineups, participants were asked to fill out a short demographics survey including questions about age, sex, and race. Finally, participants were thanked for their involvement and given a debriefing.

Results

Perceived Confidence. The primary goal of Experiment 2 is to understand how participants' ability to perceive the intended numeric confidence of an eyewitness's

identification is influenced by: 1) Justification Type (Confidence Only, Familiarity, Observable Feature, Multiple Observable Features, and Unobservable Feature); 2) Lineup Race (Same Race vs. Cross Race); and 3) Level of Intended Confidence (60 – 100).

To investigate this issue, I fit a series of six linear mixed models predicting Perceived Confidence from different combinations of the three factors of interest: Justification Type, Lineup Race, and Intended Confidence (the latter of which was centered and scaled prior to analysis). Random effects included combinations of Participant and Stimulus, the latter of which was nested within Lineup Race. Table 5 below includes a list of the six models tested, with the best model (i.e., the one with the lowest AIC) highlighted in bold.[14]

Model Formula	df	AIC
JustificationType * IntendedConfidence + (1 | Participant)	12	26964.79
JustificationType * IntendedConfidence.f + (1 | Participant)	17	26954.27
JustificationType * IntendedConfidence * LineupRace + (1 | Participant)	22	26949.16
JustificationType * IntendedConfidence + (1 + LineupRace | Participant)	14	26965.68
JustificationType * IntendedConfidence + (1 | Participant) + (1 | LineupRace / Stimulus)	**14**	**26507.20**
JustificationType * IntendedConfidence.f + (1 | Participant) + (1 | LineupRace / Stimulus)	19	26513.15

Table 5. Model formulae, degrees of freedom, and AICs for the model comparison conducted in Experiment 2 to find the best predictors of perceived confidence. In Wilkinson-Rogers notation, an asterisk (i.e., *) represents an interaction. Mixed effects terms are always included as the rightmost term within the model (see the caption for Table 3 in Experiment 1 for further details).

[14] Intended Confidence—consisting of 60s, 80s, and 100s—was modeled in two different ways: as a simple linear effect, IntendedConfidence, and as a factor, IntendedConfidence.f. More flexible methods like natural splines were unnecessary because Intended Confidence consisted of only 3 discrete points (60, 80, 100), making even a 2nd order natural spline redundant. Note that it was not possible to run a 3-way interaction of JustificationType, IntendedConfidence.f, and LineupRace because this combination was rank deficient: not all levels were represented because some decision types (e.g., familiarity justifications to cross-race lineups made with 100 confidence) were exceedingly rare. Lastly, based on the model comparisons and visual inspection, Intended Confidence was very clearly a 1st order linear effect.

See footnote 14 for a discussion of the different ways in which Intended Confidence was modeled.

The best model included the fixed effect interaction of Justification Type and Intended Confidence (modeled continuously) and random effects consisting of a Participant intercept as well as an intercept of Stimulus Number nested within Lineup Race. Random effects were generally normally distributed, although a handful of participants gave lower mean ratings of perceived confidence than others (see Appendix A for figures for all random effects).[15] As shown in Figure 12 below, Lineup Race did not show a strong effect, consistent with previous findings (e.g., Dodson & Dobolyi, 2015).

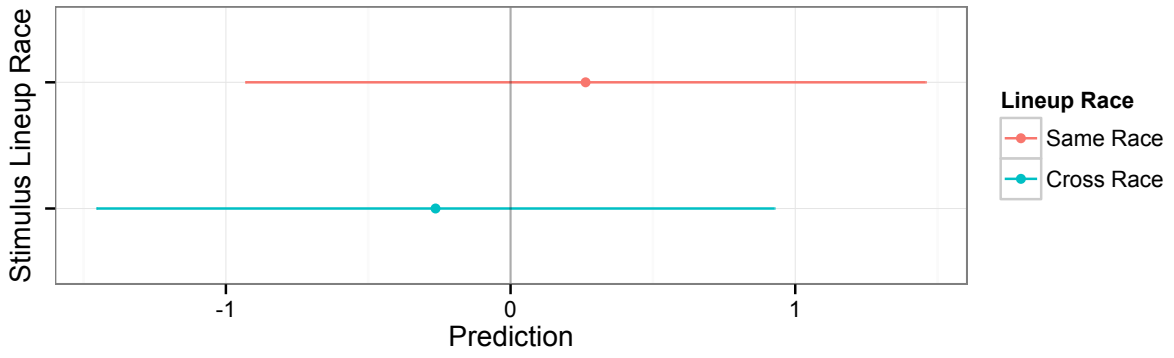


Figure 12. The random effect of the nesting factor of lineup race relative to stimulus in a model predicting perceived confidence. Error bars indicate a 95% prediction interval. The estimates

[15] A separate analysis was conducted on variability of perceived confidence that is not reported due to length concerns. However, several of the greatest outliers in the random effect of Participant on both perceived confidence and variability overlapped across the models (eight of 10), suggesting that these participants may not have completed the task correctly. However, after dropping these eight participants and refitting models, conducting additional likelihood ratio tests, and re-plotting, the results remained unchanged, so I have left these participants in the sample. An advantage of mixed modeling is that including the random effect of participants helps to control for the effect of these outliers relative to the fixed effects.

for same race lineups are slightly higher than for cross race lineups, but the prediction intervals for both overlap zero.

A likelihood ratio test conducted on the best model showed that both main effects of Justification Type and Intended Confidence were significant, but not the interaction. The absence of a significant interaction means that there is no variation in perceived confidence across Justification Types at different levels of Intended Confidence. Table 6 summarizes these results.

Effect	χ^2	**df**	**p**
JustificationType	17.02	4	<.01
IntendedConfidence	143.05	1	<.0001
JustificationType:IntendedConfidence	4.08	4	.40

Table 6. The results of a likelihood ratio test on the best model from the analysis of perceived confidence in Experiment 2.

The significant effect of Intended Confidence is illustrated in Figure 13 below: participants' perceived confidence ratings were consistently lower by roughly 20 points than the eyewitness's intended confidence. For example, when the eyewitness intended their confidence statement to mean 80% the participants perceived the statement as meaning 60% confidence. It is worth noting that this down-rating of perceived confidence relative to intended confidence accelerates over the scale, such that perceived confidence is 23.59 points lower at an intended confidence of 100, 19.51 points lower at 80, and 15.44 points lower at 60, respectively.

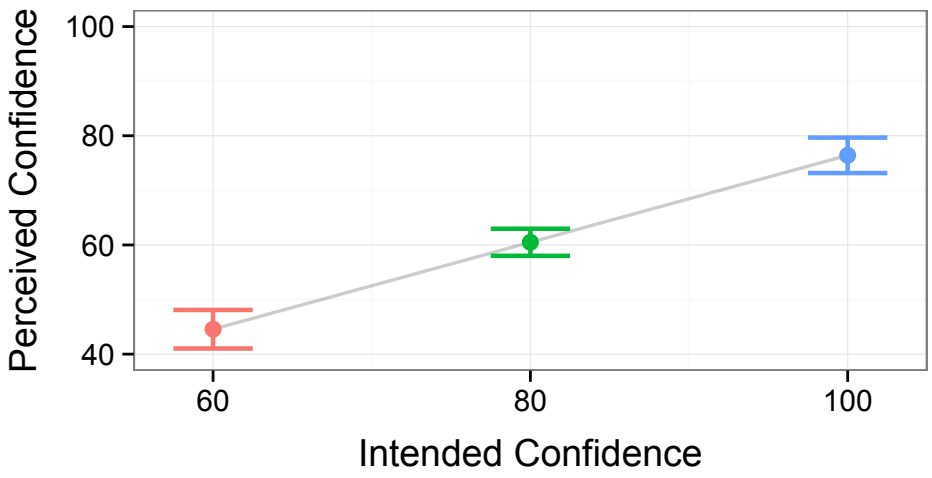


Figure 13. Perceived confidence at different levels of intended confidence in Experiment 2. Note that while intended confidence was inherently categorical (i.e., with ratings of 60, 80, and 100), it was modeled as a continuous predictor based on the results of a multi-model selection process. Perceived confidence is consistently lower than intended confidence, although particularly so at higher levels of intended confidence. Error bars indicate a 95% confidence interval.

The significant effect of Justification Type is shown in Figure 14 below. A series of four planned contrasts compared confidence alone to (a) familiarity (F), $\chi^2(1, N = 300) = 2.95$, $p = .09$; (b) a single observable feature (O1), $\chi^2(1, N = 300) = 0.34$, $p = .56$; (c) multiple observable features (Omany), $\chi^2(1, N = 300) = 10.27$, $p < .01$, and (d) a single unobservable feature (U1), $\chi^2(1, N = 300) = 1.89$, $p = .17$. Only the comparison between confidence only and multiple observable features was significant such that the latter was associated with perceived confidence 3.89 points higher on average. A fifth contrast compared a single observable feature to multiple observable features (i.e., O1 vs. Omany); this contrast was also significant, with multiple observable features producing a perceived confidence that was on average 4.65 points higher on

the perceived confidence scale than a single observable feature, as shown in the figure below, $\chi^2(1, N = 300) = 8.57, p < .01$.[16]

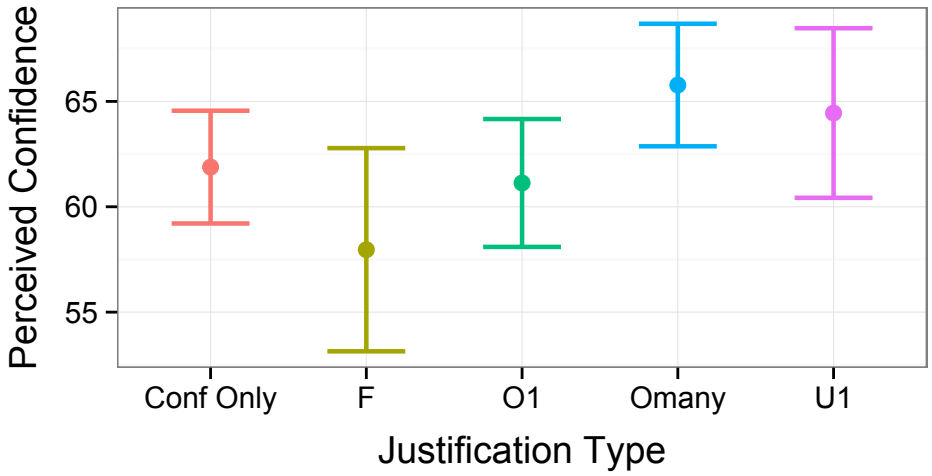


Figure 14. Perceived confidence across different justification types in Experiment 2. The categories are confidence only (Conf Only), familiarity (F), observable feature (O1), multiple observable features (Omany), and unobservable feature (U1). Perceived confidence is highest for multiple observable features and lowest for familiarity. Error bars indicate a 95% confidence interval.

Discussion

In terms of how others perceive an eyewitness's statement of confidence, Experiment 2 showed that participants consistently perceived eyewitnesses as less confident than was intended by the eyewitness. In addition, participants' perceptions were also influenced by how eyewitnesses justified their statement of confidence. Specifically, participants perceived confidence statements as indicating a higher value when the statement included a justification

[16] Out of curiosity, a sixth unplanned contrast showed that a single unobservable feature produced a perceived confidence 6.49 points higher than familiarity, $\chi^2(1, N = 300) = 5.18, p = .02$, but this contrast should be considered only in post-hoc terms.

that referred to multiple observable features than a single observable feature. By contrast, the other four categories—confidence alone, familiarity, single observable feature, and a single unobservable feature—showed no significant differences in the overall level of perceived confidence.

Overall, these results are surprising because they are different from what was previously shown in Dodson and Dobolyi (2015). Specifically, based on their experiments, I expected to find that perceived confidence would be comparable for confidence alone and unobservable feature statements, but significantly lower for observable feature statements, particularly at high confidence.

There is however an important difference between the design of the present experiment and that used by Dodson and Dobolyi (2015) that may explain this failure to replicate. Specifically, they used a between-subjects design in which justification type was varied between subjects (i.e., some participants saw confidence statements alone and others saw confidence statements with justifications); by contrast, the present experiment used a within-subjects design such that participants saw every justification type as well as confidence alone. The original proposal for this experiment also used a between-subjects design, but because a within-subjects design was potentially more powerful, I opted to change it without realizing the influence this decision choice would have on the outcome.

Thus, Experiment 2 may have inadvertently identified an important boundary condition regarding how others perceive confidence based on different types of justifications: specifically, a within-subjects design allows for comparative judgments that are relative in nature. For example, if a participant sees the confidence only expression “I’m very sure” after just having encountered “I’m very sure. I remember his chin,” it is likely that the perceived confidence of

the former would be reduced because it contains less information overall (i.e., no justification). This is very different from a design where participants are comparing only confidence statements in isolation to one another (e.g., "I'm very sure" versus "I'm pretty sure") or a series of justifications that all include an observable feature (e.g., "I'm very sure. I remember his chin" vs. "I'm pretty sure. I remember his nose"). In other words, when participants encounter multiple justifications types in a within-subjects design, it is possible to make relative comparisons based on the content of these justifications such that a particular justification type becomes the absolute baseline reference point; by contrast, when justification type only varies between subjects, then no such comparative evaluations can be made across justification type since only a single justification type is represented.

Thus, the results of Experiment 2 suggest an important follow-up study to compare how justifications are perceived depending on whether they vary within-subjects or between subjects. Such a study could reuse the materials from the present experiment and simply present the materials in either a between-subjects or a within-subjects manner.

This follow-up experiment would then answer an important question regarding how perceived confidence varies depending on the way in which justifications are presented: do people make comparative evaluations across justification types when interpreting the confidence of others and does this have real world consequences relative to cases with multiple eyewitnesses? For example, if one eyewitness says, "I'm very sure. I remember his chin" (i.e., observable feature) and another says, "I'm very sure. He looks like a friend of mine" (i.e., unobservable feature) would these two statements interact with one another in contrast to when both statements involved the same justification type? This question is extremely relevant,

especially considering that the next experiment (i.e., Experiment 3) varies justification type within-subjects to evaluate differences in behavior.

Experiment 3

Experiment 2 focused on understanding the perceived meaning of eyewitness expressions of confidence. A crucial next step is examining behavior. Police often need to seek corroborating evidence to support eyewitness testimony, and they particularly need to do so when eyewitnesses are *perceived* as being not very confident in their identification decision. However, police resources are finite and so allocating them to search for corroborating evidence for a crime means that fewer resources remain for other tasks. So, this experiment examines the prediction that different types of justifications of confidence will cause changes in behavior.

Participants in this experiment pretended to be police investigators with the task of assigning department resources across a series of case files. They were told to close as many cases as possible by assigning the highest priority to cases that represent the strongest evidence and the lowest to cases that represent the least evidence. Cases consisted of lineup identifications, and each case included a verbal expression of certainty with 1) either high (e.g., "I'm completely certain") or moderate (e.g., "I'm pretty sure") statement strength; and 2) varying justification types (i.e., confidence alone, unobservable justification, or observable justification).[17]

In regards to statement strength, assuming perceived confidence influences behavior then participants should assign highly confident statements a higher priority than moderately

[17] Note that familiarity justifications are not included in this experiment due to the number of available fair lineups: within the six white lineups that have previously been vetted for fairness, a 2x3 combination of statement strength by justification type uses all six lineups.

confident statements. In other words, if participants wish to close as many cases as possible, they will prioritize cases associated with the highest levels of perceived confidence.

More important is the issue of how participants will prioritize cases that contain confidence statements referring to different types of justifications. One possibility is that Experiment 3 will show a *featural justification effect* (Dodson & Dobolyi, 2015): cases involving observable featural information will receive a lower priority ranking than will cases involving unobservable featural information because observable featural statements tend to be viewed skeptically by others – especially when paired with a highly confident statement. Thus, according to this account, cases involving observable justifications should require more follow-up investigation relative to those cases with an unobservable justification or no justification.

However, it is important to remember that thus far the *featural justification effect* has been observed using between-subjects designs. The current experiment used a within-subjects design similar to Experiment 2, and in this preceding experiment there were few significant differences across justification type relative to confidence statements alone.[18] In addition, the preceding experiment differed from the current one with respect to stimuli. Whereas Experiment 2 used the actual lineups, confidence statements, and justifications that were generated by participant-eyewitnesses in Experiment 1, the current experiment used variations of pre-generated statements from Dodson and Dobolyi (2015). Notably, unobservable justifications were defined slightly differently in the current experiment compared to the preceding ones; for example, whereas an unobservable feature statement in Experiment 2 tended to refer to something specific (e.g., "He looks like my cousin"), in the current experiment unobservable

[18] In Experiment 2, only multiple observable features significantly differed from confidence only; however, numerically unobservable statements were also associated with a higher mean level of perceived confidence.

justifications consisted of simplified statements (e.g., "I remember him"). In this respect, the current experiment is more similar to Dodson and Dobolyi (2015) overall, with the major exception of the use of a within-subjects design rather than a between-subjects design.

Method

Participants. Participants were 112 white individuals between the ages of 18 and 40 (M = 28.34, SD = 5.38, range = 18.5 – 39.8; 46.43% female) who were recruited via Amazon Mechanical Turk in exchange for payment. An a-priori power analysis conducted using G*Power (Faul et al., 2007) using $\alpha = .05$ showed that we would have over 99% power to detect medium-sized effects (Cohen's $f = .25$; Cohen, 1988) in the context of a repeated measures ANOVA with 100 participants. All participants gave consent and completed a brief demographic questionnaire.

Design. The experiment was entirely within-subjects consisting of a 2 (Statement Strength: High Confidence vs. Moderate Confidence) x 3 (Justification Type: Confidence Only, Unobservable, Observable Feature) design.

Materials. The experiment was conducted using a custom browser-based framework built using PHP, jQuery/JavaScript, MySQL, and HTML. The entire experiment is available online at http://www.dodsonlab.com/studies/faces_e3/ and via a debug link that allows the full task to be seen in a single page: http://dodsonlab.com/studies/faces_e3/?debug.

Statement Strength	Confidence Statement	Justification Phrase	Observable Feature Modifiers	Unobservable Modifiers
High	I am very certain	I remember ___	his nose	him
High	I am very sure	I recall ___	his eyes	him
High	I am absolutely confident	I recognize ___		
Moderate	I am mostly certain	I recollect ___		
Moderate	I am pretty sure			
Moderate	I am fairly confident			

Table 7. Confidence and justification statements for Experiment 3 for each of the six lineups. Note that there are only four justification phrases because only four of the six lineups included a justification (i.e., two were Confidence Only). These four justification phrases were randomly assigned across the Observable Feature and Unobservable modifiers to complete the justification phrases.

Stimuli for the experiment consisted of the six white lineups used in the preceding studies and in Dodson and Dobolyi (2015). Table 7 above shows the confidence statements, which were similar to those used in Experiment 2 and 3 from Dodson and Dobolyi (2015). As is clear in the table, the high and moderate confidence phrases were designed to be as similar as possible: for example, the high confidence statement "I am very certain" was paired with a moderate confidence statement "I am mostly certain" (i.e., the phrases were identical besides "very" vs. "mostly"). Furthermore, because four out of the six lineups included a justification phrase (i.e., except for the two Confidence Only conditions), four unique phrases were constructed to initiate a justification phrase (e.g., "I recall ____"). These phrases were then completed with one of two observable featural justifications ("his nose" or "his eyes") or with "him" for both of the two unobservable justifications.

Procedure. The general procedure of Experiment 3 is similar to Experiment 2. Participants were instructed to pretend they were police investigators. They were assigned a series of six case files that each included an eyewitness lineup with a suspect highlighted with a red border. Moreover, each case file included an eyewitness' written expression of certainty regarding their identification. The participants' task was to assign each case a priority in an effort to close as many cases as possible based on the following instructions: "Police departments seek to close as many cases as possible, and your task will be to assign a priority to each case. Based on the eyewitness's testimony, you should give cases that represent the strongest evidence the highest priority and cases that represent the weakest evidence the lowest priority."

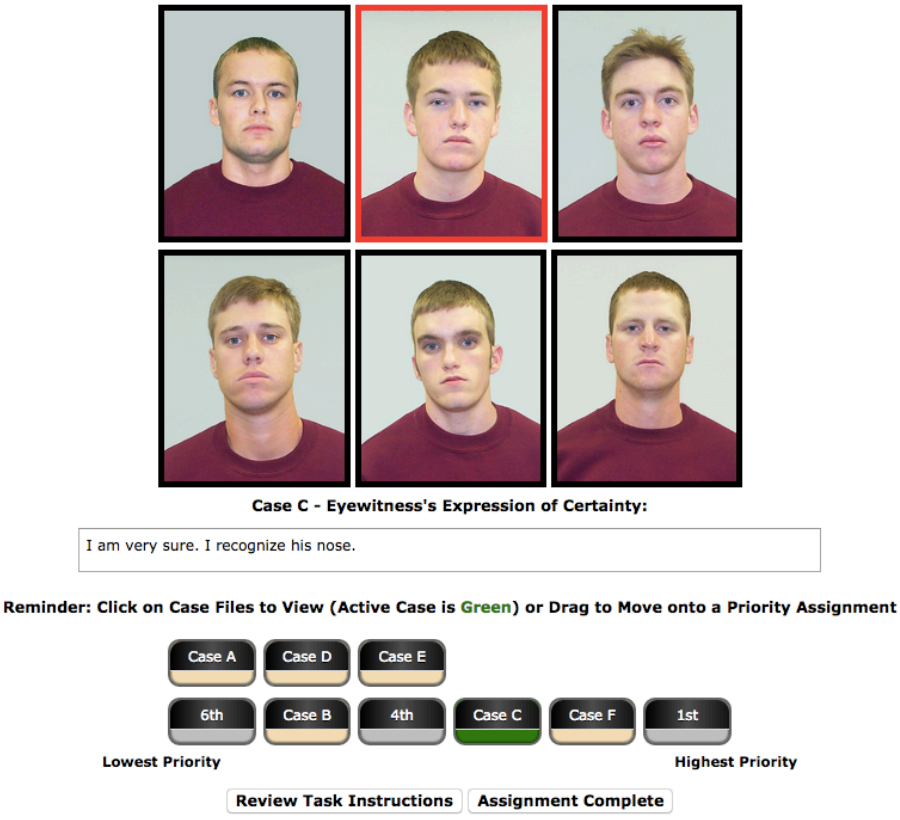


Figure 15. The task from Experiment 3. The button for Case C is green indicating that it is currently active, with the lineup and eyewitness's written expression of certainty in view.

Moreover, Case C has been given an assignment of 3rd priority based on it having been dragged into that slot on the priority assignment scale, which ranges from Lowest Priority (6th) on the left to Highest Priority (1st) on the right.

After reading the initial instructions, they proceeded to the main task, an example of which is shown in Figure 15 above. The main task began with a white screen and a series of buttons at the bottom of the screen, one for each case file, i.e., Case A – Case F. Clicking on a case file displayed the lineup and expression of certainty associated with that case, allowing participants to view and switch back and forth to compare case files. Once participants made a decision regarding how a case should be prioritized, they could then drag the case file buttons onto the priority ranking at the bottom of the screen. The priority assignment ranged from Lowest Priority (6th) to Highest Priority (1st), and cases could be freely rearranged among the six slots. At the very bottom of the screen a "Review Task Instructions" button that allowed participants to re-read the task instructions. Finally, the "Assignment Complete" button allowed participants to submit the task, but only when all cases were successfully prioritized.

Upon completing the task, participants were asked to fill out a short demographics survey including questions about age, sex, and race. Finally, participants were thanked for their involvement and debriefed.

Results

Cumulative Link Mixed Model. Because the priority rating scale involved data that are inherently ordinal in nature, the best approach for modeling involved a cumulative link mixed model (CLMM), which I conducted via the *ordinal* package (Christensen, 2015; version 2015.1-21). Although a linear mixed effects model (LMM) could also be used, such a model would not

be ideal for two reasons: 1) the spacing between the ordinal priority assignment points is not necessarily equidistant; and 2) more importantly, each participant's rankings are not independent (e.g., only a single case file can be given first priority, meaning that the remaining case files are necessarily given a priority of 2nd through 6th).

As a reminder, the design consisted of a 3 (Justification Type: Confidence Only, Unobservable, Observable Feature) x 2 (Statement Strength: High Confidence vs. Moderate Confidence) interaction, with all factors represented within subject. Because participants saw a total of six lineups (i.e., one per each unique combination of the within-subjects factors), random effects necessarily consisted of a simple intercept within participant.

A series of five models, from the full interaction to the intercept only, were fit to the data and compared using AIC to select the best model. The model with the lowest AIC included the full interaction, as shown in the table below. All five models included a random effect of participant and were fit using 10 quadrature points and flexible theta threshold points for the Priority Assignment scale. Follow-up analyses of the distance between scale points suggested that participants viewed the difference from one point to the next across the scale as approximately equal (e.g., the distance from 1st priority to 2nd priority is roughly equivalent to the distance from 5th priority to 6th priority).[19] A table of scale point distances for the flexible interaction model is provided in Appendix A.

[19] In fact, a full interaction model with an equidistant threshold had a lower AIC than the same model using flexible thresholds (2111.58 vs. 2115.44, respectively, a difference of 3.86 AIC). However, the *effects* package only recently (i.e., version 3.0-4, released March 25, 2015) provided a method for plotting CLMMs with confidence intervals, but only when using flexible threshold CLMMs (via conversion of the fixed effects to a *polr* model from the *MASS* package [Venables & Ripley, 2002; version 7.3-40]). For equidistant threshold CLMMs, predicted effects can be computed from model coefficients, but there is no recommended approach for computing confidence intervals. However, a plot of the model estimates for the equidistant CLMM of the full interaction is included in Appendix A for comparison.

Model Formula	Degrees of Freedom	AIC
StatementStrength * JustificationType	11	2115.44
StatementStrength + JustificationType	9	2129.34
JustificationType	8	2414.78
StatementStrength	7	2142.05
1 (Intercept Only)	6	2420.13

Table 8. AIC table for the cumulative link mixed models in Experiment 3. Model formulae are provided in Wilkinson-Rogers notation where an asterisk (i.e., "*") signifies an interaction that includes all lower order effects.

As shown in Figure 16 below and as predicted, participants were significantly more likely to assign higher priority to cases involving high confidence statements and vice versa for moderate confidence statement. However, highest priority (i.e., an assignment of 1st) was most likely to be given to highly confident Unobservable justifications, whereas lowest priority was most likely to be given to moderately confident Confidence Only statements.

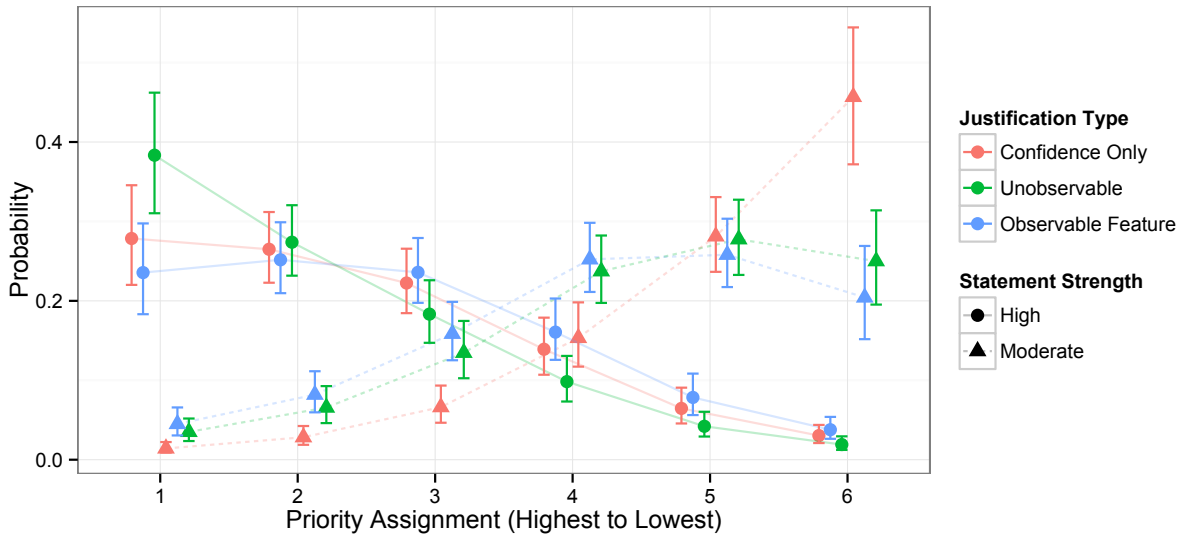


Figure 16. Model estimates for the interaction between Justification Type and Statement Strength in Experiment 3. The circular points show high-confidence statements, whereas the

triangular points show moderate-confidence statements; different colors represent different Justification Types. The x-axis indicates scale points on the Priority Assignment scale: points to the left (i.e., towards 1st) represent higher priority and points to the right lower priority; the y-axis indicates probability of assignment. Error bars represent a 95% confidence interval.

Linear Mixed Model. One issue with the CLMM approach is that it is difficult to conduct follow-up tests on the interactions. For this purpose, I also report the LMM that treats priority assignment as a continuous factor. Given that the scale points appear to be roughly equally spaced relative to one another as mentioned earlier, this assumption of the LMM is less likely to be problematic despite the ordinal nature of the data. However, it is worth noting that the LMM is unable to take the relative rankings of cases along the priority scale into account, although the overall results from the LMM model are consistent with the CLMM approach.

A likelihood ratio test showed significance for all factors within the full interaction model: a significant interaction between Justification Type and Stimulus Strength, $\chi^2(2) = 15.52$, $p < .001$, a significant main effect of Justification Type, $\chi^2(2) = 14.49$, $p < .001$, and a significant main effect of Stimulus Strength, $\chi^2(1) = 312.33$, $p < .0001$. Because both significant main effects are marginal to the significant interaction, I focus on describing exclusively the interaction, which is displayed in Figure 17 below.

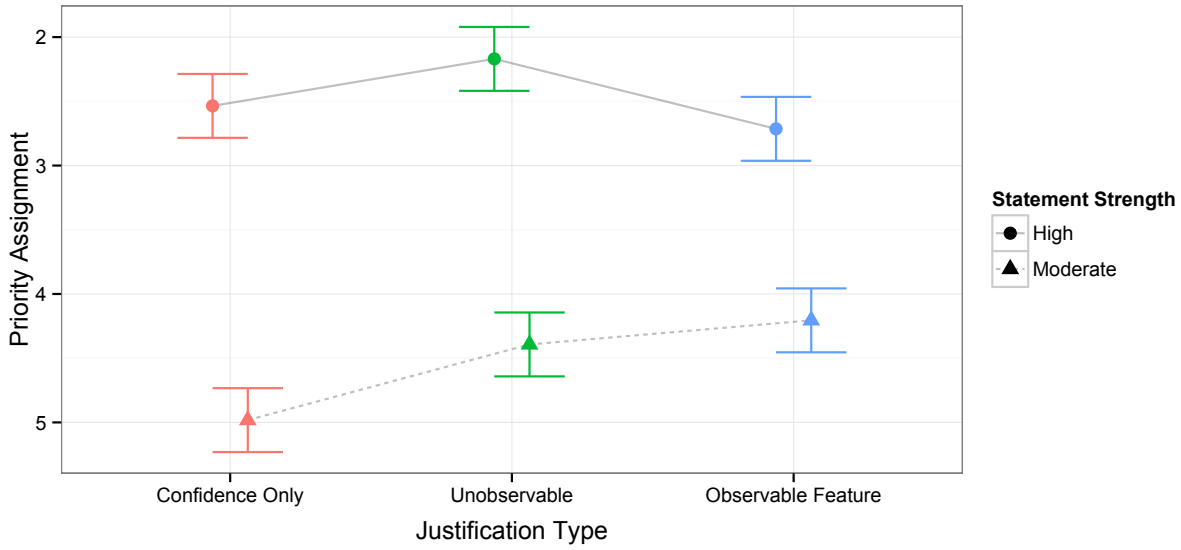


Figure 17. Model estimates for the interaction between Justification Type and Statement Strength in Experiment 3. The circular points show high-confidence statements, whereas the triangular points show moderate-confidence statements. The x-axis consists of the different justification types and the y-axis indicates Priority Assignment, modeled as a continuous variable, where higher values (i.e., approaching 1) indicate higher priority and lower values (i.e., approaching 6) lower priority. Error bars represent a 95% confidence interval.

Similar to the CLMM, the LMM showed a strong and expected effect of Statement Strength, such that highly confident cases were associated with higher priority ratings. However, for highly confident statements, Confidence Only and Observable Feature justifications were not significantly different from one another, $\chi^2(1, N = 112) = 0.99$, $p = .32$, whereas Unobservable justifications produced higher priority assignments compared to Confidence Only, $\chi^2(1, N = 112) = 4.17$, $p = .04$, and Observable Feature justifications, $\chi^2(1, N = 112) = 9.24$, $p < .01$,

respectively.[20] By contrast, for moderately confident statements, Confidence Only statements received a lower priority compared not only to Unobservable justifications, $\chi^2(1, N = 112) = 10.82$, $p < .01$, but also to Observable Feature justifications, $\chi^2(1, N = 112) = 18.79$, $p < .0001$, but Unobservable and Observable Feature justifications were no different from one another, $\chi^2(1, N = 112) = 1.10$, $p = .30$.

Discussion

This experiment focused on behavior rather than perceived confidence. Consistent with expectations, participants assigned priority to cases differently depending upon the overall level of confidence: highly confident statements were associated with a higher priority ranking than moderately confident statements, showing that participants understood the task and that eyewitness expressions of certainty had the capacity to influence behavior in this paradigm.

The effects of justification type were more complex. For moderately confident statements, both observable and unobservable justifications were assigned a higher priority than confidence only statements, and the latter received the lowest ratings overall. By contrast, for highly confident statements, observable feature statements were rated on par with confidence alone, whereas unobservable justifications received higher priority ratings than either observable feature justifications or confidence alone. These results do not support the Perceived Diagnosticity account, which argues that observable featural justifications are viewed skeptically by others, causing them to be devalued in terms of perceived numeric value relative to both unobservable justifications and confidence alone (Dodson & Dobolyi, 2015). Instead, the results suggest a new phenomenon akin to an *unobservable inflation effect*, given that highly confident unobservable justifications were more highly prioritized than any other condition.

[20] An alternative contrast of unobservable justifications versus the combination of observable justifications and confidence only was also significant, $\chi^2(1, N = 112) = 8.61, p < .01$.

It is important to emphasize that this effect was observed using a design that varied justification type within subjects whereas Dodson and Dobolyi (2015) varied justification type between-subjects. Moreover, Experiment 3 revolved around switching back and forth among lineups to determine which represents the strongest evidence—a task that focuses on making relative comparisons across different justification types. This also sets Experiment 3 apart from Experiment 2, despite both involving within-subjects designs: in Experiment 2, participants viewed lineups in isolation with no explicit way of comparing one expression of certainty to another beyond having encountered preceding stimuli.

This issue of relative comparisons is important, because it helps explain the underlying mechanisms of the *unobservable inflation effect*. For example, in the context of such a task it might have been reasonable to expect that expressions of certainty involving justifications of confidence would be rated more highly than confidence statements alone even when the overall level of confidence was high: the inclusion of a justification clearly represents additional information beyond just the confidence statement (e.g., "I'm very sure" versus "I'm very sure. I remember his nose"). However, whether or not observable featural justifications are varied within- or between subjects, it is still true that observable justifications are open to interpretation by others in ways that unobservable justifications are not. Thus, it is possible that in a within-subjects design, the addition of an observable featural justification adds no value above confidence alone for highly confident statements because the justification is open to doubt (i.e., when an eyewitness says "I'm very sure. I remember his nose," others may doubt that the nose could help discriminate among a fair lineup of faces and thus rate this statement similarly to a statement that included confidence alone). By contrast, unobservable justifications are not open to reevaluation. If an eyewitness says, "I'm very sure. I remember him," then there is no

obvious reason to doubt the eyewitness's justification. Thus because unobservable justifications are simultaneously not open to reinterpretation and also more informative than confidence alone, these statements represent the strongest type of evidence. This is ultimately important knowledge from a criminal and legal perspective.

Experiment 4

Experiments 2 and 3 have shown that observable featural justifications are not weighted differently than confidence alone, whereas unobservable justifications are associated with higher perceived confidence. This pattern differs from the one identified by Dodson and Dobolyi (2015), who found that observable featural justifications were associated with lower perceived confidence than both unobservable justifications and confidence alone—a finding referred to as the *featural justification effect*. As previously discussed, this discrepancy is likely due to differences in design: Experiments 2 and 3 varied justification type within-subjects, whereas Dodson and Dobolyi (2015) varied justification type between-subjects.

An important question is what is the underlying mechanism of the *featural justification effect*? According to the Perceived Diagnosticity account, it is a result of individuals' experience with faces that allows them to make judgments about the relative memorability of features (Dodson & Dobolyi, 2015). But a basic question is what drives this judgment about relative memorability? In other words, does it require experience and familiarity with an object? Or alternatively, can individuals develop an expectation "on the fly" about the likely memorability of a feature of a novel object?

One approach to answering these questions involves manipulating the novelty and familiarity of the stimuli that participants encounter. Since most people have expertise with faces, it is reasonable to expect that individuals should be able to assess the diagnosticity and

potential memorability of a specific facial feature. Moreover, the size of the effect should be strongest for highly confident statements: it is only in these instances that individuals have reason to be skeptical of a justification involving a feature that does not appear particularly memorable (e.g., "how can the eyewitness be so confident on the basis of remembering a chin that does not seem particularly memorable").

If expertise is responsible for the effect, then it also ought to occur for stimuli other than faces if the objects are relatively common, such as cars or weapons, which are often the subject of eyewitness identifications, albeit potentially to a lesser degree. Critically however, if expectations about the memorability of a feature require expertise and familiarity with the object then the *featural justification effect* ought not to occur for stimuli that are novel because individuals would have no way to assess the memorability of a feature for these stimuli.

To evaluate if this is the case or not, Experiment 4 showed participants stimuli that included not only faces, but also novel "greebles" and other non-novel stimuli such as cars and weapons. Again, if expertise drives the *featural justification effect*, I would not expect to find a difference in perceived confidence when participants evaluate different justification types involving greebles. By contrast, for non-novel stimuli including faces, cars, and weapons, I expect to replicate the *featural justification effect*: lower perceived confidence for observable featural justifications compared to unobservable justifications or confidence alone, particularly for faces where expertise is strongest.

Alternatively, I may find that greebles produce the same pattern of perceived confidence that the Perceived Diagnosticity account predicts for faces (i.e., lower perceived confidence for observable feature statements than other justification types). If this were the case, it would

suggest that people are able to develop expectations "on the fly" about the perceived distinctiveness or memorability of objects even if those objects are completely novel.

Participants were shown a series of objects from one of four stimulus types: faces, greebles, cars, and weapons. Each object was presented in conjunction with an expression of confidence of either moderate or high statement strength. Moreover, these expressions either included or not a justification based on the following three justification types: confidence only (e.g., "I am very certain"), unobservable justification (e.g., "I am very certain. I remember him"), or observable justification (e.g., "I am very certain. I remember his nose"). Each participant encountered eight objects: one for each combination of stimulus type by statement strength.

One difference between Experiment 4 and the previous experiments is that objects were presented in isolation, not in lineups. There are two reasons why I presented the objects in isolation: 1) Dodson and Dobolyi (2015) showed that the featural justification effect occurred regardless of whether participants evaluated lineups or faces in isolation (i.e., participants do not require a lineup to judge the relative memorability of feature since they can use their memory for faces encountered in the past to make these determinations); and 2) presenting the objects in isolation avoids a potential confound involving lineup fairness across different types of objects.[21]

Finally, it is worth noting that justification type was varied between subjects, in contrast to Experiments 2 and 3. In other words, participants encountered the same justification-type for each of eight objects. Thus, given that the design is essentially identical to the faces in isolation

[21] More specifically, lineup stimuli had been previously normed for fairness by Dodson and Dobolyi (2013) using a standard mock eyewitness paradigm such that each lineup was equally likely to be chosen relative to a model description (e.g., Malpass & Lindsay, 1999). However, it is unclear if a similar procedure would produce "fair" greeble lineups since greeble similarity can only be judged abstractly.

condition of Experiment 3 in Dodson and Dobolyi (2015), I expect to exactly replicate the results of that experiment by observing a *featural justification effect* for facial stimuli.

Method

Participants. Participants were 202 white individuals between the ages of 18 and 40 (M = 28.72, SD = 5.37, range = 18.2 – 39.9; 49.01% female) who were recruited via Amazon Mechanical Turk in exchange for payment. An a-priori power analysis conducted using G*Power (Faul et al., 2007) using $\alpha = .05$ showed that we would have 99% power to detect medium-sized effects (Cohen's $f = .25$; Cohen, 1988) in the context of a mixed factorial ANOVA with 200 participants. All participants gave consent and completed a brief demographic questionnaire. Participants were divided across the between-subjects factor of Justification Type as follows: Confidence Only (n = 67), Unobservable (n = 68), and Observable Feature (n = 67).

Design. The experiment was a mixed factorial design consisting of a 4 (Stimulus Type: Faces, Greebles, Cars, Weapons) x 2 (Statement Strength: High Confidence vs. Moderate Confidence) x 3 (Justification Type: Confidence Only, Unobservable, Observable Feature). Stimulus Type and Statement Strength varied within-subjects and Justification Type varied between-subjects.

Materials. The experiment was conducted using a custom browser-based framework built using PHP, jQuery/JavaScript, MySQL, and HTML. The entire experiment is available online at http://www.dodsonlab.com/studies/faces_e4/ and via a debug link that allows the full task to be seen in a single page: http://dodsonlab.com/studies/faces_e4/?debug.

Stimuli for the experiment consisted of two exemplars within each of the following four categories: white faces used in the preceding studies and in Dodson and Dobolyi (2015), 3D models of asymmetric greebles from the CNBC Stimuli Repository

(http://wiki.cnbc.cmu.edu/Novel_Objects), and images of cars and weapons collected via image searches. Greebles were pre-categorized based on feature similarity: I selected one greeble from two of five pre-constructed categories.

For cars, I used two different models (i.e., one sedan and one SUV) of identical year, make, and color, matched such that both had the same optional features (e.g., both cars were painted the same shade of white and were of identical trim levels, featuring similar fog lamps and moon roofs). Weapons consisted of a revolver and a shotgun: both were black and matched for features (i.e., textured grip, prominent sights, exposed barrel, and visible triggers). The photographs of the cars were taken from the same orientation: facing the car directly from the front at roughly eye-level and against a white background. Similarly, the photographs of the weapons were taken against a white background and visible in "profile": handles to the left and barrels pointing right.

Confidence statements were generated in advance in a manner consistent with Dodson and Dobolyi (2015). A series of statements were designed for each stimulus type in an effort to make them as similar possible. For example, greeble observable feature justifications consisted of statements like "I remember the horns" while unobservable justifications included "I remember it." Analogous phrases for faces included "I remember his chin" and "I remember him," respectively. To keep the statements as similar as possible across stimuli, all justifications began with an identical phrase regardless of stimulus type (e.g., "I remember ____"). The phrase was then completed for each stimulus type: e.g., for a featural justification involving a face, one might encounter "I remember his nose"; alternatively, for a car, the phrase might be "I remember its headlights." A complete table of the phrases used within Experiment 4 is provided in Table 9 below.

			Observable Feature Modifiers				Unobservable Modifiers	
Statement Strength	**Confidence Statement**	**Justification Phrase**	**Face**	**Greeble**	**Car**	**Weapon**	**Face**	**Greeble, Car, and Weapon**
High	I am very certain	I remember ___	his nose	its horns	its headlights	its trigger	him	it
High	I am totally positive	I recall ___	his eyes	its arms	its grille	its grip	him	it
High	I am absolutely confident	I recognize ___	his mouth	its head	its mirrors	its barrel	him	it
High	I am very sure	I recollect ___	his eyebrows	its base	its hood	its sights	him	it
Moderate	I am mostly certain	I remember ___	his nose	its horns	its headlights	its trigger	him	it
Moderate	I am pretty sure	I recall ___	his eyes	its arms	its grille	its grip	him	it
Moderate	I am fairly confident	I recognize ___	his mouth	its head	its mirrors	its barrel	him	it
Moderate	I am moderately sure	I recollect ___	his eyebrows	its base	its hood	its sights	him	it

Table 9. Confidence and justification statements for Experiment 4. Note that only Confidence Level and Confidence Statements vary across all eight rows. Justification phrases and modifiers are repeated within levels of Statement Strength. Justification phrases were randomly assigned to modifiers across participants (e.g., one participant may have encountered "I remember his nose" while another may have encountered "I recall his nose").

Procedure. The general procedure of Experiment 4 is similar to the previous experiments. Participants were instructed to pretend they were police officers and were told that they would see a series of objects eyewitnesses identified as having previously seen. Moreover, for each object, participants were told that they would see the eyewitness' written expression of certainty regarding his or her identification. The participants' task was "to translate the eyewitnesses' written expressions onto a numeric confidence scale" that ranged from 0 (Not at All Certain) to 100 (Completely Certain) on a six-point scale in intervals of 20.

Participants viewed a total of nine objects. The first object was always a yellow smiley face with the following written expression of certainty: "I am completely certain this was the yellow smiley face I saw earlier. The color was yellow." This smiley face served as a manipulation check: only participants who correctly translated this statement to 100 (Completely Certain) were included in the final sample.

The following eight objects consisted of each combination of the within-subjects factors presented in a randomized order: Stimulus Type (Faces, Greebles, Cars, Weapons) x Statement Strength (High Confidence vs. Moderate Confidence). As shown in Figure 18 below involving a greeble, beneath each object was a label—"Did you see this?"—and two buttons indicating a "YES" or "NO" response, with the "YES" option always highlighted with a red border to remind participants that these objects were previously encountered by an eyewitness.

Figure 18. The main task from Experiment 4. Similar to Experiment 2, participants were asked to translate the verbal expression of certainty onto a numeric scale. Unlike Experiment 2, participants encountered faces as well as greebles, cars, and weapons.

Furthermore, beneath the label and buttons, each object included a text box containing a confidence statement that may or may not have included an additional justification. Regardless of the type of stimulus encountered, participants saw statements that varied based on Statement Strength: some conveyed high confidence (e.g., "I am positive") and others moderate confidence (e.g., "I am pretty sure"). In addition, Justification Type also varied between-subjects, with all

expressions of certainty involving one of the following three categories: 1) Confidence Only; 2) Unobservable; or 3) Observable Feature. Finally, beneath each expression of certainty was the six-point numeric confidence scale. Again, participants were reminded of their instructions: "Translate the written expression of certainty above onto the following numeric scale." Clicking on a radio button on the numeric scale advanced the study to the following object.

Upon translating each of the eight critical expressions of certainty, participants were asked to fill out a short demographics survey including questions about age, sex, and race. Finally, participants were thanked for their involvement and debriefed.

Results

Given that the design included both within- and between-subject factors, I fit a linear mixed model of the full interaction, which consisted of a 3 (Justification Type [Between]: Confidence Only, Unobservable, Observable Feature) x 2 (Statement Strength [Within]: High Confidence vs. Moderate Confidence) x 4 (Stimulus Type [Within]: Faces, Greebles, Cars, Weapons). Because participants saw a total of eight lineups (i.e., one per each unique combination of the within-subjects factors), random effects necessarily consisted of a simple intercept within participant.

A likelihood ratio test showed three significant effects for this model: a significant interaction between Justification Type and Stimulus Strength, $\chi^2(2) = 11.32$, $p < .01$, a significant main effect of Justification Type, $\chi^2(2) = 15.83$, $p < .001$, and a significant main effect of Stimulus Strength, $\chi^2(1) = 779.59$, $p < .0001$; no other effects reached significance (all ps $> .30$). Because both significant main effects are marginal to the significant interaction, I focus the description of the results exclusively on the latter.

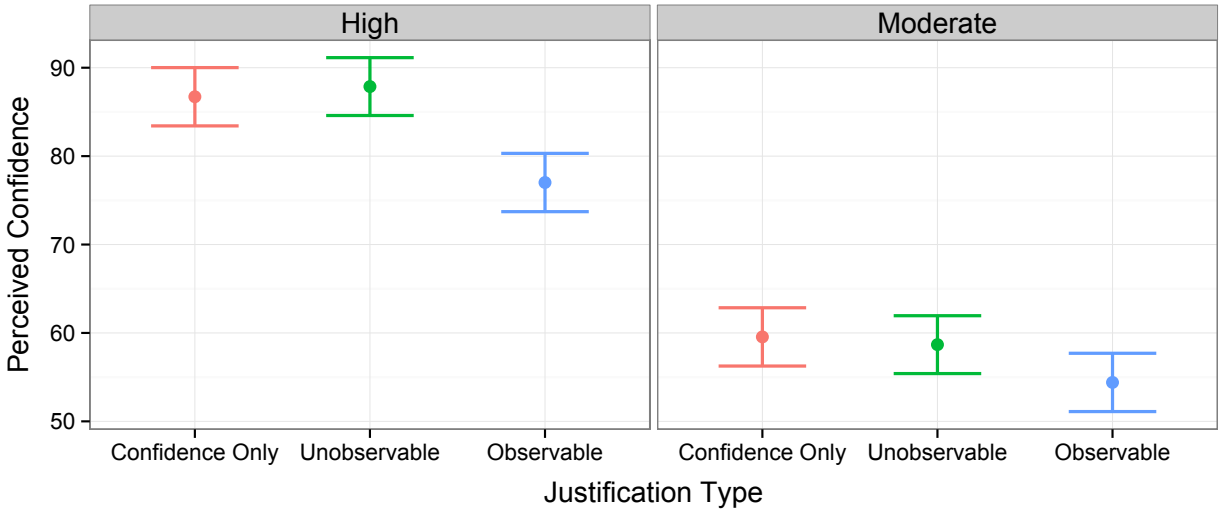


Figure 19. Model estimates for the significant two-way interaction between Justification Type and Statement Strength in Experiment 4. The left panel shows high-confidence statements, whereas the right panel shows moderate-confidence statements; as expected, high confidence statements are perceived as more confident than low confidence statements. However, Observable Feature statements are associated with lower perceived confidence than Unobservable or Confidence Only statements, and this effect is most pronounced for highly confident statements. Error bars indicate a 95% confidence interval.

As shown in Figure 19 above, the results replicate the *featural justification effect* described in Dodson and Dobolyi (2015): while there was no difference in perceived confidence between Unobservable or Confidence Only statements, $\chi^2(1, N = 135) = 0.14$, $p = .95$, Observable Feature statements were associated with lower perceived confidence than Unobservable and Confidence Only statements combined, $\chi^2(1, N = 202) = 16.21$, $p < .0001$. This effect of Justification Type varied based on Statement Strength: specifically, it was more pronounced for highly confident statements, $\chi^2(1, N = 101) = 24.99$, $p < .0001$, than for moderately confident statements, $\chi^2(1, N = 101) = 5.25$, $p = .02$.

One implication of our Perceived Diagnosticity account is that the *featural justification effect* on perceived confidence should show a strongest effect for stimuli for which people have expertise or experience (i.e., for faces in particular, and for cars and weapons) and less so for completely novel stimuli such as greebles. Experiment 4 does not support this assumption of the Perceived Diagnosticity account: the three-way interaction between Justification Type, Statement Strength, and Stimulus Type was non-significant ($p = .96$), as were all effects involving Stimulus Type (all $ps > .31$). A plot of the non-significant interaction is shown in Figure 20 below.

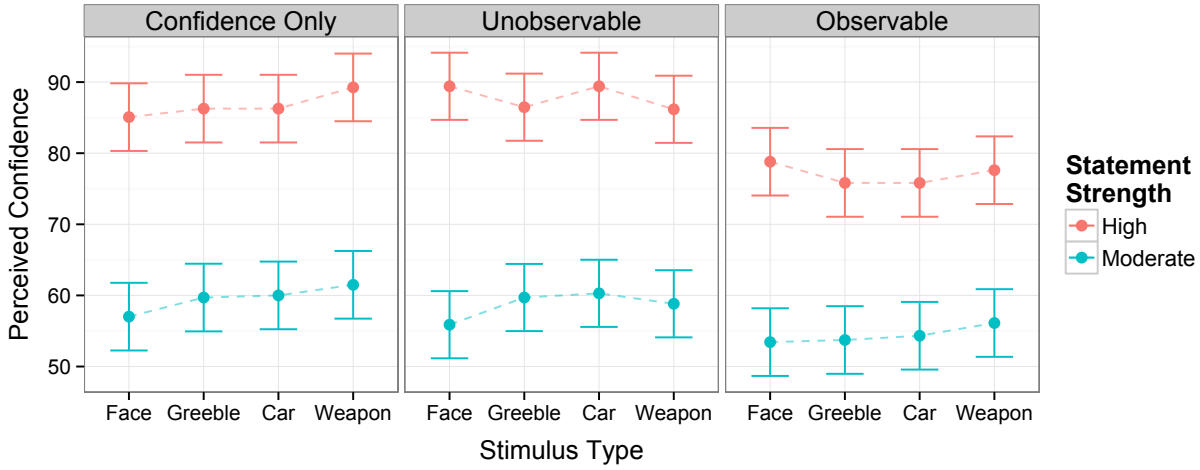


Figure 20. Model estimates for the non-significant three-way interaction between Justification Type, Stimulus Type, and Statement Strength in Experiment 4. The red lines indicate perceived confidence for high-confidence statements, whereas the blue lines indicate moderate-confidence statements. Justification Type is split across the three facets, and within each facet the x-axis indicates Stimulus Type. Although the two-way interaction between Justification Type and Statement Strength was significant in the model, there are not significant differences relative to Stimulus Type. Error bars represent a 95% confidence interval.

Discussion

Experiment 4 replicates a key finding of Dodson and Dobolyi (2015): within the context of a between-subjects design, the identical confidence statement about the identification of a face is interpreted as less confident when it is justified by referring to an observable feature than by either an unobservable feature or with no accompanying justification at all (i.e., confidence only). The proposed explanation for this finding is the Perceived Diagnosticity account. Justifications that refer to an observable feature cause individuals to evaluate the relative memorability or diagnosticity of this feature. Individuals are likely to devalue a confidence statement—especially a high confidence one—when the referenced observable feature does not appear particularly memorable. Dodson and Dobolyi (2015) had assumed that expectations about the relative memorability of a feature depended on familiarity and experience with the object, such as faces.

Results of Experiment 4 suggest that the Perceived Diagnosticity account is not specific to faces however: regardless of stimulus type (i.e., faces, completely novel greebles, and common, crime-relevant, non-facial stimuli such as cars and weapons), the *featural justification effect* still occurs: perceived confidence was lower when justifications included an observable feature compared to an unobservable justification or confidence alone. Thus, the Perceived Diagnosticity account needs to be modified to account for the effect without relying on expertise.

An alternative explanation is that for novel objects, people are judging the memorability (or non-memorability) of features spontaneously, or "on the fly." This would be consistent with research on metamemorial inference based on the perceived memorability of words: several studies have shown that individuals judge the subjective memorability of words in recognition tasks, and these judgments are often quick or automatic (Brown, Lewis, & Monk, 1977;

Guttentag & Carroll, 1998; Hintzman, Caulton, & Curran, 1994; Wixted, 1992). Wixted's (1992) study is perhaps the most relevant to the current experiment: participants were shown a series of low frequency and high frequency words and asked to imagine that these words had been previously studied; the task was to judge the relative memorability of these words under the assumption that their memory would be tested later in a recognition task. Ultimately participants' judgments proved to be quite poor since they judged high frequency words as more memorable than low frequency words; by contrast, in standard recognition paradigm, the opposite pattern is typically found (i.e., low frequency words are more accurately recognized than high frequency words).

In the present experiment participants were tasked with judging the perceived confidence of identifications made by others, and in this respect, this task was similar to Wixted's (1992) in that it also involved simulation. Assuming participants used their imagination to infer the basis of recognition judgments made by others, it is likely that they would have tried to evaluate what would be useful for making accurate judgments. Thus, even though they may have had no prior experience with greebles, they could still attempt to infer the relative distinctiveness or memorability of greeble features. Given that observable featural justifications for greeble identifications were perceived as less confident in a pattern similar to observable features statements involving faces, it is likely that participants viewed greeble features as not particularly diagnostic (e.g., if the justification statement was "I remember its horn," participants may have felt that the horn would not be particularly useful for recognizing one particular greeble relative to another).

To test this possibility, it would be interesting to conduct a new experiment involving a series of greebles identifications that vary based on featural distinctiveness. For these greebles,

horns could vary widely, such that a particular horn would be very obviously unique to one and only one greeble; by contrast, body shapes would only vary slightly across all the greebles. In this example, I would expect that an observable feature mentioning a greeble's horns would be associated with higher perceived confidence than one that referred to the body shape because the former feature is highly distinctive and far more diagnostic for the identification of one greeble versus another. I expect this effect would be particularly strong if the task allowed for cross-lineup comparisons similar to Experiment 3.

Finally, it is worth noting that Experiment 4 exactly replicates the *featural justification effect* observed by Dodson and Dobolyi (2015): confidence for observable features was consistently lower than for unobservable justifications or confidence alone across all stimulus types. There is a straightforward explanation: the present study—unlike Experiments 2 and 3 that preceded it—varied justification type between subjects. Thus, there are now several studies using a between-subjects design that show one pattern of results (i.e., the *featural justification effect*), and others showing a different effect (i.e., higher perceived confidence for unobservable justifications than observable feature justifications or confidence alone). Still, to verify that this between versus within difference accounts for these differences directly, a follow-up study that includes both designs would be invaluable.

It is also worth noting that Experiment 4—as well as Experiment 3—defined unobservable justifications in a manner consistent with Dodson and Dobolyi (2015) but different from Experiments 1 and 2. For example, in this experiment "I remember him" was categorized as unobservable, whereas in Experiment 1 "I definitely remember this guy" was categorized as an unknown statement. This difference exists because there are multiple types of unobservable statements: 1) those that refer to a specific feature (e.g., "He looks like a friend of mine") or 2)

those that refer more generally to a statement of recognition (e.g., "I recognize him/it"). Given the constraints of (a) using different stimuli (i.e., faces, greebles, cars, and weapons) and (b) the need to standardize the unobservable descriptions so that they could apply universally across stimulus types, it was necessary to settle on the latter type (i.e., statements of recognition) for unobservable statements. This does raise an important question however: is the statement "I recognize him" treated differently than "He looks like a friend of mine"? In other words, is a statement of recognition weighted similarly or differently compared to a statement that mentions a specific unobservable feature? This question will be answered in a future experiment comparing both types of unobservable statements directly.

General Discussion

I conducted four experiments that advance our knowledge of the interpretation of eyewitness confidence. These experiments focused on four key questions: (1) are eyewitness justifications—when combined with confidence and decision time—meaningful postdictors of identification accuracy?; (2) how accurately can observers interpret the intended meaning of an eyewitness's confidence statement given a particular type of justification?; (3) what is the consequence of a particular kind of justification on an observer's behavior?; and (4) are differences in perceived confidence across different types of justifications a result of expertise with faces or do these findings represent a more general memory phenomenon?

Regarding the latter two points, Experiments 3 and 4 showed a consistent pattern: unobservable justifications were perceived as a stronger piece of evidence than observable justifications. This was true regardless of whether or not justification type varied within-subjects (Experiments 3) or between-subjects (Experiment 4), although the between versus within distinction did influence how others perceived confidence only statements. Specifically, when

justification type varied within-subjects (i.e., in Experiment 3), highly confident observable featural justifications were ranked similarly to confidence alone. In other words, despite the fact that an observable justification was added to an otherwise identical high confidence statement (e.g., "I am very sure. I remember his nose" versus "I am very sure," respectively), this additional observable justification did not lead to a higher rating relative to confidence alone. Rather, only the addition of an unobservable justification led to a rating higher than confidence alone: a finding I refer to as the *unobservable inflation effect*. Because unobservable justifications—relative to observable featural justifications—are not open to interpretation (e.g., when an eyewitness says "I'm very sure. I remember him," there is no clear reason to doubt this statement), they represent stronger evidence than confidence alone. By contrast, observable featural justifications are open to interpretation (e.g., when an eyewitness says "I'm very sure. I remember his nose," others may wonder if that nose is particularly diagnostic) and thus more open to doubt, meaning that they will be associated with lower perceived evidentiary when compared to an unobservable justification.

This latter pattern (i.e., lower perceived value for an observable justification versus an unobservable justification) is consistent with the Perceived Diagnosticity account (Dodson & Dobolyi, 2015). Experiment 4 replicated the *featural justification effect* described by Dodson and Dobolyi (2015) using an identical between-subjects design: both unobservable justifications and confidence alone were associated with higher perceived confidence than observable featural justifications. More importantly, Experiment 4 showed that the *featural justification effect* is not specific to faces: participants showed the same pattern of reduced perceived confidence for observable featural justifications for faces, greebles, cars, and weapons. The fact that the same pattern appeared for greebles—which are completely novel and thus obviously not associated

with expertise—is the most surprising, since it suggests that people are able to surmise the perceived memorability of a feature "on the fly" without the need for expertise.

While Experiments 2, 3 and 4 were focused on perceived confidence, Experiment 1 used a standard eyewitness memory paradigm (e.g., Meissner et al., 2005) and evaluated how participants justified their eyewitness identifications and non-identifications (i.e., a response of "not present"). Previous studies have shown that postdictors of eyewitness identifications such as confidence and decision time are predictive of accuracy, particularly when choosing a face within a lineup (e.g., Dodson & Dobolyi, under review; Sauerland & Sporer, 2009; Sauerland et al., 2012). I replicated this pattern in Experiment 1, but also showed that 1) familiarity justifications (e.g., "He's very familiar") were much more likely to occur in the context of a non-identification response that was only moderately confident; 2) for chooser responses, accuracy was lower for familiarity-based justifications than for both observable and unobservable featural justifications; and 3) the confidence/accuracy relationship was weaker for familiarity-based responses at higher levels of confidence (i.e., 60 and above) compared to other justification types.

This third point was the most interesting in the context of Experiment 2, which involved showing the identifications generated in Experiment 1 to a new set of participants in Experiment 2 who evaluated the intended numeric confidence of the statement. Surprisingly, Experiment 2 showed that perceived confidence was relatively similar across all justification types,[22] which is inconsistent with the pattern I expected to find: reduced perceived confidence for observable featural justifications compared to other types, consistent with the *featural justification effect*.

[22] Only multiple observable features showed statistically higher perceived confidence than confidence alone or a single observable feature; otherwise, all other categories (i.e., familiarity, a single observable feature, and a single unobservable feature) were perceived similarly to confidence alone.

Nevertheless, Experiment 2 identified an important boundary condition for studies involving perceived confidence: results differ depending on whether justification type varies within subjects or between subjects as discussed in regards to Experiments 3 and 4.

This between versus within effect raises important questions for a follow-up investigation: in cases involving multiple eyewitnesses, do different justification types interact with one another? In other words, if there were two eyewitnesses to a crime and one says, "I'm very certain. I remember him," (i.e., an unobservable justification) and another says, "I'm completely sure," (i.e., confidence alone) will this lead to different levels of perceived confidence than if both had expressed unobservable justifications or confidence alone?

Results of the current set of studies suggest that they will. Specifically, when justification type varies within the two eyewitnesses (i.e., the former scenario), I would expect the pattern of perceived confidence to be consistent with the findings of Experiment 3: perceived confidence should be higher for the unobservable justification than confidence alone at high levels of confidence due to the *unobservable inflation effect*. By contrast, when the two eyewitnesses both express confidence alone or unobservable justifications, I would expect the overall levels of confidence to be perceived similarly across all statements consistent with Experiment 4 (i.e., a between-subjects design). I look forward to investigating this issue further in a future study.

Overall, then, these findings from these four experiments suggest that 1) unobservable justifications represent the strongest form of evidence regardless of the context in which they occur; 2) familiarity-based justifications should be met with additional scrutiny in criminal and legal proceedings; and 3) the *featural justification effect* is not particular to faces and may be the result of "on the fly" judgments of perceived memorability rather than expertise.

References

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

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*(3), 257-269. doi: 10.1002/for.3980010305

Brewer, N., Weber, N., Clark, A., & Wells, G. L. (2008). Distinguishing accurate from inaccurate eyewitness identifications with an optional deadline procedure. *Psychology, Crime & Law*, *14*, 397-414. doi: 10.1080/10683160701770229

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

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency, and negative recognition. *Quarterly Journal of Experimental Psychology*, 29**,** 461-473. doi: 10.1080/14640747708400622

Brun, K., & Teigen, W. H. (1988). Verbal probabilities: Ambiguous, context-dependent or both? *Organizational Behavior and Human Decision Processes, 41*, 390-404. doi: 10.1016/0749-5978(88)90036-2

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,* 299-308. doi: 10.1111/j.1467-9280.2009.02284

Budescu, D. V., Karelitz, T. M., & Wallsten, T. S. (2003). Predicting the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making, 16*, 159-180. doi: 10.1002/bdm.440

Budescu, D. V., Por, H. H., & Broomell, S., (2012). Effective communication in the IPCC reports. *Climatic Change, 113*, 181-200. doi:10.1007/s10584-011-0330-3

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag. doi: 10.1007/b97636

Christensen, R. H. B. (2015). ordinal: Regression Models for Ordinal Data. R package version 2015.1-21. http://www.CRAN.R-project.org/package=ordinal/.

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

De Rosario-Martinez, H. (2015). phia: Post-Hoc Interaction Analysis. R package version 0.2-0. http://CRAN.R-project.org/package=phia.

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. doi: 10.1037/a0034596

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

Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting Eyewitness Expressions of Confidence: The Justification Effect. *Law & Human Behavior*. doi: 10.1037/lhb0000120

Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identifications. *Journal of Applied Psychology*, *87*, 951-962. doi: 10.1037/0021-9010.87.5.951

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), 818-835. doi: 10.1037/0022-3514.67.5.818

Erev, I., & Cohen, B.L. (1990). Verbal versus numerical probabilities: Efficiency, biases and the preference paradox. *Organizational Behavior and Human Decision Processes, 45*, 1-18. doi: 10.1016/0749-5978(90)90002-Q

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. doi: 10.3758/BF03193146

Flugstad, A.R., & Windschitl, P.D. (2003). The influence of reasons on interpretations of probability forecasts. *Journal of Behavioral Decision Making*, *16*, 107-126. doi: 10.1002/bdm.437

John Fox (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, *8*(15), 1-27.

Gurmankin, A.D., Baron, J., & Armstrong, K. (2004). Intended message versus message received in hypothetical physician risk communications: Exploring the gap. *Risk Analysis*, *24*, 1337-1347. doi:10.1111/j.0272-4332.2004.00530.x

Guttentag, R. E., & Carroll, D. (1998). Memorability judgments for high- and low-frequency words. *Memory & Cognition*, *26*, 951-958. doi: 10.3758/BF03201175

Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*, 275-289. doi: 10.1037/0278-7393.20.2.275

Malpass, R. S., & Lindsay, R. C. L. (1999). Measuring lineup fairness. *Applied Cognitive Psychology*, *13*(S1), S1-S7. doi: 10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-9

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reinitz, M. T., Peria, W. J., Seguin, J. A., & Loftus, G. R. (2011). Different confidence-accuracy relationships for feature-based and familiarity-based memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 507-515. doi: 10.1037/a0021961

Reinitz, M. T., Seguin, J. A., Peria, W., & Loftus, G. R. (2012). Confidence-accuracy relations for faces and scenes: Roles of features and familiarity. *Psychonomic Bulletin and Review*, *19,* 1085-1093. doi: 10.3758/s13423-012-0308-9

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. doi: 10.1007/s10979-009-9192-x

Sauerland, M., Sagana, A., & Sporer, S. L. (2012). Assessing nonchoosers' eyewitness identification accuracy from photographic showups by using confidence and response times. *Law and Human Behavior*, *36*(5), 394-403. doi: 10.1037/h0093926

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46-62. doi:10.1037/a0014560

Semmler, C., Brewer, N., & Douglass, A. B. (2012). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), *Conviction of the innocent: Lessons learned from psychological research.* (pp. 185-209). Washington, D.C.: APA Press. doi: 10.1111/j.1556-4029.2012.02228.x

Singmann, H., Bolker, B., & Westfall, J. (2015). afex: Analysis of Factorial Experiments. R package version 0.13-145. http://CRAN.R-project.org/package=afex.

Sporer, S. L., Penrod, S. D., Read, D., & Cutler, B. (1995). Choosing, confidence and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*(3), 315-327. doi: 10.1037/0033-2909.118.3.315

Technical Working Group on Eyewitness Evidence (2003). Eyewitness evidence: A trainer's manual for law enforcement. U.S. Department of Justice, Office of Justice Programs. National Institute of Justice.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York, NY: Springer.

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. doi: 10.3758/BF03334162

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. doi: 10.1016/0749-596X(86)90012-4

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

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

Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: The unruly 10-12 second rule. *Journal of Experimental Psychology: Applied*, *10*, 139–147. doi: 10.1037/1076-898X.10.3.139

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

Windschitl, P. D., Martin, R., & Flugstad, A. R. (2002). Context and the interpretation of likelihood information: The role of intergroup comparisons on perceived vulnerability. *Journal of Personality and Social Psychology*, *82*, 742-755. doi: 10.1037/0022-3514.82.5.742

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 681-690. doi: 10.1037/0278-7393.18.4.681

Young, S., & Oppenheimer, D. M. (2006). Percentages matter: Framing risk information can affect fear of side effects and medication compliance. *Clinical Therapeutics*, *28*, 129-139. doi: 10.1016/j.clinthera.2006.01.013

Young, S., & Oppenheimer, D. M. (2009). Effect of communication strategy on personal risk perception and treatment adherence intentions. *Psychology, Health & Medicine*, *14,* 430-442. doi: 10.1080/13548500902890103

Appendix A

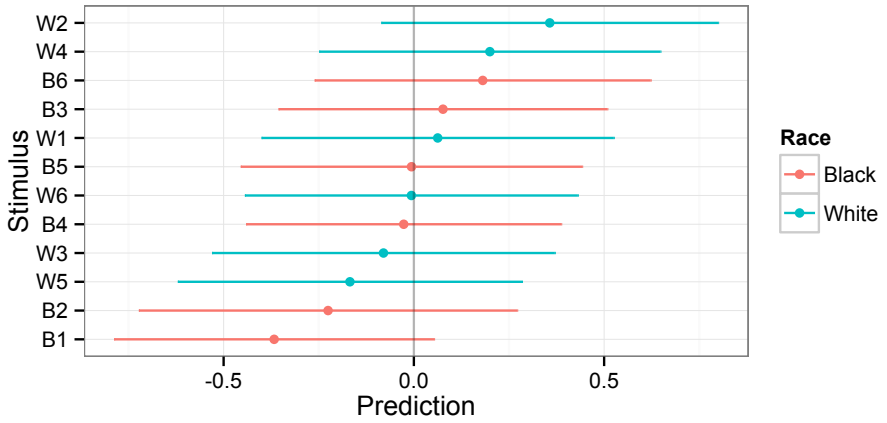


Figure 21. The random effect predictions for lineup stimulus nested within lineup race (i.e., the six black and six white lineups), including a 95% prediction interval. The prediction intervals of all stimuli overlap zero, although two white lineups (i.e., W2 and W4) trend towards higher accuracy and two black lineups (i.e., B1 and B2) trend tower lower accuracy. In Wilkinson-Rogers notation, the formula of the binomial model that generated these random effects was the following: IdentificationAccuracy ~ JustificationType * Confidence * DecisionTime + (1 | LineupRace / LineupNumber) + (1 | Participant).

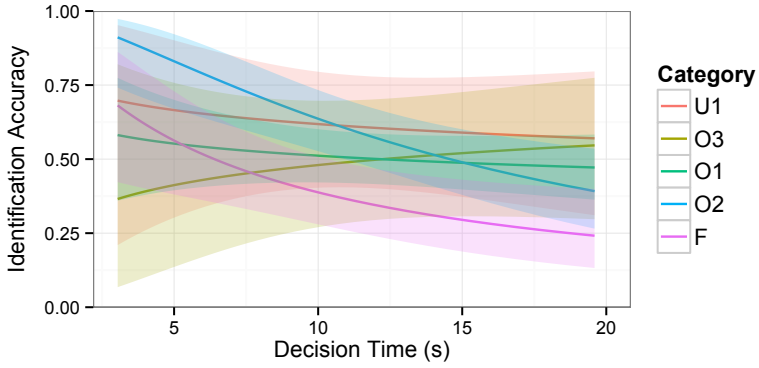


Figure 22. The marginally significant interaction between Decision Time and Justification Type in the best model of chooser identification accuracy in Experiment 1. F refers to familiarity, O

refers to observable feature, and U refers to unobservable feature statements; numbers next to O and U refer to the number of features mentioned. I focus the results on responses made within 20 seconds, as differences become increasingly indiscernible at longer decision times where there are fewer observations. There is a trend towards higher identification accuracy with faster decision times for familiarity and observable feature statements mentioning two features; by contrast, for all other justification types, there is little to no relationship between decision time and identification accuracy (i.e., as evidenced by the flat slopes). Error bars represent a 95% confidence interval.

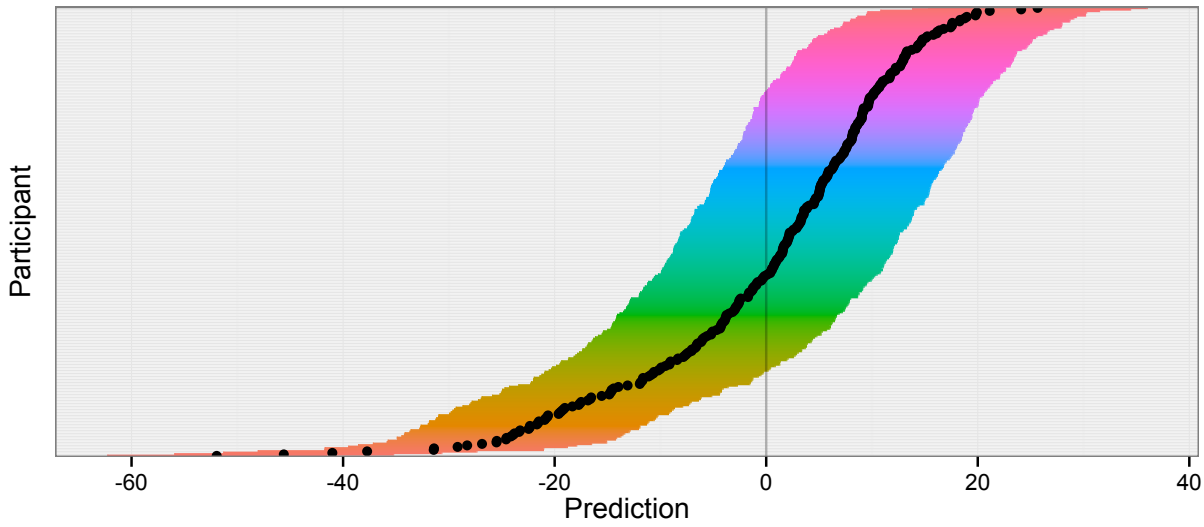


Figure 23. The random effect of participant in a model predicting perceived confidence. Error bars indicate a 95% prediction interval. While generally normally distributed, a handful of participants on the bottom right (i.e., around -40) have lower estimates than the rest of the set. Note that because the upper end of the scale is bounded at 100 at the mean perceived confidence is 61.37, some lower end skew is expected.

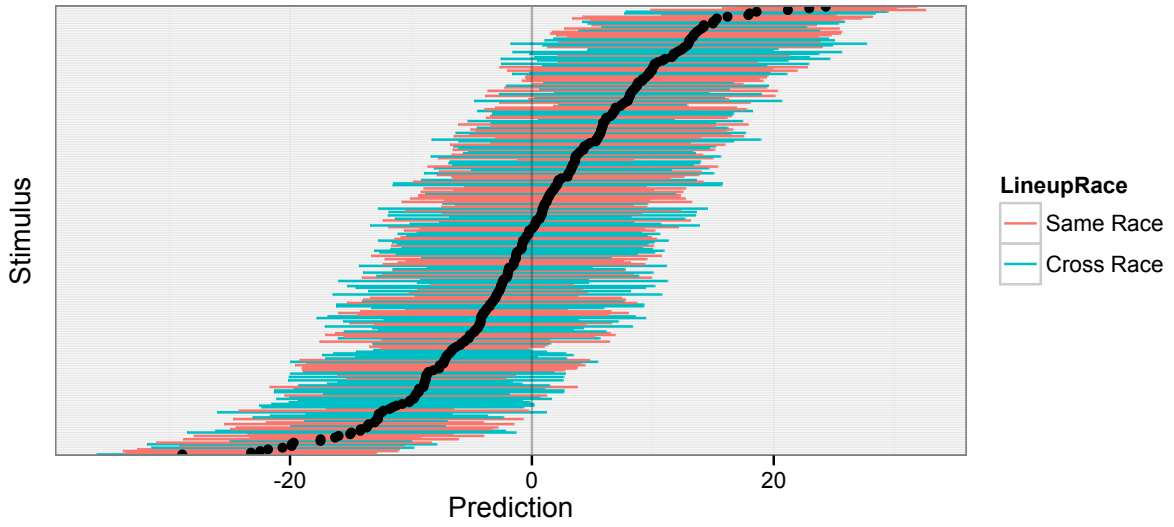


Figure 24. The random effect of stimulus nested within lineup race in a model predicting perceived confidence. Error bars indicate a 95% prediction interval. The pattern is normally distributed with no strong effect of lineup race.

Priority Assignment Threshold	Estimate
1 | 2	-0.95
2 | 3	0.17
3 | 4	1.19
4 | 5	2.25
5 | 6	3.46

Table 10. Scale threshold cut off (theta) points estimated by the flexible cumulative link mixed model of the interaction between Justification Type and Statement Strength. Note that the spacing (i.e., distance from point to point) does not strongly deviate across the scale, suggesting equidistant spacing could be appropriate.

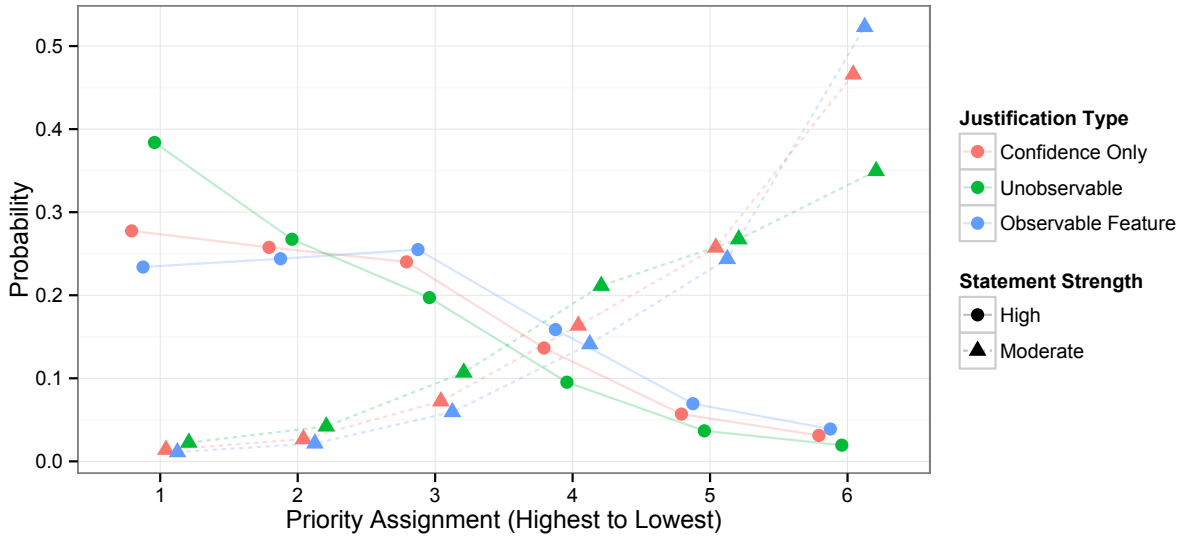


Figure 25. Predicted model estimates of probability for the full interaction of Justification Type by Statement Strength across the six-point Priority Assignment scale for the data in Experiment 4. This cumulative link mixed model (CLMM) uses equidistant threshold spacing. Note that as described in footnote 19, there is currently no recommended method for computing confidence intervals when fitting a CLMM with equidistant spacing.

Appendix B: Instructions for Coding Eyewitness Statements

Basic Information: we collected data from 277 participants who saw a series of 12 eyewitness lineups (2x3 grid arrangements of six faces, with a "Not Present" option). For each lineup, they provided us with: 1) a choice of either one of the six faces or "Not Present"; 2) a verbal confidence statement about their certainty (e.g., "I'm pretty sure"); 3) an additional explanation of this statement (e.g., "I remember his eyes"); and 4) a numeric confidence rating that translates these statements.

The Task: to analyze these data, we would like to code the verbal confidence statements (i.e., parts 2 and 3, above) into different categories. This is where we need your help. The Excel file we provide will have raw participant data, which consists of rows (one per participant) and columns—within the columns are the verbal responses for each lineup. For example, one statement could be: "I'm pretty sure. I remember his eyes."

Each of these statements needs to be coded into one of the following five categories, which are described below, with examples:

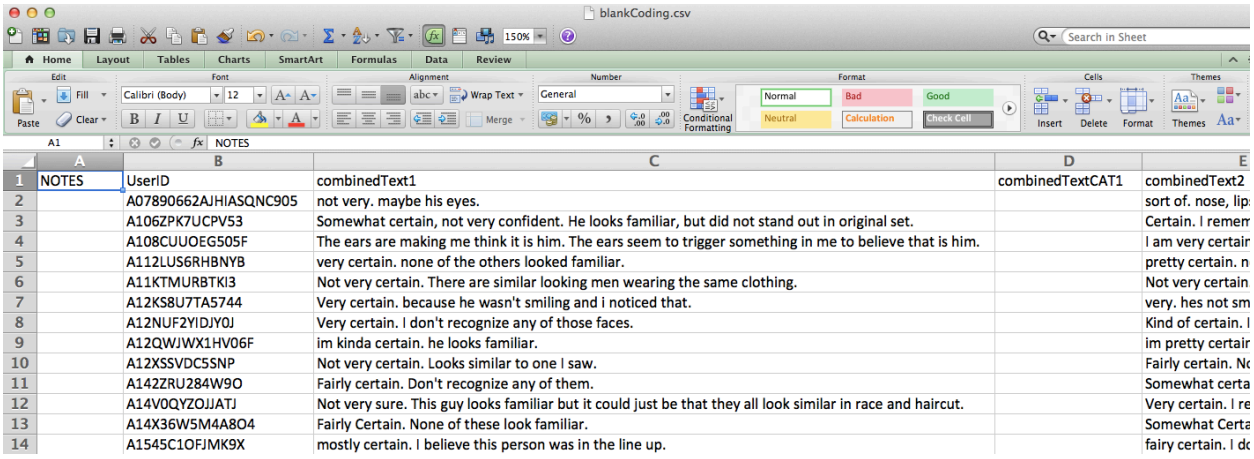
Category	Code	Description	Examples
Pure Familiarity	F	Statements within this category constitute familiarity, i.e., a feeling of having experienced a prior event but without recollecting specific details. This is sometimes thought of as "knowing." *Note:* the statements must include the term "familiar" to be included in this category	"he's familiar" "none of them are familiar"
Observable Feature	O_{count}[23]	Statements that mention specific, observable facial features (and do not mention a feeling of familiarity). *Note:* for this category, you can count the number of features mentioned (see parentheses on the examples to the right).	"I remember his chin" *(1 feature; O1)* "I remember his chin, eyebrow shape and face" (*3 features; O3*)
Unobservable Feature	U_{count}[24]	Statements that mention specific features that are not directly observable by you, the interpreter, assuming you saw the same lineup.	"he looks like a friend of mine" *(U1)*
Mixed	M	Statements include both feature and familiarity aspects. Also use when statements include both types of feature statements (i.e., O and U).[25]	"the shape of the head seems familiar, four of them don't seem familiar at all"
Other	?	The "when in doubt" category. Use this when not sure. Also use this category if the participant mentions some specific detail about the study (see example on the right).	"none are recognized" "I saw each face four times and none had this structure"

[23] For "Observable Feature" and "Unobservable Feature," include a feature count next to the code. To better explain, look at the examples: O1 refers to the number of features counted in the statements "I remember his chin." Also, consider the word "face" to count as one feature. For more on this, see the instructions below for an example.

[24] See the third page of these instructions for some comments about "Unobservable Feature."

[25] We are curious about how often people combine O and U categories into an M response. You can mention these occurrences in the notes.

Coding the Excel Data: the Excel sheet you receive will look something like this:



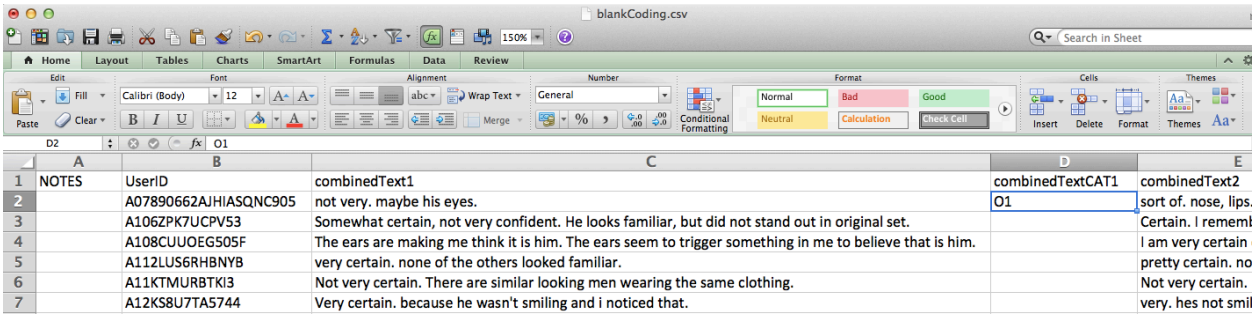
blankCoding.csv

	A	B	C	D	E
1	NOTES	UserID	combinedText1	combinedTextCAT1	combinedText2
2		A07890662AJHIASQNC905	not very. maybe his eyes.		sort of. nose, lip
3		A106ZPK7UCPV53	Somewhat certain, not very confident. He looks familiar, but did not stand out in original set.		Certain. I remem
4		A108CUUOEG505F	The ears are making me think it is him. The ears seem to trigger something in me to believe that is him.		I am very certair
5		A112LUS6RHBNYB	very certain. none of the others looked familiar.		pretty certain. n
6		A11KTMURBTKI3	Not very certain. There are similar looking men wearing the same clothing.		Not very certain
7		A12KS8U7TA5744	Very certain. because he wasn't smiling and i noticed that.		very. hes not sm
8		A12NUF2YIDJY0J	Very certain. I don't recognize any of those faces.		Kind of certain. I
9		A12QWJWX1HV06F	im kinda certain. he looks familiar.		im pretty certair
10		A12XSSVDC5SNP	Not very certain. Looks similar to one I saw.		Fairly certain. No
11		A142ZRU284W9O	Fairly certain. Don't recognize any of them.		Somewhat certa
12		A14V0QYZOJJATJ	Not very sure. This guy looks familiar but it could just be that they all look similar in race and haircut.		Very certain. I re
13		A14X36W5M4A8O4	Fairly Certain. None of these look familiar.		Somewhat Certa
14		A1545C1OFJMK9X	mostly certain. I believe this person was in the line up.		fairy certain. I do

The **UserID** column you can basically ignore, but the **NOTES** column exists for you to write any comments you have for that participant (e.g., things you want to bring to my attention or problems you encountered). The important columns start with **combinedText1** and **combinedTextCAT1**:

- For each row of the Excel sheet, there will be 12 **combinedText** columns, i.e., combinedText1 through combinedText12. ***These are the verbal statements participants made that you need to code.***
- Next to each of these columns is a blank column, **combinedTextCAT**. Again, there will be twelve of these, one for each combinedText column, so combinedTextCAT1 through combinedTextCAT12. ***These are the columns in which you should put in the coding from the table above.***

For each column within the Excel sheet, you would work from left to right (since each participant has 12 observations) and enter a total of 12 codes. For example, in row 2 (UserID A07890662AJHIASQNC905), the first relevant column, combinedText1, says “not very. maybe his eyes.” This is an “Observable Feature” statement, and thus you would enter code “O1” in the corresponding column, i.e., combinedTextCAT1:



blankCoding.csv

	A	B	C	D	E
1	NOTES	UserID	combinedText1	combinedTextCAT1	combinedText2
2		A07890662AJHIASQNC905	not very. maybe his eyes.	O1	sort of. nose, lips.
3		A106ZPK7UCPV53	Somewhat certain, not very confident. He looks familiar, but did not stand out in original set.		Certain. I rememl
4		A108CUUOEG505F	The ears are making me think it is him. The ears seem to trigger something in me to believe that is him.		I am very certain
5		A112LUS6RHBNYB	very certain. none of the others looked familiar.		pretty certain. no
6		A11KTMURBTKI3	Not very certain. There are similar looking men wearing the same clothing.		Not very certain.
7		A12KS8U7TA5744	Very certain. because he wasn't smiling and i noticed that.		very. hes not smil

Be sure to work on coding responses one row at a time, mainly because a given participant is likely to use similar statements, and each row corresponds to one participant's 12 responses.

"Observable" vs. "Unobservable" Feature

It can sometimes be difficult to decide if a feature is observable or unobservable. For a feature to be unobservable, it would mean that someone looking at the lineup, given a set of verbal statements, would not be able to see the feature the eyewitness had mentioned. For example, consider this screen, which is what participants will see in the next phase of the study:

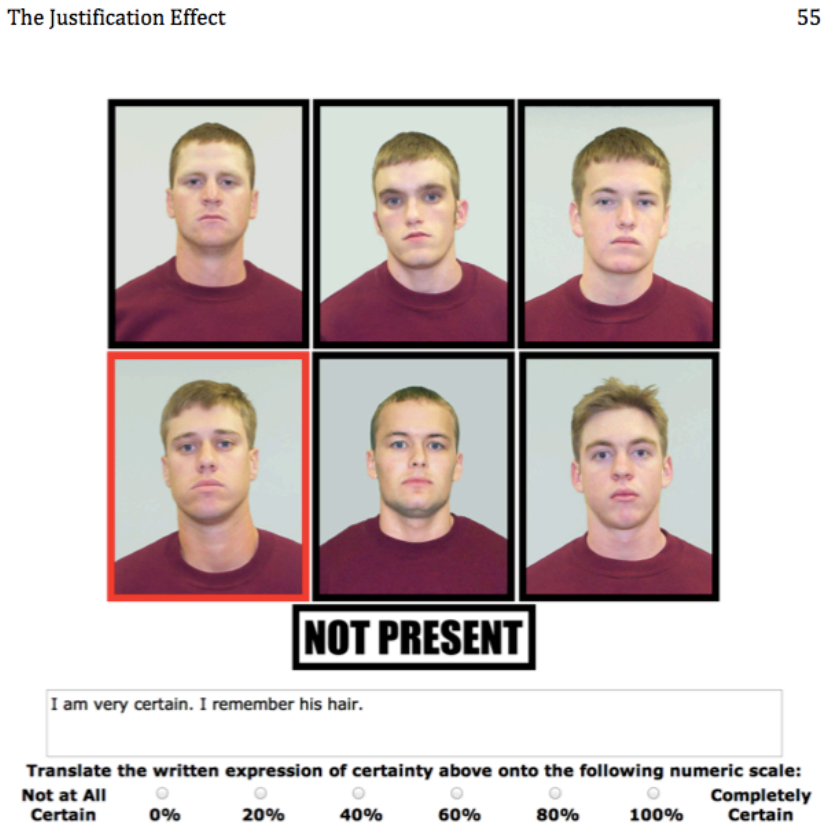


Figure 1. An example of the participant's task in Experiment 1 of translating an expression of certainty about a lineup identification into a number.

The verbal statements here are "I am very certain. I remember his hair." This would be an O1 statement, because "his hair" refers to hair visible in the lineup shown. However, if the statement had read "I am very certain. He was wearing a different shirt before," then this would be a U1 statement, because it refers to a "different shirt" and not one of the red shirts within the lineup.

Additional Examples

combinedText1	**combinedTextCat1**
not very. maybe his eyes.	O1
Somewhat certain, not very confident. He looks familiar, but did not stand out in original set.	F
The ears are making me think it is him. The ears seem to trigger something in me to believe that is him.	O1
very certain. none of the others looked familiar.	F
Not very certain. There are similar looking men wearing the same clothing.	?
Very certain. because he wasn't smiling and i noticed that.	U1
Very certain. I don't recognize any of those faces.	?
im kinda certain. he looks familiar.	F
Not very certain. Looks similar to one I saw.	?
Fairly certain. Don't recognize any of them.	?
Not very sure. This guy looks familiar but it could just be that they all look similar in race and haircut.	M
Fairly Certain. None of these look familiar.	F
mostly certain. I believe this person was in the line up.	?
Fairly certain. I am almost positive that I do not recognize these people, but there's a small chance I'm wrong.	?
A little certain. I believe I recall seeing that specific hair style.	O1
Fairly confident. One face looked familiar however the mouth was not right.	M
pretty sure. don't recognize any of the faces.	?
Pretty Certain. I think I remember.	?
I think I might have seen this guy. I am not sure though. I remember seeing a guy with his eye shape, and his face seemed very familiar, but the more I compare my memory with this person in front of me, the less sure I am.	M
pretty certain. I don't recognize any of these people.	?
fairly. didnt recognize any of them.	?
Somewhat certain due to the eyes and chin. the eyes and chin seem very familiar, but cannot be sure.	M
completely certain. because no one is there from earlier.	?
very certain. i recognize his hair, and the indents in his collar bone.	O2
fairly certain. I remember the facial hair - shirt is different.	M
I am extremely certain I saw this person. I know for a fact I saw this person before.	?
I am fairly certain. the skin tone is familiar.	M
None of the faces look familiar. Because I wasn't certain I didn't make a decision.	F
Fairly certain. The facial expression is quite familiar as well as the facial hair.	M

To Be Done Together

combinedText	combinedTextCat1
I know I saw this man. I remember seeing him in a blue plaid shirt.	
None of these faces look even remotely familiar. I am somewhat certain (I got ahead of myself. sorry sorry. Please switch this response with the previous).	
Not at all. Not sure if any were there.	
fairly. looks the same to me.	
I am pretty sure I haven't seen this photo before. He does not look familiar.	
somewhat. I think it's the same but I'm not positive.	
fairly certain. same nose and eye shape.	
Almost Certain. His shirt color, and his head is tilted to the side slightly.	
Very certain. It's the same face, nose, ears eyes and mouth.	
I am close to completely certain. He has a very rectangular face, and a strong jaw that I remember.	
very certain. I remember his eyes and nose.	
almost certain. I think I remember seeing him.	
not very certain. seeing everyone in the same clothes makes it difficult, along with the same color hair and eyes.	
Very certain. Doesn't look like any of the line up.	
pretty certain. none of the noses or eyes look familiar.	
ehhh kind of. i notice hair more, no hair differences.	
I am pretty certain. The eyebrows and shading around the eyes is familiar.	
I am fairly certain. The face shape looks like one that I saw before, but the shoulders look weird. The eyes are flat; the stubble is sparse and only slightly shaped.	
I am very confident in my selection. I do not recall any red shirts among the lineup.	
pretty certain. i don't recognize any of the faces in the lineup.	
I'm fairly certain. It looks like someone from the previous photos.	
Mostly Certain. Shirt color was not in the lineups.	
very. remember the face shape.	
mildly certain. same eye, hair color and texture.	
Very. I recognize his vacant stare.	
Somewhat certain. I remember a similar face and maybe the haircut.	
Fairly certain. I don't remember any of the faces from before.	
very. I don't remember any of these faces.	
I am fairly certain. None of the people have the hairstyles I recognize or facial features.	
I am very certain. I recognize the hairstyle and facial hair, and the rest of the face matches.	

combinedText	combinedTextCat1
I know I saw this man. I remember seeing him in a blue plaid shirt.	U1
None of these faces look even remotely familiar. I am somewhat certain (I got ahead of myself. sorry sorry. Please switch this response with the previous).	**?**[26]
Not at all. Not sure if any were there.	?
fairly. looks the same to me.	?
I am pretty sure I haven't seen this photo before. He does not look familiar.	F
somewhat. I think it's the same but I'm not positive.	?
fairly certain. same nose and eye shape.	O2
Almost Certain. His shirt color, and his head is tilted to the side slightly.	?
Very certain. It's the same face, nose, ears eyes and mouth.	O5
I am close to completely certain. He has a very rectangular face, and a strong jaw that I remember.	O2
very certain. I remember his eyes and nose.	O2
almost certain. I think I remember seeing him.	?
not very certain. seeing everyone in the same clothes makes it difficult, along with the same color hair and eyes.	?
Very certain. Doesn't look like any of the line up.	?
pretty certain. none of the noses or eyes look familiar.	M
ehhh kind of. i notice hair more, no hair differences.	O1
I am pretty certain. The eyebrows and shading around the eyes is familiar.	M
I am fairly certain. The face shape looks like one that I saw before, but the shoulders look weird. The eyes are flat; the stubble is sparse and only slightly shaped.	O4
I am very confident in my selection. I do not recall any red shirts among the lineup.	?
pretty certain. i don't recognize any of the faces in the lineup.	?
I'm fairly certain. It looks like someone from the previous photos.	?
Mostly Certain. Shirt color was not in the lineups.	?
very. remember the face shape.	O1
mildly certain. same eye, hair color and texture.	O3
Very. I recognize his vacant stare.	O1
Somewhat certain. I remember a similar face and maybe the haircut.	O2
Fairly certain. I don't remember any of the faces from before.	?
very. I don't remember any of these faces.	?
I am fairly certain. None of the people have the hairstyles I recognize or facial features.	U2
I am very certain. I recognize the hairstyle and facial hair, and the rest of the face matches.	O3

[26] This would normally be F, but the subject mentions details that would confuse a new participant rating the statement.