6-20-2022

Improving Witnesses' Predictive Confidence Judgments by Enhancing Test Domain Familiarity

Laura J. Shambaugh
Florida International University, lshambau@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

IMPROVING WITNESSES' PREDICTIVE CONFIDENCE JUDGMENTS BY ENHANCING TEST DOMAIN FAMILIARITY

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PSYCHOLOGY

by

Laura J. Shambaugh

2022

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This dissertation, written by Laura J. Shambaugh, and entitled Improving Witnesses' Predictive Confidence Judgments by Enhancing Test Domain Familiarity, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Jaqueline Evans

_____
Amy Hyman Gregory

_____
Bennett Schwartz

_____
Stephen Charman, Major Professor

Date of Defense: June 20, 2022

The dissertation of Laura J. Shambaugh is approved.

_____
Dean Michael R. Heithaus
College of Arts, Science and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

This dissertation is dedicated to my parents (David and Sharon Shambaugh) and brother (Jonathan Shambaugh), who have unwaveringly supported my academic journey over the years regardless of how far it has taken me from home. Thank you for being my biggest cheerleaders and for always believing in me.

This dissertation is also dedicated to my grandparents, Harold and Evelyn Shambaugh. Thank you for caring enough about my education long before it began to ensure that I would have the funding to attend university. Without your foresight, I may have never discovered the field of eyewitness memory. I love and miss you both, and hope that I have made you proud.

ACKNOWLEDGMENTS

There is a famous African proverb that says, “it takes a village to raise a child.” I would argue that the same is true for completing a dissertation project: such a feat can only occur with the help of others. The list of people who have assisted me in some capacity during the dissertation process is probably longer than this manuscript, but here are a few individuals that I would like to thank in particular:

First, I want to thank my dissertation committee – Dr. Stephen Charman, Dr. Jacqueline Evans, Dr. Bennett Schwartz, and Dr. Amy Hyman Gregory. I appreciate each of your unique perspectives on this dissertation and am grateful for your feedback on its design and execution. A special thanks to Steve, who has served as my primary advisor during my graduate career. Thank you for your support and patience over the last five years, and for teaching me everything I could want to know about lineup research.

Next, I would like to thank two of the most important men in my life: my fiancée Carlos Cortesi, and my cat Theodore (Theo). Carlos: I am forever grateful for your love, encouragement, and ability to make me laugh; at the end of the day, there is no one else I would rather come home to. Theo: thank you for providing the occasional distraction from work with your playful mischief – it was a good reminder for me to take breaks, too.

I also want to thank my fellow legal psychology graduate students and my research assistant Dominica Musiet. To my fellow graduate students: thank you for being there to commiserate with me during this process and for stepping in to help me troubleshoot the occasional hang-up in R. Without your assistance and moral support, I

might not have made it through my Results section. To Dominica: your assistance with the logistics of collecting data in an online two-part study was invaluable.

Finally, thank you to Nostra CrossFit and its phenomenal coaching staff – Gaby, Nikki, Andrea, Cali, David, and Karen. Nothing makes you forget dissertation woes faster than showing up to a WOD programmed with 21-15-9 barbell thrusters and 60-second sprints on the assault bike – but ultimately, these workouts kept me sane by providing a much-needed break from the computer screen (and an outlet for data-related frustration).

ABSTRACT OF THE DISSERTATION

IMPROVING WITNESSES' PREDICTIVE CONFIDENCE JUDGMENTS BY ENHANCING TEST DOMAIN FAMILIARITY

by

Laura J. Shambaugh

Florida International University, 2022

Miami, Florida

Professor Stephen Charman, Major Professor

Recent research on witnesses' pre-identification confidence ("predictive confidence") suggests that these judgments are moderately related to identification accuracy when witnesses experience encoding variability and appropriate statistical techniques are used. However, other research shows that under ecologically valid conditions involving a single identification, the relationship between predictive confidence and accuracy deteriorates. One potential explanation for this lack of a meaningful confidence-accuracy relationship is that witnesses are unfamiliar with the lineup task leading them to underestimate its difficulty. Identification difficulty is partly determined by the similarity of lineup fillers to the suspect, which witnesses cannot anticipate when they make a predictive confidence judgment; in light of this, the current study tested whether witnesses' predictive confidence could be improved by exposing participants to "sample fillers" that matched (or did not match) the similarity of fillers in the actual lineup. The current study also explored whether witnesses' self-reported memory strength predicted their identification accuracy. Finally, to overcome limitations of using continuous measures (such as memory strength or predictive confidence) to

make a dichotomous decision as to whether to show a witness a lineup, the present experiment evaluated whether witnesses' dichotomous judgments about their ability to make an accurate identification decision could predict their subsequent identification accuracy. Witnesses viewed a mock crime under one of eight encoding conditions, and one week later were shown "good", "poor", or no sample fillers prior to reporting their predictive confidence, memory strength, and dichotomous lineup prediction, and attempting to make a lineup identification. Results indicated that viewing good sample fillers did not significantly improve the predictive confidence-accuracy relationship, and that although exposure to either type of sample filler decreased witnesses' predictive confidence, they were largely overconfident relative to their level of accuracy (thereby harming calibration). Witnesses' self-reported memory strength and dichotomous prediction also failed to successfully differentiate accurate and inaccurate eyewitnesses. Results suggest that real-world decisions as to whether to present witnesses with a lineup based on their predictive confidence are misguided. Implications of retention interval on the use of predictive measures are discussed.

Keywords: Confidence, accuracy, calibration, lineup, prediction

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

I. INTRODUCTION

Mistaken eyewitness identifications are the number one contributor to wrongful convictions: recent statistics indicate that they have played a role in approximately 69% of the 375 cases that have been overturned by post-conviction DNA evidence (Innocence Project, 2020). Inaccurate identifications are problematic for two primary reasons: first, they waste time and resources. When a witness selects an innocent person from an identification procedure, investigators' attention becomes diverted from the actual perpetrator. In turn, time and resources are utilized building a case against the innocent suspect while the actual perpetrator continues to walk free. Second, inaccurate identifications are problematic because eyewitness evidence is known to be a compelling factor in court – jurors rely heavily on this evidence when determining a suspect's innocence or guilt (Benton et al., 2006; Brigham & Wolfskeil, 1983; Lindsay et al., 1981). Unfortunately, research has demonstrated that jurors are not always able to accurately conclude whether an identification was reliable (Semmler et al., 2012; Wells et al., 1979). If jurors are presented with inaccurate but highly confident identification evidence, this could contribute to the possibility of a wrongful conviction (Nicholson et al., 2018).

Because eyewitnesses' lineup decisions can strongly influence the course of an investigation, it is important that witness identifications brought as evidence be reliable. In other words, suitable witnesses are those who can successfully discriminate between innocent and guilty suspects in an identification procedure. In a perfect world, law enforcement officers would only permit witnesses who possess this ability to make an identification; this would ensure that the identification obtained (or not) was diagnostic of

a suspect's guilt (or innocence). However, research examining how to reduce false identifications has focused almost exclusively on methods of obtaining eyewitness evidence that involve exposing *all* witnesses to suspect identification procedures (including witnesses whose identification decisions may not be reliable; Charman & Cahill, 2012; Malpass & Devine, 1979; Sporer et al., 1993; Steblay et al., 2003; Wells & Bradfield, 1998; Wixted & Wells, 2017). From an applied standpoint, then, it is important to know whether there is a way of identifying witnesses who are likely to make an incorrect decision in an identification procedure (such as a lineup). This requires researchers to determine whether eyewitnesses can accurately assess the likelihood that they will make an accurate identification decision in the future, and to explore the conditions that may facilitate the reliability of this prospective judgment.

Approaches to Enhancing Witness Reliability

Historically, researchers have attempted to improve the reliability of identifications by either (1) uncovering postdictor variables (i.e., variables associated with the identification decision that are related to eyewitness accuracy), or (2) developing empirically based lineup procedures. Common postdictors of eyewitness accuracy include response latency (quick decisions are more likely to be accurate than slow decisions; Sporer, 1993; 1994), judgment strategy (an absolute judgment is more likely to be accurate than a relative judgment; Dunning & Stern, 1994), memory for lineup fillers (strong memory for lineup fillers is associated with a higher probability of an inaccurate identification; Charman & Cahill, 2012), and post-identification confidence (higher decision confidence indicates a greater likelihood of accuracy; Kassin, 1985; Sporer et al., 1995; Wixted & Wells, 2017). However, although these variables may be helpful in

estimating the reliability of an identification, they all share a common disadvantage in that they are evaluated after a person has already been selected from a lineup, at which point an innocent person is at risk of incarceration – even in the presence of postdictors suggesting an inaccurate identification. Furthermore, postdictors may not be sensitive to biased identification procedures (e.g., lack of unbiased pre-lineup instructions, presence of a non-blind administrator, etc.), resulting in potentially questionable identifications with unreliable postdictors as case evidence.

Consequently, in addition to uncovering postdictors, researchers have developed lineup procedures based in psychological science that maximize eyewitness accuracy. Such procedures give law enforcement an ability to intervene before an identification is made. Such procedural interventions include administration of unbiased lineup instructions (warning the eyewitness that the true perpetrator may or may not be present in the lineup; Malpass & Devine, 1981), using a blind administrator (who is unaware of whether or where the suspect may be in the lineup; Wells & Bradfield, 1998), simultaneous or sequential lineup presentation (the superior method is somewhat debated in the field; Clark, 2005; Steblay et al., 2003), and selecting appropriate lineup fillers (via the match-description method; Fitzgerald et al., 2013).

Even when precautions are taken in lineup construction and administration, however, false identifications may still occur, especially for witnesses with a weak memory of the perpetrator. For instance, weak-memory witnesses are susceptible to making identification errors because of a lowering of their decision criterion (Bornstein et al., 2012; Smith et al., in press). Furthermore, weak memory witnesses are particularly susceptible to influence from extraneous factors such as biased instructions, non-blind

administrator behavior, and post-identification feedback, which can both increase false identifications (in the former two cases) and increase confidence in those false identifications (in all cases; Bradfield & Wells, 1998; Charman et al., 2018; Greathouse & Kovera, 2009). If exposed to one or more of these factors during or after an identification, weak memory witnesses are susceptible to making false identifications with inflated confidence. At trial, an inflated confidence report can have serious consequences as it may distort jurors' perceptions of the witness' reliability (Tenney et al., 2007). To prevent this problem from occurring, it would be advantageous to find a method of improving identification reliability that does not require that all witnesses view a lineup – for example, by using a pre-identification screening tool that provides law enforcement officers with a means of identifying witnesses whose memories may be weak (and who are therefore likely to be inaccurate).

Pre-Identification Confidence as a Screening Tool

Perhaps the most obvious technique available to screen out witnesses who have a weak memory of the perpetrator is to ask the witness their confidence in their ability to later identify the perpetrator from a lineup. However, the utility of witnesses' pre-identification confidence ("predictive confidence"; Nguyen et al., 2018) rests on the critical assumption that witnesses are able to monitor and accurately assess their own memory quality, thus producing a reliable relationship between the predictive judgment and subsequent identification accuracy. Eyewitness researchers have emphasized the potential importance of a predictive confidence-accuracy relationship (Brewer, 2006; Cutler & Penrod, 1989; Wixted et al., 2015), and the literature has provided reasons to believe that this relationship can exist when witnesses experience varied encoding and

appropriate analytic techniques are used (Molinaro et al., 2021; Sauerland & Sporer; 2009; Shambaugh & Charman, in preparation; Valentine & Mesout, 2009). Thus, it is critical for both theoretical and applied reasons to continue exploring this topic. The goals of the present manuscript were therefore (1) to introduce and test a novel technique that is aimed at improving the extent to which predictive confidence is related to subsequent identification accuracy; (2) to evaluate the extent to which an alternative variable – witnesses' metamemory strength – predicts their subsequent lineup accuracy; and (3) to test whether witnesses themselves are able to determine whether they should be shown a lineup by exploring the extent to which a dichotomous predictive confidence judgment predicts their subsequent lineup identification accuracy. Each of these goals are discussed in turn.

Improving the Calibration of Predictive Confidence

From a theoretical perspective, we would expect witnesses' predictive confidence to be related to their subsequent identification accuracy. The question of whether people can accurately predict their future lineup performance is related to basic cognitive research on judgments of learning (JOL). A JOL is a metacognitive prediction of the likelihood that a person will correctly remember studied material in the future (Arbuckle & Cuddy, 1969). In basic JOL research, metacognitive monitoring is assessed by having participants study a set of material (often word lists featuring a target item and paired associates; Rhodes, 2016), and asking them to predict how likely they are to correctly recall the target item in the future. These predictions are made either immediately after studying each item or after a slight delay (ranging from a few seconds to a few minutes).

Although people tend to be overconfident when making general predictions about future performance ability (Cauvin et al., 2018), JOLs can be reliable when made under the appropriate conditions (Fleming & Dolan, 2012; Townsend & Heit, 2011). Most notably, the within-subject utility of JOLs improves when the prediction is made after a slight delay from the time of study (as compared to immediately; Nelson & Dunlosky, 1991; Roediger et al., 1989). The primary benefit of delaying the JOL from the time of study is that doing so causes the assessment to be made under cognitive conditions that are more reflective of the conditions present at test. When a JOL is delayed slightly in relationship to the study phase, a person's memory trace is more reflective of long-term memory (making it more "transferrable" to test circumstances). This is highly relevant for real-world use of predictive confidence, in which there is often a delay between witnessing a crime and being questioned by police (let alone lineup administration). In addition to delay, concreteness of study items (cf. abstract; Tauber & Rhodes, 2012) and memory for prior testing (Finn & Metcalfe, 2007) also affect JOL magnitude and accuracy, such that higher JOLs are assigned for concrete items and items on which learners have previously been tested. Thus, the JOL literature provides valuable evidence that people possess an ability to monitor the impact of encoding on their memory quality.

The notion that people are able to monitor their own memory is also reflected in eyewitness theory. For instance, Semmler et al. (2018) proposed the *constant-likelihood ratio model* to explain how postdictive eyewitness confidence relates to lineup identification accuracy across witnesses. According to the model, witnesses adjust their decision criteria across variations in memory strength to maintain a constant likelihood ratio of correct to incorrect lineup identification decisions for any given confidence level.

Therefore, when witnessing conditions are good and witnesses have a strong memory of the perpetrator, they tend to adopt higher decision criteria than when witnessing conditions are poor, resulting in an equally strong confidence-accuracy relationship regardless of encoding condition. In other words, according to this model, witnesses are appropriately sensitive to variations in memory strength when assessing their post-identification confidence. It is reasonable to assume, therefore, that witnesses' estimations of predictive confidence, too, vary appropriately with variations in their underlying memory strength. As a result, we would expect to observe a significant relationship between witnesses' predictive confidence and their subsequent identification accuracy.

What does empirical literature show regarding the relationship between witnesses' predictive confidence and their subsequent identification accuracy? Before addressing this question, it is important to first discuss the analytic techniques that have historically been used to examine this relationship to have a better understanding of the magnitude of the predictive C-A relationship as it exists in the literature.

Analytic Approaches to Assessing Witnesses' Predictive Confidence

Analytic approaches to studying the reliability of witnesses' predictive confidence judgments have undergone transformation over the years. Initially, researchers relied on point-biserial correlations, but over time, developed more sophisticated analytic techniques such as calibration and confidence-accuracy characteristic (CAC) analyses. Each of these techniques is discussed in turn.

Point-biserial Correlation

Early research examined the predictive confidence-accuracy relationship using point-biserial correlations, in which witnesses' dichotomous accuracy was correlated with their pre-identification confidence. These early studies typically indicated that the magnitude of these point-biserial correlations did not differ significantly from zero, leading researchers to conclude that there was no significant relationship between pre-identification confidence and accuracy (e.g., see Cutler & Penrod, 1989). However, it was later noted that point-biserial correlations are not the most appropriate way to assess the relationship between confidence (whether predictive or postdictive) and accuracy. In particular, Juslin et al. (1996) showed that point-biserial correlations can severely underestimate the true magnitude of a relationship and are thus a poor measure of the confidence-accuracy relationship.

Calibration

Rather than utilizing point-biserial correlations to study confidence, Juslin et al. (1996) suggested the use of calibration analyses, which evaluate witnesses' objective identification accuracy as a function of their subjective probability of accuracy (i.e., their confidence). Calibration analyses were first used to study the relationship between post-identification confidence and accuracy, and address the question "Given that the witness reported X% confidence, what is the likelihood that their identification decision was accurate?" Calibration analyses can also be used to study predictive confidence, in which case the question addressed by the analyses concerns an identification yet to be made (i.e., "Given that the witness reported X% confidence, what is the likelihood that their later identification decision *will be* accurate?").

Calibration necessitates that witness confidence be collected on a ratio scale (with a true zero) in order to plot an appropriate calibration curve. Such scales often range from 0% (*Not at all Confident*) to 100% (*Extremely Confident*). Confidence ratings are then binned into continuous, increasing categories when constructing the plot itself (e.g., 0-20%, 30-40%, 50-60%, 70-80%, and 90-100%), and the observed accuracy rate is plotted at each level of binned confidence by dividing the number of witnesses who made a correct identification decision by the sum of correct and incorrect identification decisions ($\frac{\#\ correct\ IDs}{\#\ correct\ IDs + \#\ incorrect\ decisions}$).

Perfect calibration exists when witnesses' stated expression of confidence corresponds to the percentage of witnesses who make an accurate identification decision at that confidence level (Mickes, 2015). For example, perfect calibration would exist when witnesses who state that they are 90% confident are accurate 90% of the time, accurate 50% of the time when stating 50% confidence, and so forth. Researchers plot the proportion of accurate witnesses in each confidence bin to produce the calibration curve. In a within-subjects analysis, witnesses would provide multiple confidence judgments (e.g., across various mock crimes) allowing for a comparison of changes in calibration within the individual witness. In a between-subjects analysis, witnesses provide one confidence judgment and calibration is compared across witnesses.

Calibration analysis involves the use of three inferential statistics. Calibration (*C*) indicates the extent to which the calibration curve represents perfect calibration, with 0 representing perfect calibration. Over/under-confidence (*O/U*) indicates how far (whether below or above) the calibration curve is from perfect calibration. The *O/U* value can range from -1 (indicating extreme under-confidence) to 1 (indicating extreme over-

confidence). The adjusted normalized resolution index (*ANRI*) is a measure of discriminability indicating how well witnesses' confidence judgments on the calibration curve distinguish correct and incorrect identification decisions. An *ANRI* value of 1 represents perfect discrimination.

In addition to its ability to remedy limitations of point-biserial correlations, calibration was adapted by eyewitness researchers because the question addressed by these analyses more closely maps onto the question that jurors and triers of fact are primarily interested in: What is the accuracy of a witness given that they have provided some level of confidence? However, although calibration analyses are particularly useful for assessing the theoretical relationship between witness confidence and accuracy, they tend to be less useful for measuring the applied value of witness confidence. It is almost always witnesses who identified a suspect that testify in court; therefore, the most applied question involves looking at only decisions made by suspect identifiers. This applied focus on suspect identifications in particular led to the formulation of Confidence-Accuracy Characteristic (CAC) analyses.

Confidence-Accuracy Characteristic (CAC) Analyses

Much like calibration, CAC analyses (Mickes, 2015) involve plotting identification accuracy rates as a function of witnesses' stated level of confidence, but with two differences. First, CAC analyses focus on suspect identifiers (witnesses who selected the target from target-present lineups, or the innocent suspect from target-absent lineups). CAC analyses do not consider witnesses who rejected the lineup or selected a known-innocent filler, but are instead directed at answering the applied question "Given that the suspect was identified with a specific level of confidence, what is the probability

that he/she is guilty?" CAC analyses can also be used to assess the relationship between predictive confidence and identification accuracy, although suspect identifiers can only be determined after the lineup is eventually shown to the witness. In this case, it addresses the question "If the witness identifies the suspect, then given their stated level of predictive confidence, what is the probability that the suspect is guilty?"

Second, in contrast to calibration, CAC does not require a ratio scale; rather, any ordinal scale will suffice (e.g., "low", "medium", and "high" confidence). Common confidence cutoffs observed in the literature are 0-60% (low), 70-80% (medium), and 90-100% (high; see Mickes, 2015 for an overview). For each confidence bin, accuracy (*A*) is computed as $\frac{\#\ correct\ suspect\ IDs}{\#\ correct\ suspect\ IDs + \#\ incorrect\ suspect\ IDs}$.

Predictive Confidence in the Eyewitness Literature

Although both theory and research on JOLs provides reasons to believe that a predictive confidence-accuracy relationship can exist, empirical findings from the eyewitness literature have been less promising, with the bulk of studies finding no or only a weak association (Cutler & Penrod, 1989; Hourihan et al., 2012; Nguyen et al., 2018; Whittington et al., 2019).

Cutler and Penrod (1989)

In this meta-analysis, authors evaluated nine empirical eyewitness studies that included some measure(s) of witnesses' pre-lineup and post-lineup confidence. Of primary interest were the various point-biserial correlations between pre-lineup confidence, post-lineup confidence, and accuracy. Results indicated that in five of the nine studies, witnesses' post-lineup confidence had a significantly stronger relationship with accuracy than did pre-lineup confidence. Overall, the authors concluded that

confidence measures collected after the lineup were stronger predictors of identification accuracy than pre-lineup confidence, and that the relationship of pre-lineup confidence with accuracy was trivial (mean r = .10). Ultimately, they discouraged consideration of witnesses' pre-lineup confidence when deciding whether they should attempt an identification.

Hourihan et al. (2012)

Hourihan et al. (2012) used a standard recognition paradigm coupled with a judgment of learning to assess witnesses' metamemory and memory judgments for own-race and other-race faces. In this study, researchers recruited Caucasian and Asian participants. Half of the participant pool was exposed to same-race faces, and the other half were exposed to other-race faces. Each group studied an equal number of photographs (25); after viewing each face, participants rated the likelihood of accurately recognizing the face during a later testing period. These scores were then analyzed in a 2 x 2 mixed ANOVA (with face as the repeated-measures factor and participant race as the between-subjects factor). Caucasian subjects displayed greater JOL accuracy (as indicated by planned comparisons) for Caucasian faces compared to Asian faces. Likewise, Asian participants had greater JOL accuracy for Asian faces than Caucasian faces, though this was not statistically significant. The authors concluded that relative metamemory accuracy was greater for faces of the participants' own race compared to those of another race, but the predictive confidence-accuracy relationship was still quite weak overall.

Nguyen et al. (2018)

Nguyen et al. (2018) tested the relative utility of predictive and postdictive confidence in witnesses' estimations of discriminating between same- and cross-race faces in two experiments. Twenty faces (10 White and 10 Black) were shown to participants one at a time; after viewing each face, participants rated (either immediately or after a 30-second delay) how likely it was that they would later accurately recognize the face that they just studied in increments of 20% (0, 20, 40, 60, 80, and 100). Approximately one minute after the last face was studied, participants engaged in the test phase (an old/new recognition memory test that included 40 faces – 20 old, 20 new).

In their analyses, Nguyen et al. (2018) conducted a series of 2 x 2 ANOVAs to determine how judgment delay and face race affected identification accuracy at three levels of predictive confidence: *low* (0% and 20%), *medium* (40% and 60%), and *high* (80% and 100%). In their results, they focused on the high predictive judgment group (as these witnesses are most likely to be asked to attempt an identification; Cutler & Penrod, 1989). Amongst witnesses who reported a high level of predictive confidence (80% or 100%), the effect of judgment delay was consistent across experiments: delayed predictive confidence judgments were associated with a higher proportion of identification accuracy compared to immediate judgments. Across experiments, the proportion of accurate identifications was higher for same-race faces than cross-race.

The authors also examined the relative utility of predictive to postdictive confidence. They found that even at the highest level of predictive confidence, witnesses' overall identification accuracy was still objectively low (M = 64%). However, witnesses reporting the highest level of postdictive confidence (80% and 100%) were much more

accurate (*M* = 91%). Thus, postdictive confidence was a more reliable predictor of identification accuracy compared to predictive confidence – one of the key takeaways from this study.

Whittington et al. (2019)

Most recently, Whittington et al. (2019) tested the predictive confidence-accuracy relationship in two experiments. In Experiment 1, they used a multiple-block lineup recognition paradigm in which participants were exposed to a series of faces and houses (on a green backdrop) with the target face (on a red backdrop) embedded somewhere amidst this series. After viewing all items, participants completed a yes/no recognition test for filler items followed by a target-present or target-absent lineup. In addition, participants provided a confidence rating (1-10) either before viewing the lineup, after viewing the lineup, or both before and after the lineup (indicating the likelihood of making a correct identification decision).

In their analyses, the authors constructed calibration curves and calculated two of the standard inferential statistics (*C* and *ANDI* [an estimate of resolution similar to *ANRI*]). Results of Experiment 1 demonstrated that witnesses' post-lineup confidence was more strongly calibrated with accuracy than witnesses' pre-lineup confidence. The authors also noted that predictive confidence not only underperformed postdictive confidence, but it actually appeared to harm the postdictive confidence-accuracy relationship: the postdictive confidence-accuracy calibration was weaker when witnesses had also reported predictive confidence (compared to when they reported only postdictive confidence).

In Experiment 2, researchers tested predictive confidence with a mock witness paradigm. Participants viewed a video depicting a male perpetrator stealing a cellphone from an unlocked vehicle. They were then asked to report their confidence (0-100%) either before and after viewing a target-absent or -present lineup, or only after (Experiment 2 did not have a pre-identification-only condition). Postdictive confidence (taken from the Pre/Post condition) exhibited the best calibration with accuracy, followed by the Post-only condition, and finally predictive confidence (taken from the Pre/Post condition). As in Experiment 1, predictive confidence was outperformed by postdictive confidence in both calibration and resolution. However, across experiments, authors did note that the predictive confidence-accuracy relationship *was* strong at the highest levels of confidence (consistent with other research; Semmler et al., 2018; Shambaugh & Charman, in preparation).

In sum, the reviewed literature tends to show a weak or negligible relationship between witnesses' predictive confidence and accuracy. There are, however, two studies that have demonstrated an association between predictive confidence and accuracy. Sauerland and Sporer (2009) conducted a study in which participants (passers-by) were asked for directions to a particular location by one of ten confederate experimenters (the "target"). This interaction lasted approximately 15-60 seconds. Afterward, participants were approached by an interviewer (a second experimenter) who gave them a questionnaire asking for an indication of their pre-identification confidence (0-100%). The interviewer then gave participants unbiased lineup instructions and presented them with either a target-present or target-absent lineup. Across participants, calibration analyses indicated that predictive confidence was reliably related to subsequent accuracy:

participants asserting high predictive confidence (M = 64%) were more accurate than witnesses exhibiting low predictive confidence (M = 57%).

That same year, Valentine and Mesout (2009) exposed mock witnesses to one of 50 different "scary" targets amidst a labyrinth. After the encounter, witnesses completed various anxiety measures, questions regarding their memory for the target, and reported confidence in their ability to accurately identify the target. They next received unbiased instructions and attempted an identification from a target-present array. As with the Sauerland and Sporer study, Valentine and Mesout found that predictive confidence was reliably related to accuracy (r = 0.39) across witnesses.

What might account for observed differences between these two studies and the earlier studies that found a weak or negligible relationship? Although this question is difficult to answer given that there are various methodological differences across predictive confidence studies, one of the most notable differences is there was more encoding variability across witnesses in the latter two studies. In both studies that found a reliable predictive confidence-accuracy relationship, witnesses were exposed to a variety of different targets under non-uniform viewing conditions. In contrast, studies finding a weak relationship utilized paradigms in which witnesses were exposed to the same target under the same viewing conditions (Whittington et al., 2019, Experiment 2), or to a series of still images that also failed to vary encoding (Hourihan et al., 2012; Nguyen et al., 2018; Whittington et al., 2019, Experiment 1). This observation led researchers to take a closer look at the role of encoding variability in more recent predictive confidence studies.

Encoding Variability and the Predictive C-A Relationship

Recent work by Molinaro et al. (2021) provides compelling evidence that properly inducing encoding variability across witnesses may be one of the keys to producing a reliable predictive confidence-accuracy relationship. They make two arguments for the importance of varied encoding. First, they contend that exposing all witnesses to the same mock crime and perpetrator under the same viewing conditions may artificially weaken the predictive confidence-accuracy relationship because it restricts the range of predictive confidence values elicited from witnesses. Second, encoding variability is important for real-world legal practitioners; these individuals are more interested in knowing how predictive confidence performs *across* witnesses who have undergone various degrees of encoding (as no two eyewitnesses will have had the *exact* same experience of a crime).[1]

In their study, Molinaro et al. exposed witnesses to eight different mock crimes under varying encoding conditions. After each mock crime, witnesses provided their predictive confidence and also answered secondary memory questions about their memory clarity, viewing quality, ability to make out specific features of the perpetrator's face, and other measures before moving on to the next mock crime. As hypothesized, results revealed a negligible relationship between predictive confidence and accuracy within individual encoding conditions. However, a significant relationship *was* produced when analyses were aggregated across viewing conditions.

[1]In the memory literature, "encoding variability theory" (see Johnston & Uhl, 1976) refers to variation in how information is encoded, particularly differences in learners' cognitive environments. The proposed study uses the term "encoding variability" to instead refer to differences in the viewing quality (viewing distance and duration).

For calibration analyses, witnesses' confidence was binned into five groups (0-20%, 30-40%, 50-60%, 70-80%, and 90-100%). Choosers were significantly better calibrated compared to non-choosers: whereas choosers who gave higher predictive confidence ratings were more likely to make a correct target identification compared to choosers who provided lower predictive confidence ratings, no such relationship was found for non-choosers. Importantly, CAC analyses revealed that amongst suspect identifiers, predictive confidence was good at predicting subsequent accuracy (equitable to postdictive confidence – something not found in previous studies).

For this analysis, confidence was binned into three categories: low (0-60%), medium (70-80%), and high (90-100%). Witnesses asserting high predictive confidence displayed the same proportion of identification accuracy (98%) as witnesses asserting high postdictive confidence (98%); these proportions were not significantly different. The accuracy rate of medium predictive confidence witnesses (96%) also did not significantly differ from that of medium postdictive confidence (95%). Finally witnesses who provided a low predictive confidence statement did not significantly differ in accuracy from witnesses providing a low postdictive confidence statement (89% and 87%, respectively).

Molinaro et al. (2021) made significant progress in our understanding of the predictive confidence-accuracy relationship, particularly the role of witness encoding. However, their conclusions are limited for two reasons. First, they used a within-subjects methodology. Each witness completed eight trials in which they viewed a mock crime, made a predictive confidence judgment, and were then exposed to a lineup. Thus, there is a possibility that the observed results were an artifact of participants' practice at making multiple pre-identification judgments across trials; as they progressed through the study,

they may have learned to differentiate between encoding conditions that produced a weak versus strong memory trace (making subsequent predictive confidence judgments more calibrated). JOL research has shown a similar pattern wherein participants' predictive judgments become better calibrated with their accuracy across trials (Rhodes, 2016). Second, their witnesses were shown a lineup immediately after providing a predictive confidence judgment. In the real world, however, there is frequently a significant delay between providing a predictive confidence judgment and subsequent lineup task. Therefore, in response to these limitations, Shambaugh and Charman (in preparation) conducted a follow-up study using a between-subjects design to test the performance of predictive confidence under more ecologically valid conditions in which there is a delay between the predictive confidence judgment and lineup task.

Shambaugh and Charman (In Preparation)

Shambaugh and Charman (in preparation) manipulated the delay periods between crime encoding and predictive confidence judgment, and predictive confidence judgment and lineup task. Witnesses (N = 885) were exposed to one of eight different mock crime versions that varied in terms of encoding quality (10 versus 60 seconds in length, and about 5 versus 15 feet in viewing distance). Witnesses were then randomly assigned to either make an immediate predictive confidence judgment and an immediate lineup identification, an immediate predictive confidence judgment and a delayed lineup identification (one week later), or a delayed predictive confidence judgment and a delayed lineup identification (both one week later).

Witness calibration was binned into five categories in the same manner as Molinaro et al. (0-20%, 30-40%, 50-60%, 70-80%, and 90-100%). Researchers found

reliable calibration (C = 0.09) amongst witnesses who made an immediate predictive confidence judgment followed by an immediate lineup identification (replicating Molinaro et al., 2021). However, when a one-week delay was introduced (either between the event encoding and the predictive confidence judgment, or between the predictive confidence judgment and lineup identification), the relation was harmed. Across the board, witnesses were severely overconfident in their future memory ability, though witnesses in the Immediate/Immediate condition less so than witnesses in either the Immediate/Delayed condition or Delayed/Delayed condition. Similar results were obtained with a CAC analysis.

Shambaugh and Charman (in preparation) theorized that witnesses' global judgments about their likelihood of future accuracy should tend to be decoupled from their actual task performance when either (a) the predictive confidence judgment is not based on the same underlying memory as the identification decision (which explains why they observed a stronger relationship when the identification task was obtained immediately after the predictive confidence judgment), or (b) to the extent that predictive confidence judgments are unable to incorporate relevant information about the lineup task. Concerning this latter point, eyewitnesses may hold inappropriate beliefs about the difficulty of the lineup task due to a lack of experience with lineups. Most people have not been eyewitnesses before when they are asked to make an identification, and therefore the identification process is novel. Critically, prior to actually viewing the lineup, it is impossible for witnesses to anticipate the quality of choice alternatives (fillers) which affect the difficulty of the task (Fitzgerald et al., 2013). To the extent that witnesses over- or underestimate the difficulty of the future lineup task when making

their predictive confidence judgments, these judgments will not strongly predict their actual accuracy.

Given that witnesses in Shambaugh and Charman (in preparation) exhibited overconfidence (noted by the magnitude of the *O/U* statistic in each delay condition), they clearly underestimated the difficulty of the lineup task. This pattern of witness overconfidence suggests important new avenues of research. For instance, can witness overconfidence be reduced (and therefore the calibration of witnesses' predictive confidence judgments be improved) if they are given a way to more accurately anticipate the difficulty of the lineup task? Based on the metacognitive monitoring literature, giving witnesses a way to "preview" the difficulty of a lineup judgment may be an important means of improving the predictive confidence-accuracy relationship.

Problem Characteristics and the Accuracy of Metacognitive Monitoring

In the metacognition literature, *problem characteristics* refer to characteristics specific to the memory task at hand (its demands and complexity) that can affect metacognitive monitoring. Examples of problem characteristics include whether test items are learned actively or passively (better monitoring accuracy occurs when items are actively generated during study rather than passively read; Mazzoni & Nelson, 1993), task difficulty (monitoring accuracy tends to decrease as task difficulty increases; Lichtenstein et al., 1982; Suantak et al., 1996), and testing experience (monitoring accuracy improves following a practice test with the material; Glenberg et al., 1987; King et al., 1980; Lovelace, 1984; Shaughnessy & Zechmesiter, 1992).

Perfect calibration of metacognitive judgments and test performance rarely exists, especially when a person has low familiarity with the testing domain and the task is

moderately or extremely difficult (Glenberg et al., 1987; Glenberg & Epstein, 1985; 1987; Keren, 1991). In fact, task difficulty is one of the primary determinants of people's ability to accurately predict (or not) their memory performance (Schraw & Roedel, 1994). According to the "hard/easy effect" (Bjorkman, 1992; Lichtenstein et al., 1982), learners have an increased likelihood of exhibiting overconfidence in prospective judgments as a task becomes more difficult (Keren, 1991; Newman, 1984) because even though actual performance accuracy declines with task difficulty, people expect to do equally well as they would on an easier task; they fail to adjust expectations of their performance (Schraw & Roedel, 1994). Recognition tests (compared to recall) may be especially likely to produce this pattern, as recognition tests can inflate people's feeling of material mastery (Ghatala et al., 1989).

Despite the potential harm task difficulty can have on monitoring accuracy, there are ways to reduce it or prevent it from occurring. Research has indicated that increasing a learner's prior knowledge in a domain improves prospective judgment calibration (i.e., by reducing over- or underconfidence). Prior knowledge and experience in a domain serve as a basis to help the learner appropriately adjust their performance expectations (Glaser & Chi, 1988). Additionally, previous experience yields more automated problem solving, thus leaving more cognitive resources available for performance monitoring (Allen & Casbergue, 1997). When attempting a novel task, people lack sophisticated schemas to provide information about task difficulty and their personal performance. Exposing novice learners to additional task information may substitute for the missing schemas and provide them with important insight into how they will ultimately perform on the subsequent task (Kalyuga et al., 2003). Empirical lab studies have demonstrated

that perceived readiness for testing improves metacognitive monitoring (Pressley et al., 1987) and increasing a learner's familiarity with a test improves monitoring accuracy (Nietfeld & Schraw, 2002). JOL research also attests to the benefit of prior domain knowledge on prospective judgments. For instance, expert chess players (cf. novices) have higher JOL accuracy (measured by rating their confidence in accurately predicting a particular set of chess moves in the future), and they perform better in a chess endgame simulation (De Bruin et al., 2007). Additionally, studying worked examples and solving practice problems improve JOL accuracy amongst elementary students and adolescents (Baars et al., 2014; 2016).

Given these findings, it stands to reason that having prior knowledge about a lineup task may benefit eyewitnesses' predictive confidence judgments by making them less overconfident about their future lineup performance. Thus, the first goal of the current study was to determine whether the calibration of witnesses' predictive confidence judgments can be improved by providing witnesses a means of anticipating the difficulty of the lineup task. Identification difficulty is in part determined by the similarity of the various lineup fillers to the suspect (Wells et al., 2020). However, before viewing a lineup, witnesses cannot know how similar these fillers will be to the suspect. This leads to an important applied research question: can exposure to 'sample' fillers prior to making a predictive confidence judgment help witnesses adjust their performance expectations by leading them to better anticipate the difficulty of the task?

Based on findings from metacognition and JOL research, exposure to sample fillers should improve witnesses' calibration. However, this prediction comes with an important caveat: the sample fillers must be *good* fillers (i.e., as well-matched to the

suspect as fillers in the actual lineup). Predictive confidence queries witnesses' discrimination ability – viewing well-matched sample fillers should help weak-memory witnesses realize that they cannot discriminate between their memory trace of the suspect and the photos in front of them, in turn making them aware that they are also likely to experience similar difficulty during the actual lineup task. As a result, exposure to good sample fillers should aid calibration by reducing confidence. However, if the innocent filler photos shown are poor (i.e., not well-matched to the suspect) then such exposure is unlikely to provide any benefit to prospective confidence calibration (because the sample fillers do not represent the actual lineup task whose difficulty will ultimately determine the witness' accuracy). The first goal of the current study was therefore to test whether exposing witnesses to high quality sample fillers prior to providing a predictive confidence judgment improves the calibration of those judgments with subsequent lineup identification accuracy.

Using Metamemory as a Predictor of Identification Accuracy

Predictive confidence, however, is not the only measure that can be used to predict subsequent lineup identification accuracy. In recent predictive confidence studies, researchers have also examined the relationship between witnesses' self-reported memory strength and subsequent lineup identification accuracy (Molinaro et al., 2021; Shambaugh & Charman, in preparation). Objective memory strength may be a more reliable means of predicting accuracy compared to confidence because it overcomes two important limitations of predictive confidence. First, whereas predictive confidence involves incorporating beliefs about lineup difficulty (which are often incorrect), assessments of objective memory strength tap directly into the underlying construct that

drives identification accuracy. Second, predictive confidence typically asks witnesses to determine the likelihood of making an accurate identification from a target-present lineup, *or* the likelihood of accurately rejecting the lineup in the event that the perpetrator is absent. If witnesses are asked only one of these questions, their answer may not be very applicable to their subsequent lineup task (depending on whether the lineup is target-present or target-absent; the witness' answer addresses only one of these two scenarios). If witnesses are asked both questions, it still leaves the problem of determining which question should be used to predict accuracy; because in the real world it is unknown whether the lineup contains the actual perpetrator, it is unclear which measure should be used to predict their accuracy. Objective memory strength measures, however, get around this issue because it does not matter whether witnesses are shown a target-present or target-absent lineup – the ratings of memory strength, clarity of view, attention paid, etc. are applicable regardless.

Findings from Molinaro et al. (2021) and Shambaugh and Charman (in preparation) support this assertion. Molinaro et al. performed multilevel logistic modeling and found that witnesses' responses to the secondary memory questions greatly improved the ability to predict lineup accuracy beyond predictive confidence alone – that is, the model including both witnesses' metamemory and witnesses' predictive confidence as factors was a much better fit compared to the model with only predictive confidence as a factor. Results of Shambaugh and Charman reflected findings from Molinaro et al.: witnesses' metamemory scores appeared to serve as more reliable predictors of lineup accuracy than predictive confidence. Importantly, this pattern was maintained in the face of delay (though only trending and it did not reach statistical significance; $p = .165$).

Shambaugh and Charman (in preparation) theorized that the potential predictive superiority measures of metamemory (cf. predictive confidence) in their delay conditions may have partly been an artifact of how witnesses were binned into memory strength groups for analyses. Witnesses' composite memory scores were calculated by adding their *Likert* scale ratings on four encoding-related questions (view quality, length of time perpetrator's face was in sight, ability to make out specific features of the perpetrator's face, and how much attention the witness paid to the perpetrator) and dividing this sum by four. The composite scores (somewhere between 1 and 7) were then binned into roughly equal thirds with scores 1-3 indicating "weak memory", 3.25-5 indicating "moderate memory", and 5.25-7 indicating "strong memory". However, this binning is arbitrary; there are a number of ways witnesses could be binned into various memory strength categories (which could alter the extent to which memory strength predicts accuracy). Thus, the second goal of the current research was to continue exploring the use of metamemory measures (e.g., memory strength, quality of view, etc.) as predictors of subsequent identification accuracy.

Exploring a Dichotomous Predictive Judgment

Even if we did find a significant relationship between predictive confidence and accuracy, or between memory strength measures and accuracy, deciding whether to present witnesses with a lineup based on continuous measures (such as predictive confidence or memory strength measures) involves an arbitrary decision as to what the appropriate cut-off point is. Real world policing involves a dichotomous decision about whether to show a witness a lineup; it is unclear how best to make that decision using continuous measures. It is therefore useful to test various methods of imposing a cut-off

point to dichotomize witnesses into groups theoretically representing those who are more versus less likely to be accurate. In the present study we tested three such possibilities. First, we split witnesses into dichotomous groups using the median predictive confidence and composite memory values in each sample filler condition. Second, we grouped witnesses by identifying the predictive confidence and composite memory values that maximally differentiated witness accuracy in each sample filler condition. However, both methods have limitations: median split relies on a distribution of witness responses which police officers would not have for a single-witness crime. A maximally differentiating point would likely vary between witnesses, crimes, viewing conditions, (etc.), making it difficult to generalize. Due to the practical shortcomings of using either a median split or a maximally differentiating point, we also tested a third method of dichotomizing witnesses: Allowing witnesses to determine themselves whether they will be able to make an accurate lineup identification decision. In other words, would witnesses' responses to a simple dichotomous yes/no question as to whether or not they believed they could make an accurate identification decision if shown a lineup predict lineup identification accuracy?

The third goal of the present study was therefore to explore whether witnesses themselves can appropriately determine whether they will be able to make an accurate lineup identification by asking them a simple dichotomous question about whether they will make an accurate decision or not. A dichotomous judgment directly meets the demands of real-world policing in which the decision to show a witness a lineup is in and of itself dichotomous.

Present Study: Hypotheses

There were four hypotheses for the present study: first, based on recent eyewitness research on the predictive confidence-accuracy relationship, we expected to observe a significant relationship between predictive confidence and identification accuracy among witnesses who experience varied encoding conditions. Second, we hypothesized that exposure to high quality sample filler photographs prior to making a predictive confidence judgment would improve calibration by reducing witnesses' confidence (cf. exposure to poor sample fillers or no exposure to sample fillers). Third, we hypothesized that witnesses' metamemory scores would be less sensitive to the filler manipulation, as witnesses' impressions of task difficulty should not influence their judgments of memory strength, viewing experience, etc. As a result, witness metamemory should predict subsequent lineup accuracy in all three filler conditions. Fourth, we hypothesized that, assuming witnesses have the ability to monitor their own memory, witnesses who responded "yes" to the dichotomous lineup performance question would exhibit greater accuracy compared to witnesses who responded "no". Furthermore, we expected the magnitude of this effect to be greater among witnesses who were shown high quality sample fillers (cf. poor quality sample fillers or no sample filers), as high-quality fillers should reduce witnesses' chronic overconfidence.

II. METHOD

Participants

Adult college students 18 years and older (N = 411) were recruited from two large southeast universities. Participation was completed on a voluntary basis in exchange for course credit. Informed consent was obtained prior to interaction with any study

materials, and the consent document informed participants that both Part 1 and Part 2 of the study were obligatory to receive full compensation. All participation was completed online (remotely) using the Qualtrics Survey System.

Design

The current study conformed to a 3 (Sample Filler Exposure: None vs. Poor vs. Good) x 2 (Lineup: Target-present vs. Target-absent) between-subjects design. Witnesses' encoding conditions were manipulated for generalizability purposes but were not analyzed as an independent variable.

Stimuli

Mock Crime Videos

Mock crime videos were borrowed from Molinaro et al. (2021). These videos depicted either a male perpetrator graffitiing a building, or a different male perpetrator committing a carjacking. Each video was shot from a close or far viewing distance (about 5 versus 15 feet, respectively), and for a long or short viewing duration (10 versus 60 seconds, respectively). Therefore, participants were exposed to one of eight possible mock crime variations (serving as our encoding variability induction).

Lineups

Lineup photographs were also borrowed from Molinaro et al. (2021). All lineups consisted of six color photographs presented simultaneously. Lineup photos were edited to display only colored images of the lineup members' heads against a blank (white) background. For each mock crime, there was one target-absent lineup and six target-present lineups (whereby the target replaced each filler in turn). Filler photographs (6) were selected using the match-to-description method, and all lineup photos (7) were

tested by Molinaro et al. using a group of students (separate from the study sample) to estimate the effective size of the target-absent (6 photos) and target-present (7 photos) lineups for each crime (Tredoux's *E*; Malpass & Lindsay, 1999; Tredoux, 1998). For the graffiti mock crime, the effective size of the target-present lineup was $E = 3.56$ [2.99, 4.41], and for the target-absent lineup $E = 4.65$ [3.73, 6.18]. For the carjacking, the effective size of the target-present lineup was $E = 4.76$ [3.77, 6.46], and for the target-absent lineup $E = 4.29$ [3.26, 6.27]. See Appendices E and F for sample lineups for the graffiti and carjacking mock crimes, respectively.

Sample Filler Photographs

Sample filler photographs were selected from the New Jersey and Kentucky criminal offender databases available on the Government Documents Roundtable webpage (https://godort.libguides.com/prisonerdbs). For each mock crime, twenty "good" filler faces were selected using the match-description method (Wells et al., 1993). These descriptions were generated by undergraduate pilot participants from Molinaro et al. (2021). Twelve "poor" filler faces were selected based on matched sex and race to the perpetrator.

The selected filler faces underwent two rounds of pilot testing. In the first round, participants ($n = 16$) rated the similarity of each filler photograph to a photograph of the respective perpetrator. The eight faces with the highest photo-similarity rankings for each mock crime were considered "good" sample fillers. The bottom six faces with the lowest photo-similarity rankings for each mock crime were considered "poor" sample fillers. In the second round of pilot testing, two new sets of participants (graffiti $n = 40$; carjacking $n = 35$) rated the similarity of each good/poor sample filler and each actual lineup filler to

the photograph of the respective perpetrator. We then used the top six most-similar faces (serving as the "good" fillers) and poor fillers to run a series of paired samples t-Tests. See Table 1 for mean similarity ratings for the finalized sets of sample fillers. The second round of pilot testing allowed us to test whether the "good" sample fillers for each crime were as similar to the perpetrator as the actual lineup fillers, and whether the "poor" sample fillers for each crime were sufficiently less similar to the perpetrator than the actual lineup fillers. See Appendices C and D for the graffiti and carjacking sample filler lineups, respectively.

Graffiti. For the graffiti mock crime, the good sample fillers (M = 3.33, SD = 1.32) were rated equally-similar to the perpetrator as the actual lineup fillers (M = 3.22, SD = 1.22), $t(39) = 1.10$, $p = .274$. The good sample fillers were rated significantly more similar to the perpetrator than the poor sample fillers (M = 1.64, SD = 0.65), $t(39) = 9.32$, $p < .001$. Finally, the poor sample fillers were rated significantly less similar to the perpetrator than the actual lineup fillers, $t(39) = 9.61$, $p < .001$.

Carjacking. For the carjacking mock crime, the good sample fillers (M = 3.15, SD = 1.21) were rated significantly less similar to the perpetrator than the actual lineup fillers (M = 3.53, SD = 1.19), $t(34) = 3.61$, $p = .001$. However, the good sample fillers were rated significantly more similar to the perpetrator than the poor sample fillers (M = 1.45, SD = 0.49), $t(34) = 9.67$, $p < .001$. Finally, the poor sample fillers were rated significantly less similar than the actual lineup fillers, $t(34) = 11.22$, $p < .001$.

Metamemory Measures

The present study contained three primary metamemory measures of interest: predictive confidence (both in the likelihood of making an accurate identification and the

likelihood of making a correct rejection), secondary (metamemory) memory judgments, and a dichotomous lineup judgment prediction. Question order was randomized across participants (with the constraint that the two predictive confidence questions were always answered together, but in a different order).

Predictive Confidence. When asked to report predictive confidence, participants were told: "In a moment, you will be shown a lineup containing six photographs of people who may or may not have committed the crime you witnessed. Please think carefully and answer the following questions." They were then asked two questions: (1) "If the criminal who committed the crime you witnessed was among those photos, how likely would you be to correctly identify that person?" and (2) "If the criminal who committed the crime you witnessed was not among those photos, how likely would you be to correctly indicate that they are not there?" Both questions were rated on a 0% (*Not at all confident*) to 100% (*Extremely confident*) scale increasing in 10% increments.

Secondary Memory Judgments. Witnesses also reported secondary memory judgments including the amount of attention they paid to the criminal, the quality of view they had, their ability to make out specific features of the criminal's face, how long the criminal was in view, how strong their memory is, and how clear their memory was. These items were rated on a *Likert* scale ranging from 1-7, with 1 indicating no attention, an extremely poor view, extremely weak memory (etc.) and 7 indicating complete attention, an extremely good view, extremely strong memory (etc.). See Appendix A.

Dichotomous Lineup Judgment. Participants were asked to make a dichotomous decision as to whether they believed they could make an accurate identification decision. This measure was meant to approximate what real-world witnesses may be asked by

investigators. Specifically, witnesses were asked: “If you were shown a lineup of six people that may or may not contain the criminal who committed the crime you witnessed, do you believe that you could make an accurate identification decision?” (to which participants were given the option to respond yes or no).

Procedure

Participants enrolled in “Perceptions of Strangers II”, a two-part study that was reportedly interested in evaluating how people’s personality disposition and situational factors interact to affect their perceptions of unfamiliar others. Respective parts of the study were completed one week apart. After providing informed consent in Part 1, participants completed the first half of a series of personality measures resembling BuzzFeed quizzes and personality questionnaires. Completion of these measures took approximately five minutes. At this time, participants were randomly assigned to view one of eight mock crime variations. After viewing the mock crime participants were informed that they would receive an email link to complete Part 2 at a later point in time, answered a manipulation check question (“What crime was the criminal committing in the video you watched?”), and Part 1 ended.

One week later, participants were sent a link to complete Part 2 of the study. At the outset of Part 2, participants were reminded of the study’s general purpose and then they completed the second half of the “personality measures” (lasting approximately five minutes). Next, participants were reminded that they witnessed a crime during the first part of the study and were told that they would be shown a lineup momentarily. At this time, they were randomly assigned to a sample filler condition (to enhance familiarity with the testing domain or not): good, poor, or none (control). Participants in the good

and poor filler exposure conditions were shown a lineup of filler photographs accompanied by the following statement:

> "The photo array (below) contains a lineup of individuals who are **known to be innocent** (in other words, did **NOT** commit the crime that you witnessed), but **resemble the faces you will see** in the actual lineup. Please **take a moment to look at these photos** to get a sense of the task at hand. You will be moved to the next page automatically after 30 seconds."

After viewing the sample fillers, the survey proceeded to the predictive confidence judgments, metamemory measures, and dichotomous lineup question. Participants in the no-sample filler condition went straight to reporting the various memory measures after the lineup warning.

After all prospective judgments were made, participants were given unbiased pre-lineup instructions (warning them that the criminal may or may not be present) followed by a target-present or target-absent lineup (randomly assigned). They were given the option to identify one of the six lineup members, or to indicate that the criminal was "not present." Following the identification attempt, participants rated provided responses to the secondary memory judgments (post-identification confidence, difficulty of identification, willingness to testify, etc.; see Appendix B) and answered two manipulation check questions ("For this question, please select '3'", and "Select 5 for this question"). Finally, participants were debriefed regarding the true nature of the study.

III. RESULTS

Finalized Dataset

For inclusion in analysis, participants must have completed both parts of the study and correctly answered the three manipulation check questions. In total, complete data were collected from 471 participants who completed both Part 1 and Part 2. Of these, 31

individuals failed the manipulation check question in Part 1 and an additional 29 individuals failed one or both manipulation check questions in Part 2. Thus, the finalized dataset included 411 participants. Participants' average age was 22.25 years (*SD* = 5.68), and they were majority female (82%) and Latin/x (69%).

Encoding Manipulation Check

Encoding was not classified as an independent variable in the current study, but participants were exposed to different mock crime durations (10 vs. 60 seconds) and viewing distances (close vs. far) to produce encoding variability. To determine whether the manipulation of encoding duration was successful, we ran a series of 2 (Distance: Close vs. Far) x 2 (Duration: Long vs. Short) between-subjects ANOVAs on participants' perceptions of how long the criminal was in view. To assess the encoding distance manipulation, we performed another series of 2 (Distance: Close vs. Far) x 2 (Duration: Long vs. Short) ANOVAs on participants' reported ability to make out details of the criminal's face. Finally, we performed a series of chi-square tests to determine whether and how encoding variability affected identification decisions. All analyses were separated by crime.

Carjacking

In the carjacking condition, encoding duration had a significant effect on witnesses' perceptions of how long the criminal's face was in sight, $F(1) = 22.12$, $p = .002$, $\eta_p^2 = 0.05$. Witnesses with a long view ($M = 4.07$, $SD = 1.54$) reported that the criminal's face was in sight for a longer time than did witnesses with a short view ($M = 3.44$, $SD = 1.50$). The effect of encoding distance was likewise significant, $F(1) = 17.75$, $p = .005$, $\eta_p^2 = 0.04$. Witnesses with a close view ($M = 4.07$, $SD = 1.58$) reported that the

criminal's face was in sight for a longer time than did witnesses with a far view (M = 3.50, SD = 1.48). The interaction between viewing duration and viewing distance was not significant, $F(1) = 0.11$, $p = .737$, $\eta_p^2 = 0.001$.

Encoding duration also significantly affected participants' reported ability to make out specific facial features of the criminal, $F(1) = 9.38$, $p = .003$, $\eta_p^2 = 0.05$. Witnesses with a long view (M = 3.58, SD = 1.50) reported being better-able to make out the criminal's facial features relative to witnesses with a short view (M = 2.92, SD = 1.48). Witnesses' viewing distance also significantly affected their reported ability to make out the criminal's facial features, $F(1) = 5.66$, $p = .018$, $\eta_p^2 = 0.03$. Witnesses with a close view (M = 3.51, SD = 1.46) reported being better-able to make out specific features of the criminal relative to witnesses with a far view (M = 3.04, SD = 1.54). The interaction between viewing duration and viewing distance was not significant, $F(1) = 1.70$, $p = .194$, $\eta_p^2 = 0.009$.

Regarding identification outcomes, Fisher's exact test indicated that viewing distance did not significantly affect the proportion of hits ($p = .631$), misses ($p = .684$), or filler identifications ($p = 1.00$) from target-present lineups. It did, however, affect picks from target-absent lineups such that witnesses with a close view were significantly more likely to make target-absent picks (0.82) than witnesses with a far view (0.57), $p = .016$. Fisher's exact test also indicated that encoding duration did not significantly affect the proportion of hits ($p = 1.00$), misses ($p = .840$), or filler identifications, ($p = 1.00$), from target-present lineups, or the proportion of correct rejections or target-absent picks ($p = .670$). Table 2 displays the proportion of hits, filler identifications and misses (target-

present lineups), and correct rejections and picks (target-absent lineups) broken down by encoding condition for the carjacking mock crime.

Graffiti

In the graffiti condition, encoding duration had a significant effect on witnesses' perceptions of how long the criminal's face was in sight, $F(1) = 10.25$, $p = .002$, $\eta_p^2 = 0.05$. Witnesses with a long view ($M = 4.04$, $SD = 1.61$) reported that the criminal's face was in sight for a longer time than did witnesses with a short view ($M = 3.35$, $SD = 1.45$). The effect of encoding distance was not significant, $F(1) = 3.49$, $p = .063$, $\eta_p^2 = 0.02$. The interaction between viewing duration and viewing distance was likewise non-significant, $F(1) = 0.02$, $p = .902$, $\eta_p^2 = 0.00$.

Encoding duration also significantly affected participants' reported ability to make out specific facial features of the criminal, $F(1) = 10.73$, $p = .001$, $\eta_p^2 = 0.05$. Witnesses with a long view ($M = 3.71$, $SD = 1.59$) reported being better-able to make out the criminal's facial features relative to witnesses with a short view ($M = 3.01$, $SD = 1.48$). The effect of viewing distance, however, did not significantly affect witnesses' reported ability to make out the criminal's facial features, $F(1) = 0.31$, $p = .580$, $\eta_p^2 = 0.001$. The interaction between viewing duration and viewing distance was likewise non-significant, $F(1) = 1.00$, $p = .318$, $\eta_p^2 = 0.005$.

In terms of identification decisions, Fisher's exact test revealed that encoding distance did not have a significant effect on the proportion of hits ($p = .284$), misses ($p = .145$), or filler identifications ($p = .688$), from target-present lineups. Encoding distance also had no significant effect on the proportion of correct rejections or picks from target-absent lineups ($p = .437$). Fisher's exact test also indicated that encoding duration did not

significantly impact hits (p = .831), misses (p = .417), or filler identifications (p = .313), from target-present lineups, or correct rejections or picks from target-absent lineups (p = .169). Table 3 displays the proportion of hits, filler identifications and misses (target-present lineups), and correct rejections and picks (target-absent lineups) broken down by encoding condition for the graffiti mock crime.

Identification Decisions by Crime

To determine whether identification outcomes differed as a function of crime, the proportion of each identification decision was compared as a function of crime. All comparisons were non-significant (ps > .05), except that Fisher's exact test revealed that witnesses were more likely to make a target-absent pick in the carjacking condition (0.67) relative to the graffiti condition (0.47), p = .005. We elected to collapse across crime for the remainder of our analyses. Tables 4 and 5 display identification decision outcomes in each bin of predictive confidence as a function of sample filler condition (carjacking and graffiti, respectively), and Table 6 displays identification decisions in each confidence bin and sample filler condition collapsed across mock crime.

Correlative Relationships Between Outcome Measures

To assess the relationship between the outcome variables of interest, we ran a bivariate correlation including suspect identifiers' accuracy, predictive confidence (in an identification), dichotomous predictive judgment, and composite memory. Composite memory scores were computed for suspect identifiers by adding their responses to the following encoding-related questions and dividing by six: (1) "How good a view of the criminal did you have?" (2) "How long would you estimate that the criminal's face was in sight?" (3) "How well were you able to make out specific features of the criminal? (4)

How much attention did you pay to the criminal?" (5) "How strong is your memory of the criminal?" (6) "How clear an image of the criminal do you have in your mind?" Participants rated each question on a 1-7 *Likert* scale, with ratings of 1 indicating "not a good view at all", "not long at all", etc. and ratings of 7 indicating "an extremely good view", "extremely long", etc. and thus each suspect identifier's score ranged from 1-7.

Results indicated a moderately strong relationship between outcome variables. As expected, predictive confidence was significantly correlated with composite memory score ($r = 0.64$, $p < .001$) and composite memory score was significantly correlated with the dichotomous predictive judgment ($r = 0.49$, $p < .001$). Nonetheless, no outcome measure was significantly correlated with accuracy (all *p*s > .05).

Impact of Sample Filler Exposure on Predictive Confidence

A 3 (Sample Filler Condition: None vs. Poor vs. Good) x 2 (Predictive Judgment: Predicting Identification vs. Predicting Rejection) repeated measures ANOVA revealed a significant effect of sample filler exposure on witnesses' predictive confidence, $F(2) = 13.46$, $p < .001$, $\eta_p^2 = .06$. When predicting the likelihood of a correct identification, exposure to sample fillers significantly reduced witnesses' confidence such that witnesses in the poor ($M = 49.92$, $SD = 26.81$) and good ($M = 40.07$, $SD = 23.49$) sample filler conditions asserted significantly lower confidence relative to witnesses in the control condition ($M = 54.61$, $SD = 24.28$). Similarly, when predicting the likelihood of a correct lineup rejection, witnesses in the poor ($M = 46.54$, $SD = 26.52$) and good ($M = 35.43$, $SD = 23.94$) sample filler conditions were significantly less confident compared to those in the control condition ($M = 49.08$, $SD = 25.60$).

Calibration Analyses

To examine the relationship between confidence (both predictive and postdictive) and accuracy, and whether this relationship varied as a function of the sample filler manipulation, separate calibration curves were constructed for each sample filler condition. These curves were also separated between choosers and non-choosers so that the predictive confidence curves accounted for witnesses' confidence in a correct identification (choosers) or rejection (non-choosers). Witnesses are typically put into one of five confidence bins (0-20%, 30-40%, 50-60%, 70-80%, and 90-100%, but due to having few witnesses on the upper end of the confidence scale, we elected to combine the two highest bins (thus having a 70-100% bin). Witnesses' accuracy in each confidence bin (*j*) was calculated using the formula $a_j = \frac{\#\ accurate\ ID\ decisions}{\#\ accurate\ ID\ decisions + (\#inaccurate\ ID\ decisions)}$ (Mickes, 2015). We calculated Calibration (*C*), Over/Under-confidence (*O/U*), Adjusted Normalized Resolution Index (*ANRI*) statistics to determine whether and how witnesses' calibration was affected by exposure to sample fillers (see Table 7).

Calibration indicates the extent to which a calibration curve represents perfect calibration: values range from 0 to 1, with 0 indicating perfect calibration. Over/Under-confidence ranges from -1 to 1 and denotes how far (below or above) a calibration line is from perfect calibration; a negative value indicates under-confidence, and a positive value indicates over-confidence. Finally, the Adjusted Normalized Resolution Index statistic measures discriminability and represents how well a calibration curve distinguishes in/correct identification decisions. This value typically ranges from 0 to 1 (with 1 denoting perfect discrimination), though negative values may be observed. If that

is the case, such values are effectively treated as 0 (indicating very poor discrimination; Yaniv et al., 1991). The "jackknife" function in R was utilized to generate 95% inferential confidence intervals (ICIs) for each calibration metric. These confidence intervals were compared across sample filler conditions to determine statistical significance: any instance in which the confidence intervals of two experimental conditions did not overlap indicated a significant difference at a .05 alpha level (Tryon & Lewis, 2008). Figures 1 and 2 display the calibration curves for both predictive and postdictive confidence, respectively, for witnesses in each sample filler condition and for choosers and non-choosers.

Predictive confidence

Choosers

Among witnesses who made a lineup identification, those exposed to no sample fillers exhibited the strongest calibration between predictive confidence and accuracy (*C* = 0.122, *SE* = 0.034, [0.074, 0.170]). This calibration was significantly stronger compared to witnesses exposed to poor sample fillers (*C* = 0.229, *SE* = 0.042 [0.170, 0.288]) but did not significantly differ from witnesses exposed to good sample fillers. Calibration also did not significantly differ between witnesses exposed to poor or good sample fillers. Neither *O/U* nor *ANRI* values statistically differed as a function of any condition comparisons.

Non-Choosers

No statistically significant differences were found amongst non-choosers for any of the *C*, *O/U*, or *ANRI* comparisons. This was not particularly surprising, as research generally tends not to find significant differences amongst witnesses who reject a lineup (Palmer et al., 2013).

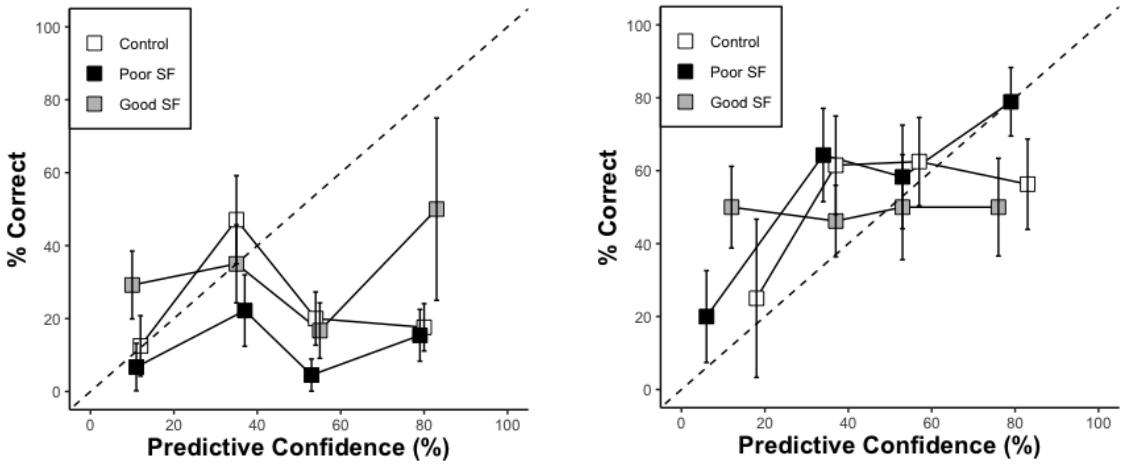


Figure 1. Calibration plots for choosers' (left) and non-choosers' (right) identification accuracy at each level of binned predictive confidence as a function of sample filler condition. Bars represent the standard error for each confidence bin, and the dotted line represents perfect calibration.

Postdictive confidence

As can be seen in Table 7, no statistically significant differences were found in any of the *C*, *O/U*, or *ANRI* comparisons for postdictive confidence across the three sample filler conditions. This was true regardless of witnesses' choosing status. The lack of a confidence-accuracy relationship amongst non-choosers is not atypical, but the lack of relationship for choosers (particularly in the control condition) is somewhat surprising. Even with delay, past eyewitness research has established a reliable postdictive confidence-accuracy relationship amongst witnesses who make identifications (Palmer et al., 2013; Sauer et al., 2010).

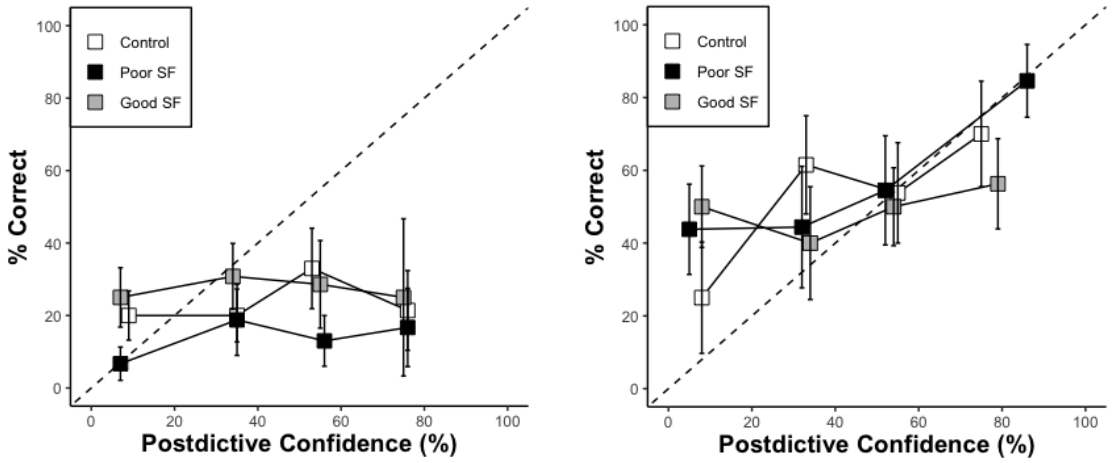


Figure 2. Calibration plots for choosers' (left) and non-choosers' (right) identification accuracy at each level of binned postdictive confidence as a function of sample filler condition. Bars represent the standard error for each confidence bin, and the dotted line represents perfect calibration.

Control vs. Collapsed Filler Conditions

Because we found no significant differences between the good vs. poor sample filler conditions on any calibration metric, we elected to collapse these conditions and examine the effects of sample filler exposure (versus none) on witnesses' calibration. For non-choosers, there were no significant differences on any of the calibration metrics for predictive or postdictive confidence. However, for choosers, there was a statistically significant difference in terms of predictive confidence: witnesses exposed to sample fillers were significantly more over-confident (*O/U* = 0.37, *SE* = 0.04 [0.31, 0.43]) compared to witnesses in the control condition (*O/U* = 0.25, *SE* =0.05 [0.18, 0.31]). Collapsed filler condition choosers did not statistically differ from choosers in the control condition on any calibration metric with respect to postdictive confidence.

Confidence-Accuracy Characteristic (CAC) Analyses

CAC analyses include only witnesses who made suspect identifications and examine the likelihood that a suspect identification is accurate at differing levels of

confidence. Witnesses' confidence is typically binned into three categories: low (0-60%), medium (70-80%) and high (90-100%) – however, as with calibration, there were very few suspect identifiers in the highest confidence bin. As such, we elected to dichotomize witnesses into two bins (low/high) and varied the methods by which this dichotomization was accomplished (performing a median split and calculating the point at which predictive confidence/memory scores maximally differentiate witness accuracy). Suspect identifier (n = 168) accuracy per confidence bin (j) was then calculated as $a_j = \frac{\#\ accurate\ suspect\ IDs}{\#\ accurate\ suspect\ IDs + \left(\frac{\#\ inaccurate\ suspect\ IDs}{6}\right)}$. The innocent suspect identification rate was estimated as 1/6th of the choosing rate because we did not have an a priori innocent suspect (Mickes, 2015). Table 8 contains the frequency of correct suspect identifications and target absent picks when suspect identifiers were dichotomized by median split (accuracy graphed in Figure 3).

Median Split

We first dichotomized suspect identifiers into two groups of predictive and postdictive confidence (low versus high) by performing a median split on witnesses' confidence in each sample filler condition, and then grouping witnesses based on whether their predictive or postdictive confidence was equal to or less than the median split value (versus above) for their respective condition. The median split value for predictive confidence was 50% in the control and poor sample filler conditions, and 40% in the good sample filler condition. The median split value for postdictive confidence was 40% in the control and poor sample filler conditions, and 30% in the good sample filler condition.

Predictive Confidence. Fisher's exact test was used to compare the accuracy of witnesses reporting low versus high predictive confidence in each sample filler condition. Comparisons revealed that in the control condition, witnesses who asserted high predictive confidence were no more accurate (0.32) than witnesses who asserted low predictive confidence (0.34), $p = 1.00$. This was also true of the poor and good sample filler conditions. Poor sample filler witnesses in the high confidence bin were not significantly more accurate (0.20) than witnesses in the low confidence bin (0.18), $p = 1.00$, and good sample filler witnesses in the high confidence bin did not differ in accuracy (0.32) from witnesses in the low confidence bin (0.47), $p = .377$.

Postdictive Confidence. Fisher's exact test was also used to compare the accuracy of witnesses reporting low versus high postdictive confidence in each sample filler condition. In the control condition, witnesses asserting high postdictive confidence were not significantly more accurate (0.38) than witnesses asserting low postdictive confidence (0.31), $p = .599$. The same was true of the sample filler conditions: in the poor sample filler condition, witnesses asserting high confidence were no more accurate (0.21) than witnesses asserting low postdictive confidence (0.17), $p = 1.00$. In the good sample filler condition, there was also no significant difference in accuracy between witnesses asserting high postdictive confidence (0.30) and low postdictive confidence (0.48), $p = .246$.

High Confidence Suspect Identifications Across Sample Filler Conditions. The accuracy of high-confidence suspect identifications was compared across sample filler conditions for predictive and postdictive confidence. These witnesses were chosen for comparison due to their forensic relevance (i.e., they are most likely to be called to

testify during trial; Mickes, 2015). Fisher's exact test revealed no significant differences in accuracy between any of the sample filler conditions for predictive confidence (p = .849). There were also no significant differences across sample filler conditions for postdictive confidence, $\chi^2(2) = 0.11$, $p = .945$, $\varphi_c = 0.03$.

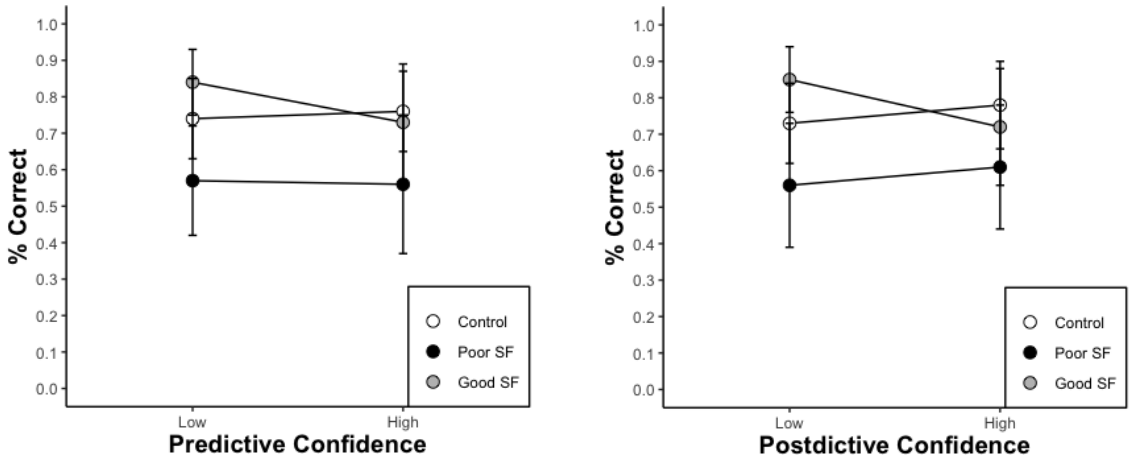


Figure 3. Suspect identifier accuracy at low vs. high (determined by median split value) predictive confidence (left) and postdictive confidence (right) as a function of sample filler condition. Bars represent standard error for each confidence bin.

Maximally Differentiating Chi-Square Values

We next dichotomized suspect identifiers into two groups of predictive and postdictive confidence (low versus high) by running a series of chi-square analyses on witnesses' confidence in each sample filler condition. First, witnesses were split into those whose confidence was equal to 0% (the minimum possible) and those whose confidence was higher, and a 2 (confidence equal to/below vs. above the split) x 2 (accurate vs. inaccurate) chi-square was conducted to determine whether accuracy varied as a function of this split. This process was repeated by splitting witnesses into those whose confidence was equal to or less than 10% and those whose confidence score was higher, and again a chi-square analysis was conducted. Similar splits were performed

incrementally at each point of witnesses' predictive and postdictive confidence, and the corresponding chi-squares were calculated, until reaching the maximum possible confidence value (100%). The peak values indicate the point at which witnesses' predictive and postdictive confidence produced the greatest differentiation between those who made accurate and inaccurate identifications. We sought to identify which chi-square value in each condition maximally differentiated witnesses such that those above the cut-off were more accurate than those equal to or below the cutoff. However, none of our chi-square values reached statistical significance (except for one, which can be attributed to low bin size; see Table 9).

Predictive Confidence. Fisher's exact test was used to compare the accuracy of witnesses reporting low versus high predictive confidence in each sample filler condition. In the control condition, witnesses in the low confidence bin did not significantly differ in accuracy (0.48) from witnesses in the high confidence bin (0.27), $p = .104$. This was also true of the poor sample filler condition, such that witnesses in the low confidence bin did not differ in accuracy (0.16) from witnesses in the high confidence bin (0.38), $p = .163$. Finally, in the good sample filler condition, witnesses in the low confidence bin did not differ in accuracy (0.75) from witnesses in the high confidence bin (0.38), $\chi^2(1) = 2.11$, $p = .147$, $\varphi_c = 0.21$.

Postdictive Confidence. In the control condition, witnesses in the low confidence bin did not significantly differ in accuracy (0.32) from witnesses in the high confidence bin (1.0), $\chi^2(1) = 2.03$, $p = .154$, $\varphi_c = 0.18$. In the poor sample filler condition, however, there was a significant difference in accuracy with respect to confidence level: witnesses in the low confidence bin were significantly less accurate (0.17) than witnesses in the

high confidence bin (1.0), $\chi^2(1) = 4.38$, $p = .036$, $\varphi_c = 0.29$. In the good sample filler condition, Fisher's exact test revealed that there was no difference between witnesses in the low and high confidence bins with regard to accuracy (0.48 and 0.30, respectively), $p = .246$.

High Confidence Suspect Identifications Across Sample Filler Conditions. The accuracy of high confidence suspect identifications was again compared across sample filler conditions. Results indicated no significant differences between any of the sample filler conditions for predictive confidence, $\chi^2(2) = 2.83$, $p = .243$, $\varphi_c = 0.12$, or postdictive confidence, $\chi^2(1) = 4.84$, $p = .089$, $\varphi_c = 0.25$.

Memory Strength-Accuracy Characteristic (MSAC) Analyses

Memory strength bin classification ("weak" versus "strong") was accomplished using the same techniques as with confidence (i.e., via a median split or via chi square analyses). Dichotomized memory strength-accuracy characteristic (MSAC) curves were then constructed using the formula $a_j = \frac{\#\ accurate\ suspect\ IDs}{\#\ accurate\ suspect\ IDs + \left(\frac{\#\ inaccurate\ suspect\ IDs}{6}\right)}$ to calculate the proportion of accurate suspect identifiers in each memory strength bin. The number of accurate suspect identifications and target-absent picks as a function of binned composite memory (determined by performing a median split) and sample filler condition is displayed in Table 10. This breakdown is not reported for the chi square analyses because no chi square value reached statistical significance. Finally, dichotomized MSAC curves are displayed in Figure 6.

Median Split

We first dichotomized suspect identifiers into memory strength bins ("weak" vs. "strong") based on the median composite memory score in each sample filler condition.

The control condition had a median split value of 3.83, the poor sample filler condition had a median split value of 4.00, and the good sample filler condition had a median split value of 3.17.

In the control condition, Fisher's exact test revealed that witnesses in the weak memory strength bin did not differ in accuracy (0.38) from witnesses in the strong memory strength bin (0.30), $p = .586$. In the poor sample filler condition, witnesses in the weak memory strength bin also did not differ in accuracy (0.23) from witnesses in the strong memory strength bin (0.13), $\chi^2(1) = 0.90$, $p = .343$, $\varphi_c = 0.13$. Finally, in the good sample filler condition, Fisher's exact test indicated that witnesses in the weak memory strength bin did not differ in accuracy (0.35) from witnesses in the strong memory bin (0.48), $p = .394$.

When strong memory suspect identifiers were compared across sample filler conditions, results revealed no significant difference in accuracy, $\chi^2(2) = 1.22$, $p = .544$, $\varphi_c = 0.09$.

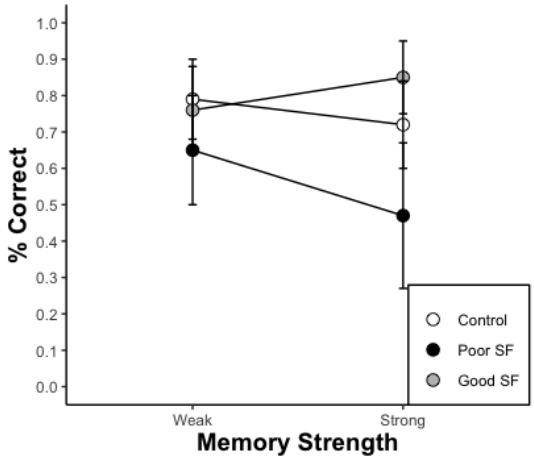


Figure 4. Suspect identifier accuracy at weak vs. strong composite memory (determined by median split value) as a function of sample filler condition. Bars represent standard error for each confidence bin.

Maximally Differentiating Chi-Square Values

We next dichotomized suspect identifiers into two memory strength bins ("weak" and "strong") based on the composite memory score that best differentiated accurate and inaccurate witnesses in each sample filler condition. The specific score for each sample filler group was determined by performing a series of chi-square analyses that split suspect identifiers into those whose composite memory score was equal to 1 (the minimum possible score) and those whose memory score was higher, and a 2 (composite memory score equal to/below vs. above the split) x 2 (accurate vs. inaccurate) chi-square was conducted to determine whether accuracy varied as a function of this split. This process was repeated by splitting witnesses into those whose composite memory score was equal to or less than 1.33 and those whose composite memory score was higher, and again a chi-square analysis was conducted. Similar splits were performed incrementally at each point of witnesses' composite memory score, and the corresponding chi-square was calculated, until reaching the maximum possible composite score (7). Unfortunately, none of the chi-square values reached significance, indicating that there was no cut-off that differentiated weak- from strong-memory witnesses (see Table 11).

Dichotomous Predictive Judgment

To determine whether witnesses who responded that they would be able to make an accurate identification decision ("yes") exhibited greater accuracy than witnesses who responded that they would not be able to make an accurate decision ("no"), we ran a series of tests. First, we analyzed witnesses from all three filler-exposure conditions – in total, there were 199 witnesses who responded "yes" and 212 witnesses who responded "no". Fisher's exact test indicated no significant difference in accuracy between

witnesses who responded “yes” and those who responded “no” (p = .754). Next, we analyzed each filler exposure group separately. Fisher’s exact test again revealed no significant differences in accuracy between witnesses who responded “yes” versus “no” in the control condition (p = .455), poor sample filler condition (p = .177), or good sample filler condition (p = .575).

Next, we examined only suspect identifiers (because they are the most forensically relevant witnesses). This time, Fisher’s exact test revealed a statistically significant difference in accuracy: contrary to expectations, witnesses who responded “no” were significantly more accurate (0.38) than witnesses who responded “yes” (0.23) p = .044 (see Table 12; results plotted in Figure 5). When broken down by sample filler condition, however, the dichotomous response did not significantly differentiate witness accuracy (all ps > .05).

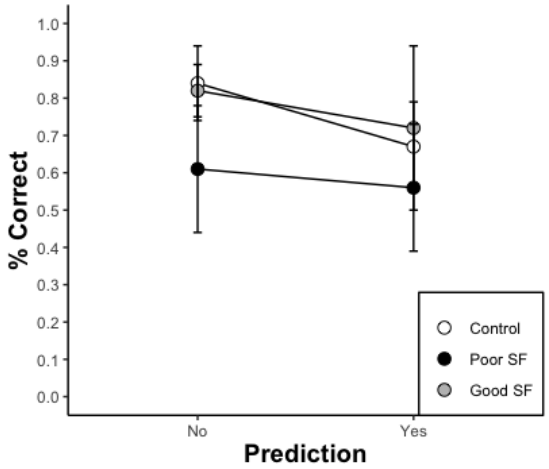


Figure 5. Suspect identifier accuracy for each dichotomous identification prediction in the as a function of sample filler condition. Bars represent standard error for each response.

Logistic Regressions

To examine whether the sample filler manipulation affected witnesses' ability to predict their future identification performance, we ran a series of logistic regressions with identification accuracy regressed onto (a) witnesses' responses to the dichotomous judgment, (b) sample filler condition, and (c) the interaction between these two variables. Separate logistic regressions were run for each pairwise comparison (control vs. poor fillers; control vs. good fillers; poor vs. good fillers). Table 13 displays parameter estimates of these regression models. For the control/poor filler comparison and the control/good filler comparison, witnesses' dichotomous predictive judgment, sample filler condition, and the interaction were all non-significant predictors of accuracy (*p*s > .05). For the good/poor filler comparison, witnesses' dichotomous predictive judgments were not a significant predictor of accuracy, but sample filler condition was: witnesses in the good sample filler condition were significantly more likely to be accurate than witnesses in the poor sample filler condition (ß = -0.82, *SE* = 0.36, Wald = 5.33, *p* = .022).

IV. DISCUSSION

Although researchers have generally failed to find that eyewitnesses' predictive confidence is a reliable indicator of subsequent identification accuracy (Cutler & Penrod, 1989; Hourihan et al., 2012; Nguyen et al., 2018; Whittington et al., 2019), recent work has challenged this conclusion by demonstrating that a moderate relationship *can* exist when witnesses experience varied encoding and appropriate statistical techniques are used (Molinaro et al., 2021; Shambaugh & Charman, in preparation). The current paper sought to further explore the use of predictive confidence as a potential screening tool for

eyewitnesses by addressing three main goals: (a) testing whether it was possible to reduce witnesses' confidence (and thus improve calibration) by exposing them to sample fillers prior to their predictive confidence judgment, (b) examining whether witnesses' memory strength could be used as an alternative predictor of identification accuracy, and (c) exploring witnesses' ability to predict their identification accuracy by asking them to make a dichotomous judgment about their subsequent identification accuracy. Our results are discussed in light of each goal.

Goal #1: Reducing Witness Overconfidence via Sample Filler Exposure

Although some studies have shown a reliable relationship between predictive confidence and subsequent identification accuracy (Molinaro et al., 2021), witnesses tend to exhibit overconfidence when predictive confidence judgments are made under real-world conditions involving a single identification attempt and a delay between the crime and the lineup task (Shambaugh and Charman, in preparation). One potential explanation for this overconfidence is that witnesses lack familiarity with the difficulty of the lineup task (which is determined in part by the similarity of the lineup fillers to the target, a factor that cannot be accounted for when providing a predictive confidence judgment). Thus, the first goal of the current study was to determine whether witnesses' confidence could be reduced by exposing them to sample fillers that accurately represented the difficulty of the lineup task prior to making their predictive confidence judgments. Given that we varied encoding of the criminal across witnesses, we expected to observe a significant relationship between predictive confidence and identification accuracy, consistent with past studies. Considering the metacognition literature on problem characteristics, we also expected that exposure to good sample fillers (cf. exposure to

poor sample fillers or no exposure to sample fillers) prior to making a predictive confidence judgment would reduce witnesses' confidence, thereby improving their calibration.

Our predictions were somewhat supported: witnesses exposed to good sample fillers asserted significantly lower predictive confidence (in an accurate identification *or* in an accurate rejection) relative to witnesses who were not exposed to sample fillers. Contrary to our hypotheses, however, we found that exposing witnesses to good sample fillers did not reduce *overconfidence*. Witnesses in our good sample filler condition had statistically equivalent overconfidence to witnesses in the poor sample filler and control conditions. Furthermore, when the sample filler conditions were collapsed, we found that exposure to sample fillers (whether good or poor) resulted in significantly *increased* overconfidence relative to the control condition. Consequently, witnesses who were exposed to good sample fillers had no better calibration than witnesses exposed to poor sample fillers or witnesses in the control condition.

We had reasoned (though did not directly hypothesize) that exposure to poor sample fillers would provide no benefit to calibration, and this was partially supported – we observed a significant decrease in calibration, and an increase in their overconfidence, amongst choosers who were exposed to poor sample fillers relative to choosers in the control condition. Interestingly, the calibration of witnesses who rejected the lineup appeared (graphically) to be superior to that of choosers; nonetheless, there were no statistically significant differences between non-choosers in any of the sample filler conditions. In terms of diagnosticity, the *ANRI* values, which represent the extent to which witness accuracy can be differentiated based on their predictive confidence, were

low and not significantly different across all three conditions. In other words, the sample filler manipulation did not improve discrimination of accurate and inaccurate eyewitnesses. Taken together, it appears that our sample filler manipulation was not sufficient to significantly improve witnesses' predictive confidence judgments – rather, showing sample fillers only made them more (over)confident in their future identification performance (thus harming calibration).

We found similar results when evaluating only suspect identifiers in our CAC analyses: suspect identifiers who had been exposed to good sample fillers were not significantly more accurate in their identification decisions compared to suspect identifiers who had been exposed to poor sample fillers or to suspect identifiers who had not been exposed to sample fillers. This was true regardless of whether witnesses were binned into confidence categories based on their condition's maximally differentiating chi square value, or their condition's median split value.

Differences in Sample Filler Quality

In all three conditions, witnesses overestimated their future memory performance. However, sample filler exposure (whether good or poor) significantly increased this overconfidence relative to the control condition. One explanation for this unexpected effect is that following a week delay, witnesses tended to have a poor memory of the target, resulting in control witnesses expressing relatively low predictive confidence judgments. However, because of this weak memory, witnesses exposed to sample fillers should have tended to lack a recognition experience when exposed to the various sample fillers, leading them to easily determine that the target was not among them; if witnesses then based their predictive confidence judgments on the ease with which they rejected the

sample fillers, then exposure to fillers (whether good or poor in quality) would tend to result in witnesses maintaining a relatively high degree of confidence in their memory performance. In other words, if witnesses attributed the ease of rejection of the sample fillers to the fact that the target was not among them (instead of to the fact that their memory was weak) then the sample fillers might have erroneously increased their confidence in their future lineup performance. However, when witnesses viewed the actual lineup, they had a similar issue (such that they were unable to distinguish lineup members from their memory trace) producing a high rate of inaccuracy.

Although witnesses in the two sample filler conditions did not differ significantly in terms of calibration, the results of the logistic regressions indicated that they did differ in accuracy: witnesses in the good sample filler condition were significantly more accurate than those in the poor sample filler condition. Although an unexpected result, this result can potentially be explained in light of diagnostic feature-detection theory (Wixted & Mickes, 2014). This theory asserts that exposing witnesses to multiple faces simultaneously (as in a lineup) improves discrimination because witnesses can compare faces to determine which feature(s) differ across lineup members, and use features that are diagnostic to rule out certain lineup members while focusing on others. Thus, it is possible that exposing witnesses to good sample fillers (who possessed similar features to the fillers in the actual lineup) led witnesses to focus specifically on diagnostic characteristics, allowing them to differentiate the target from the lineup fillers with greater accuracy. Conversely, exposure to poor sample fillers would not provide diagnostic information for the actual lineup leading to no benefit in identification accuracy. This idea could be tested in future research.

The Role of Retention Interval

Even without exposure to sample fillers, witnesses' predictive confidence did not show a reliable relationship with identification accuracy. Considering findings from Shambaugh and Charman (in preparation), however, this may not be particularly surprising – in their study, witnesses who experienced a one-week delay between encoding and predictive confidence judgment (and lineup identification attempt) also failed to show a significant relationship between predictive confidence and accuracy. There is a good explanation for this pattern of results: one of the prerequisites for a reliable predictive confidence-accuracy relationship is varied encoding (which produces varied memory strength) amongst witnesses (Molinaro et al., 2021). Although witnesses in both the current study and the Shambaugh and Charman study were exposed to varied encoding, this occurred during Part 1; however, they did not attempt to make judgments using this memory trace until one week later. Thus, immediately after viewing the mock crime, witnesses likely had significant variation in memory strength due to the encoding manipulation – but by the time they reported their predictive confidence and attempted an identification a week later, that variation had disappeared, resulting in no predictive confidence-accuracy relationship.

All witnesses presumably experienced memory degradation during the retention interval, but those with an initially strong memory should have experienced relatively more degradation than witnesses with an initially poor memory, as they had more opportunity for deterioration. Therefore, after one week, strong- and weak-memory witnesses *both* likely had weak memories, rendering them almost indistinguishable. This explanation bears out in the results. A breakdown of witnesses' mean composite memory

score in each mock crime condition shows little variation: witnesses' mean composite memory in the carjacking condition (M = 3.68, SD = 1.26) was not significantly different from witnesses in the graffiti condition (M = 3.67, SD = 1.30).

This explanation is also compatible with identification results from Part 2: when analyzing witnesses' identification decisions as a function of encoding condition, we found no significant differences. Furthermore, looking at Tables 2 and 3 makes it evident that witnesses' memory performance was quite poor following the one-week delay regardless of whether they initially experienced good or poor encoding. By and large, witnesses in the current study made a low proportion of hits from target-present lineups (0.24) and a large proportion of picks from target-absent lineups (0.57). Taken together, it is likely that the memory degradation that occurred over the retention interval hampered the possibility of a reliable predictive confidence-accuracy relationship by diluting variability in memory strength and failing to produce significant differences in identification outcomes.

Theoretical and Practical Implications

The observation that witnesses were unable to account for the effects of encoding on memory after a delay has implications for current eyewitness theory. The constant-likelihood ratio model, proposed by Semmler et al. (2018), argues that witnesses can maintain a constant likelihood ratio of correct to incorrect lineup identification decisions for any given confidence level by adjusting their decision criteria across variations in memory strength. In other words, they argue that witnesses are appropriately sensitive to variations in memory strength when assessing their confidence. In light of this theory, we would have expected that witnesses could adjust their predictive confidence as a function

of their encoding experience and the memory deterioration that took place during the retention interval. Nonetheless, that is not what we found. Rather, witnesses maintained a high level of confidence in their identification abilities while simultaneously being highly inaccurate. Importantly, Semmler et al. did not examine retention interval as an estimator variable; thus, it is important for future research to examine the effect of retention interval on the reliability of the confidence-accuracy relationship. Given the current results, it is possible that a constant likelihood ratio model is unable to account for witnesses' confidence judgments when they have experienced a significant retention interval.

Practically speaking, the results of our calibration analyses suggest that predictive confidence is not suitable for use with real-world witnesses who are asked to make a lineup identification. Because lineup identifications generally occur following a substantial delay (as they require the police to not only find a suspect, but also to find suitable fillers for inclusion in the procedure – both of which can take a substantial amount of time), witnesses presented with a lineup are likely to have a relatively poor memory. As the current study has revealed, the utility of witnesses' predictive confidence is not improved by sensitizing them to the lineup task via sample filler exposure after a delay; in fact, doing so may even make them more overconfident. However, in the event that a witness has a particularly good memory of the criminal (either because their exposure was strong or because their predictive confidence judgment is not far removed from the time of encoding), it is possible that the sample filler manipulation could work as intended.

Goal #2: Using Self-Reported Memory Strength to Predict Witness Accuracy

Prior research on predictive confidence has suggested that more direct measures of witnesses' memory strength may be a better predictor of lineup accuracy than confidence (Molinaro et al., 2021; Shambaugh & Charman, in preparation). However, the methods by which witnesses were categorized into memory strength bins (i.e., low, medium, high) in these studies was arbitrary, and it is possible that the observed results were a consequence of these quasi-arbitrary categorization methodologies. The second goal of this study, therefore, was to continue evaluating witnesses' memory strength as a predictor of lineup performance using two new approaches for binning witnesses into memory strength categories: calculating maximally differentiating chi-square values and performing median splits.

As with predictive confidence, witnesses' memory strength failed to reliably predict identification accuracy, regardless of whether witnesses were binned via maximally differentiating chi-square values or median split values, and (as expected) did not differ as a function of sample filler exposure. Given that witnesses' memory strength was significantly correlated with their predictive confidence (which did not reliably predict accuracy), it is unsurprising that memory strength also failed to predict accuracy as well. This result is consistent with findings from Shambaugh and Charman (in preparation) who found that memory strength was a reliable predictor of identification accuracy when witnesses made their predictive memory judgments and lineup identification immediately after viewing the mock crime, but not when the judgments were delayed one week.

Theoretical and Practical Implications

The immediacy between encoding and predictive memory judgments in Molinaro et al. (2021) and Shambaugh and Charman (in preparation) maintained variability in witnesses' memory strength. This variability is critical for producing a reliable relationship between predictive confidence, memory strength, and accuracy. Thus, the results of the current study suggest that the retention interval between encoding and witnesses' predictive memory judgments plays a critical role in moderating the utility of those memory judgments. Even if predictive memory judgments are made immediately before witnesses' identification attempt, a delay from the witnessing event will harm their predictive utility due to a lack of variability in memory strength across witnesses.

From a practical standpoint, the current study suggests that memory strength measures are subject to the same timing-related limitations as predictive confidence. This does not mean, however, that self-reported memory measures should be discounted altogether – these measures should still predict accuracy when memory variability is maintained (i.e., when there is minimal retention interval). Rather, researchers and practitioners should be aware that using witnesses' self-reports of memory strength to predict subsequent lineup performance may be inappropriate if these judgments were made after a delay from the time of encoding (as in the current study).

Goal #3: Exploring a Dichotomous Predictive Judgment

Although predictive confidence and memory strength have the potential to be useful screening tools (at least among witnesses whose memories exhibit substantial variability), they share one critical limitation: they are continuous measures. In the real world, however, predictive confidence and self-reported memory measures are useful

because they can be used to determine whether to present witnesses with a lineup—which is a dichotomous decision. An important practical question, then, is how should law enforcement transform a continuous measure into a dichotomous decision. In other words, what is the ideal way to impose a cut-off point on those continuous data? We examined three different ways of imposing a cut-off point. First, we performed a median split on witnesses' predictive confidence/memory scores. However, this method fails to work in the real world: to perform a median split, police would need a distribution of data to determine the appropriate cut-off values, which they clearly do not have for a single witness. Second, we looked for a maximally-differentiating point: a point on the continuous predictive confidence/memory scales that maximally differentiated the accuracy of good memory/high confidence witnesses from poor memory/low confidence witnesses. However, this method is also likely to fail in the real world, as the maximally-differentiating point is likely to vary across witnesses/circumstances, making it difficult for researchers to make practical, generalizable recommendations. Thus, to overcome these limitations, the current study examined a third method: having witnesses categorize themselves by asking them whether they could make an accurate identification decision if shown a lineup.

We hypothesized that witnesses who responded "yes" to the dichotomous lineup performance question would exhibit greater accuracy compared to witnesses who responded "no." We further expected that this effect would be stronger amongst witnesses who were shown good sample fillers (cf. poor quality sample fillers or no sample filers) due to the anticipated reduction in overconfidence. Nonetheless, when evaluating suspect identifiers' responses to this question we found that those who said

"no" were, counterintuitively, significantly more accurate overall than those who said "yes." When broken down by sample filler condition, witnesses did not significantly differ in accuracy based on their dichotomous response. Thus, witnesses' dichotomous prediction regarding their subsequent identification performance did not predict accuracy (in the expected direction).

Theoretical and Practical Implications

The results of our bivariate correlation indicated that, like predictive confidence, witnesses' dichotomous prediction was significantly correlated with their composite memory score. It is unsurprising, then, that this dichotomous prediction failed to reliably predict accuracy, given our findings that neither predictive confidence nor self-reported memory measures predicted accuracy. After a one-week delay, witnesses' memory had deteriorated to the point that they no longer had a reliable sense of how it would perform on the subsequent memory test. Because all witnesses likely had a poor memory at the time they provided a response to the dichotomous question, they were unable to predict their own performance.

Practically-speaking, these results indicate witnesses' dichotomous prediction regarding their identification accuracy should not be used to determine whether to show a witness a lineup when those predictions occur after a delay from the time of the crime. That said, it is possible that a dichotomous prediction could hold utility if made immediately after encoding (due to variability in witnesses' memory strength). This would certainly be a plausible implementation in the real world, as such a question could be asked by police during an initial crime scene response. Furthermore, in asking witnesses to predict their *decision* accuracy, our question was non-specific. If the

question directly addressed witnesses' ability to make an *identification*, perhaps their dichotomous answer could show a stronger relationship to their subsequent accuracy.

Sample Filler Exposure and Postdictive Confidence

The current study also assessed witnesses' postdictive confidence as a comparison to their predictive assessment of confidence. Interestingly, the calibration of postdictive confidence was just as bad as that of predictive confidence (i.e., there was no confidence-accuracy relationship – even in the control condition). This is largely attributable to witness overconfidence across all three sample filler conditions – in fact, the degree of postdictive overconfidence amongst choosers was consistently greater than that of their predictive confidence (though not significantly so). We found similar results amongst suspect identifiers, such that postdictive confidence did not have a significant relationship with accuracy. This replicates findings from Shambaugh and Charman (in preparation), which found that the postdictive confidence-accuracy relationship deteriorated amongst witnesses who experienced a delay between encoding and the lineup judgment. It also supports other extant research that has noted overconfidence in post-identification judgments following a retention interval (Palmer et al., 2013; Sauer et al., 2010).

The lack of postdictive confidence-accuracy relationship in the current study can, like predictive confidence, be explained by retention interval. Given that witnesses likely tended to possess a weak memory at the time of the identification decision, it is unlikely that they had a strong recognition experience when viewing the lineup. Without variability in the strength of witnesses' recognition experience, it is unlikely that they would display significant variability in their postdictive confidence. This is supported by our data such that, on average, witnesses asserted notably low postdictive confidence (*M*

= 37.64, *SD* = 26.56). Consequently, the current study provides additional evidence that a delay between encoding and memory assessment harms the relationship between witnesses' memory assessments and identification accuracy. The lack of postdictive confidence-accuracy relationship can also be explained in light of research that has observed harmful effects of predictive confidence judgments on postdictive confidence judgments (Whittington et al., 2019). Given that the current study's design prompted witnesses to make both predictive and postdictive confidence judgments, it is possible that the former judgment negatively affected the latter (as in Whittington et al.).

At the very least, the combination of calibration and CAC results in the current study leaves a question mark regarding the reliability of the postdictive confidence-accuracy relationship when the identification judgment is delayed from encoding. Given that postdictive confidence is a widely endorsed metric for estimating witnesses' likelihood of identification accuracy (Handler & Frühholz, 2021; Nguyen et al., 2018), evaluating its reliability under ecologically valid conditions (i.e., delay) certainly warrants more empirical attention. Unfortunately, the vast majority of eyewitness studies do not use an appreciable delay between encoding and lineup test, raising questions as to the generalizability of research findings to the real world. Based on the lack of postdictive confidence-accuracy relationship in the current study and Shambaugh and Charman (in preparation), we argue that future research should make a point to address ecological validity by including a delay as part of their methodology. Incorporating delay into experimental studies on a consistent basis is the only way to determine whether the postdictive confidence-accuracy is dependent on a lack of retention interval.

Study Limitations

Memory Weakness

Theoretically, the current study indicates that predictive judgments made after a delay are not reliable indicators of witnesses' future identification accuracy. We corroborated findings from Shambaugh and Charman (in preparation) such that no predictive confidence-accuracy or memory strength-accuracy relationship existed when witnesses made their predictive judgments one week after encoding. We also found no relationship between witnesses' dichotomous predictions and subsequent accuracy. Critically, familiarizing witnesses with the difficulty of the lineup task via sample filler exposure did not improve the predictive utility of any of these memory judgments (and if anything, harmed it).

The lack of reliable relationship between witnesses' predictive judgments and accuracy in the current study can be attributed to the delay between encoding and predictive judgments. During the retention interval, all witnesses experienced degradation in memory strength; when they attempted to make predictive judgments using this memory, all witnesses likely had a weak memory. This was reflected in our data, which showed that witnesses had a relatively low hit rate and high target-absent choosing rate. They still performed better than chance, however, suggesting that witnesses *did* have some memory of the crime (albeit weak).

Although witnesses experienced encoding variability during Part 1 of the study, these manipulations were not strong enough to produce significant differences in memory strength one week later. The lack of variability in memory strength made it difficult (if not impossible) to predict accuracy from any measures based on underlying memory

strength (such as predictive confidence). Consequently, testing the sample filler manipulation when witnesses had such weak memory traces to begin with was not ideal (as the lineup task was already quite difficult). It is possible that our expected effects would have been found had witnesses' memory strength had significant variability.

Nonetheless, while delay may have been a limitation to the current study, our results should not be dismissed given the ecological validity of the study's design. The current study used a delay (one week) that is not unrealistic for the real world; if anything, it may underrepresent the delay that witnesses experience – they may wait weeks or even months before viewing a lineup. Consequently, we might reasonably expect that real-world eyewitnesses have relatively weak memories much of the time, making the current research critical to informing our expectations for witnesses' memory performance after a delay.

Sample Filler Photographs

The current study did not show that sample filler exposure benefitted predictive confidence among witnesses who provided those measures one week after witnessing an event. However, it is possible that sample fillers could still benefit predictive confidence judgments if they are presented to witnesses relatively soon after a crime (when variability in memory quality is still high). If future research were to demonstrate this effect and real-world practitioners sought to adopt this recommendation, there are nonetheless practical limitations to consider in the selection of sample filler faces. For instance, the current study included two mock crimes (graffiti and a carjacking) with their corresponding "good" and "poor" filler photographs. These filler photographs were selected from criminal offender databases for the states of New Jersey and Kentucky. For

each crime, 20 “good” filler faces were selected based on matching the target’s physical description and 12 “poor” filler faces were selected based on matching only the target’s sex and race (and allowing other features to vary). The physical descriptions used to select the “good” fillers were generated by pilot participants in a previous study who had a good view of the criminal, allowing those descriptions to be relatively detailed. In the real world, however, witnesses do not always have a clear view and therefore may be able to provide only a very basic description (which could make selecting fair lineup fillers – much less additional sample fillers – difficult).

Furthermore, the final selection of sample fillers in the current study followed pilot testing that collected ratings of sample fillers’ dis/similarity to fillers in the actual lineup. In practice, this type of pilot testing would be impractical to conduct for every lineup administered to a witness. Consequently, law enforcement officers would not be able to determine how the sample fillers and lineup fillers compared (statistically speaking) to each other or to the suspect and one could imagine a situation in which the sample fillers are too dissimilar to the actual lineup photos for them to be an accurate representation of lineup difficulty.

Finally, law enforcement offers may opt to draw their sample filler photographs from their own state’s criminal offender database (rather than using the New Jersey and Kentucky databases). The quality and availability of offender photographs are likely to vary widely between state databases. For instance, New Jersey and Kentucky had the most user-friendly format for selecting physical descriptors (which is why they were selected for the current project). Other state databases that are less user-friendly might

produce additional logistical problem when attempting to find good quality sample fillers to present to witnesses.

Dichotomous Binning Methods

Witnesses were categorized into one of two bins for composite memory (weak vs. strong) and pre/postdictive confidence (low vs. high). We took two approaches to determining witnesses' bin placement: calculating a median split value for each measure within each sample filler condition and performing a series of chi-square tests in each sample filler condition to determine the value that maximally differentiated accurate from inaccurate witnesses. These binning methods share an important limitation in that they require access to data collected in a controlled experiment. Without a distribution of data points, chi square and median split values cannot be calculated. In the real world, police do not have access to this type of data since they often deal with a single witness at a time. They would therefore need pre-defined cutoff points to denote "weak" or "strong" memory and "low" or "high" confidence; considering that there is a great deal of variability in witnessing experiences, creating this metric could pose a serious challenge to researchers.

Future Directions

Although the calibration of witnesses' pre-identification confidence was not improved by enhancing lineup familiarity via sample filler exposure in the current study, researchers should continue to explore ways to harness a reliable predictive confidence-accuracy relationship. The current study, in fact, provided important new insights into potential uses of predictive confidence.

Data from the current study and that of Shambaugh and Charman (in preparation) suggest that the effects of encoding variability on witnesses' memory strength may be weakened after a delay period (such that *all* witnesses end up with weak memory). Without variability in memory strength, predictive confidence is unlikely to be reliably related to identification accuracy; therefore, using predictive confidence for lineup procedures (which are likely to be delayed from the witnessed event) may not be advisable. Importantly, however, both Shambaugh and Charman (in preparation) and Molinaro et al. (2021) found more stable relationships between predictive confidence and accuracy when witnesses underwent encoding, made a predictive judgment, and attempted an identification all in temporal proximity. Thus, it is possible that identification procedures which occur shortly after the witnessing experience are better suited for predictive confidence. For example, show-up identifications, in which a single suspect is shown to a witness, often occur shortly after a crime is committed (Sjöberg, 2016). Without a substantial delay between the witnessed event, predictive confidence judgment, and identification attempt, we would expect different witnesses to have greatly different memory qualities, and therefore we would still expect a reliable predictive confidence-accuracy relationship. Although researchers recommend avoiding show-up procedures when possible (Wells et al., 2020), it is not always feasible to avoid them, and they may still occur with high frequency (Innocence Project, 2022). Recently, researchers have called for the development of procedures that can be used to improve identification outcomes specifically from show-ups (Mook & Charman, in preparation); if predictive confidence and/or memory reports are related to show-up identification accuracy,

screening witnesses on the basis of their self-reported memory could be an effective way to answer that call. Future research should test these ideas.

On a related note, the effects of sample filler exposure should be evaluated when witnesses possess significant variability in memory strength. Given that witnesses' predictive confidence and dichotomous lineup prediction were significantly correlated with underlying memory strength, ensuring that at least some witnesses have a strong memory may give these measures a better chance at reliably predicting accuracy. Thus, future research should consider testing a sample filler manipulation without a retention interval. Alternatively, if a retention interval is desired (to mirror real-world practices), future research should adjust witnesses' encoding distance, duration, (etc.) to accentuate differences in memory strength (such that "strong-memory" witnesses have a *very good* view, and "weak-memory" witnesses have a *very poor* view). Doing so should increase the likelihood that variability in memory strength is maintained over the course of the delay.

An additional avenue for future research on predictive confidence stems from the metacognition literature. Research has shown that asking participants to use visual imagery (i.e., their "mind's eye") can improve the accuracy of metacognitive judgments made about the item(s) imagined (Rademaker & Pearson, 2012). Thus, a possible remedy for strengthening the predictive confidence-accuracy relationship could be to have witnesses picture the criminal's face in their mind's eye prior to assessing their predictive confidence. The ease (or difficulty) with which witnesses can bring the face to mind and the vividness/clarity of that image may be a useful in appropriately calibrating their predictive confidence judgments.

Finally, further clarification is needed regarding the basis of witnesses' predictive confidence judgments. In the current study, we had reasoned that a weak predictive confidence-accuracy relationship was the result of witnesses' inability to accurately anticipate task difficulty (i.e., lineup fillers' similarity to the target). Our results, however, indicate that this reasoning may not be correct. Future research should consider asking witnesses to report the logic behind their predictive confidence judgment; doing so will help researchers better understand *why* witnesses are (or are not) confident in their ability to make a correct identification or rejection, thus being able to focus future experimental manipulations on the factors that most greatly inform predictive confidence judgments.

V. CONCLUSION

Recent studies show that a reliable relationship can exist between witnesses' predictive confidence and identification accuracy when witnesses experience encoding variability and the two judgments are made in temporal proximity (Molinaro et al., 2021), but that witnesses tend to be overconfident when only one identification is made and these judgments are made after a delay (Shambaugh and Charman, in preparation). Results from the current study confirmed that witnesses tend to be substantially overconfident when predicting their future identification accuracy and suggest that this overconfidence is not reduced by familiarizing witnesses with lineup difficulty (at least among witnesses with relatively weak memories). Across the board, witnesses were overconfident; although sample filler exposure successfully reduced witnesses' predictive confidence, they were still overconfident relative to their level of identification accuracy (regardless of whether the sample fillers accurately reflected the difficulty of the lineup). Although these results may be attributed to the lack of variability in memory strength

across witnesses that occurred following a one-week retention interval (which is an important prerequisite for a reliable predictive confidence-accuracy relationship), it is also important to note that such a retention interval (and the corresponding weak memories that result) are likely to also occur among many real-world witnesses. Nonetheless, results suggest that predictive confidence and memory measures should not be used as a means of determining whether to present witnesses with a lineup.

Moving forward, researchers should explore the use of predictive confidence when witnesses possess greater variability in memory strength – for example, with identification procedures that naturally occur in proximity to the witnessed event (i.e., show-ups). Future research should also seek to further understand what specific aspect(s) of memory witnesses reference when formulating their predictive confidence. With additional research, predictive confidence may yet be improved to eventually provide law enforcement with a practical, cost-effective, and informative screening tool. Such a tool could prevent law enforcement from wasting time and resources, and may ultimately prevent a wrongful conviction based on inaccurate eyewitness evidence.

References

Allen, R. M., & Casbergue, R. M. (1997). Evolution of novice through expert teachers' recall: Implications for effective reflection on practice. *Teaching and Teacher Education*, *13*, 741–755.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*, 126-131. http://dx.doi.org/10.1037/h0027455

Aretini, M. (2011). *The effect of task difficulty and domain expertise on metacognitive accuracy* [Unpublished master's thesis]. Universiteit Leiden.

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*, 382–391. http://doi.org/10.1002/acp.3008

Baars, M., van Gog, T., de Bruin, A. & Paas, F. (2016). Effecs of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology*, *7*, 810-834. http://doi.org/10.1080/01443410.2016.1150419

Benton, T. R., Ross, D. F., Bradshaw, E., Thomas, W. N., & Bradshaw, G. S. (2006). Eyewitness memory is still not common sense: comparing jurors, judges and law enforcement to eyewitness experts. *Applied Cognitive Psychology*, *20,* 115–129. http://doi.org/10.1002/acp.1171

Brigham, J. C., & WolfsKeil, M. P. (1983). Opinions of attorneys and law enforcement personnel on the accuracy of eyewitness identifications. *Law and Human Behavior*, *7*, 337–349. https://doi.org/10.1007/BF01044736

Bjorkman, M. (1992). Knowledge, calibration, and the resolution: A linear model. *Organizational Behavior and Human Decision Processes*, *51*, 1-21.

Bornstein, B. Deffenbacher, K. & Penrod, S. (2012). Effects of exposure time and cognitive operations on facial identification accuracy: a meta-analysis of two variables associated with memory strength. *Psychology, Crime & Law*, *5*, 473-490.

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology*, *11*, 3–23. https://doi.org/10.1348/135532505X79672

Cauvin, S., Moulin, C., Souchay, C., Schnitzspahn, K., & Kliegel, M. (2018). Laboratory vs. Naturalistic prospective memory task predictions: Young adults are overconfident outside of the laboratory. *Memory.* Advance online publication. http://dx.doi.org.ezproxy.fiu.edu/10.1080/09658211.2018.1540703

Charman, S. & Cahill, B. (2012). Witnesses' memories for lineup fillers postdicts their identification accuracy. *Journal of Applied Research in Memory and Cognition*, *1*, 11-17. http://doi.org/10.1016/j.jarmac.2011.08.001

Charman, S. D., Carol, R. N. & Schwartz, S. L. (2018). The effect of biased lineup instructions on eyewitness identification confidence. *Applied Cognitive Psychology*, *32*, 287-297.

Charman, S.D. & Mook. A. (in preparation). "Improving show-up identification outcomes by undermining witnesses' beliefs in police suspicions."

Cheng, C-M. (2010). Accuracy and stability of metacognitive monitoring: A new measure. *Behavior Research Methods*, *42*, 715-732.

Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior, 29*, 575 - 604. https://doi.org/10.1007/s10979-005-7121-1

Commander, N. E., & Stanwyck, D. J. (1997). Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology*, *22*, 39–52.

Cutler, B. L., & Penrod, S. D. (1988). Improving the reliability of eyewitness identification: Lineup construction and presentation. *Journal of Applied Psychology*, *73*, 281–290. https://doi.org/10.1037/0021-9010.73.2.281

Cutler, B. L., & Penrod, S. D. (1989). Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence. *Journal of Applied Psychology*, *74*, 650-652. http://dx.doi.org.ezproxy.fiu.edu/10.1037/0021-9010.74.4.650

De Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2007). Improving metacomprehension accuracy and self-regulation in cognitive skill acquisition: The effect of learner expertise. *European Journal of Cognitive Psychology, 19*, 671–688. http://doi.org/10.1080/09541440701326204

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology, 67*, 818-835. http://dx.doi.org/10.1037/0022-3514.67.5.818

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-

analysis. *Psychology, Public Policy, and Law, 19*, 151-164. https://dx.doi.org/10.1037/a0030618

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 238-244. http://dx.doi.org/10.1037/0278-7393.33.1.238

Fleming, S. M. & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B Biological Sciences, 367*, 1338-1349.

Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior, 33*, 70-82. http://dx.doi.org/10.1007/s10979-008-9136-x

Ghatala, E. S., Levin, J. R., Foorman, B. R., & Pressley, M. (1989). Improving children's regulation of their reading PREP time. *Contemporary Educational Psychology, 14*, 49–66.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Earlbaum.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology-Learning Memory and Cognition, 11*, 702-718.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*, 119-136.

Handler, A., & Frühholz, S. (2021). Eyewitness Memory for Person Identification: Predicting Mugbook Recognition Accuracy According to Person Description Abilities and Subjective Confidence of Witness. *Frontiers in Psychology, 13*, 1-15. https://doi.org/10.3389/fpsyg.2021.675956

Hourihan, L., Benjamin, A., & Liu, X. (2012). A cross-race effect in metamemory: Predictions of faced recognition are more accurate for members of our own race. *Journal of Applied Research in Memory and Cognition, 1*, 158-162. http://doi.org/10.1016/j.jarmac.2012.06.004

Innocence Project (2020). Eyewitness Identification Reform. Retrieved November 20, 2020 from https://innocenceproject.org/eyewitness-identification-reform/

Johnston, W.A. & Uhl, C.N. (1976). The Contributions of Encoding Effort and Variability to the Spacing Effect on Free Recall. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 153–60.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1304-1316.

Kalyuga, S., Ayres, P. Chandler, P. & Sweller. J. (2003). The Expertise Reversal Effect. *Educational Psychologist*, *38*, 23-31.

Kassin, S. M. (1985). Eyewitness identification: Retrospective self-awareness and the accuracy–confidence correlation. *Journal of Personality and Social Psychology, 49*, 878-893. http://dx.doi.org/10.1037/0022-3514.49.4.878

Kelemen, W. L., Frost, P. J., & Weaver III, C. A. (2000). Individual differences in metacognition: evidence against a general metacognitive ability. *Memory and Cognition*, *28*, 92–107.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273.

King, J.F., Zechmesiter, E.B., & Shaughnessy, J.J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology*, *93*, 329-343.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting One's Own Forgetting: The Role of Experience-Based and Theory-Based Processes. *Journal of Experimental Psychology: General*, *133*, 643-656. http://doi.org/10.1037/0096-3445.133.4.643

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biase*s (pp. 306-334). Cambridge: Cambridge University Press.

Lindsay, R. C., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*, 79-89. http://dx.doi.org/10.1037/0021-9010.66.1.79

Lovelace, E. A. (1984). Metamemory: Monitoring future recall ability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 756–766.

Malpass, R., & Divine, P. (1981). Eyewitness Identification: Lineup Insructions and the Absence of the Offender. *Journal of Applied Psychology, 66*, 482-489.

Mazzoni, G., & Nelson, T. O. (1993). Metacognitive monitoring after different kinds of monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1263–1274.

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

Molinaro, P. F., Charman, S. D., & Wylie, K. (2021). Pre-identification confidence is related to eyewitness lineup identification accuracy across heterogeneous encoding conditions. *Law and Human Behavior, 45*(6), 524–541. https://doi.org/10.1037/lhb0000452

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Newman, R. (1984). Children's numerical skill and judgements of confidence in estimation. *Journal of Experimental Child psychology, 37*, 107-123.

Nguyen, T., Abed, E., & Pezdek, K. (2018). Posdictive confidence (but not predictive confidence) predicts eyewitness memory accuracy. *Cognitive Research: Principles and Implications*, *3*, 32. http://doi.org/10.1186/s41235-018-0125-4

Nicholson A.S., Yarbrough A.M., Penrod S.D. (2014). Jury Decision Making and Eyewitness Testimony. In Bruinsma G. &Weisburd D. (Eds.), *Encyclopedia of Criminology and Criminal Justice*. Springer. https://doi.org/10.1007/978-1-4614-5690-2_670

Nietfeld, J., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *Journal of Educational Research, 95*, 131-142.

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

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. *Reading Research Quarterly*, *22*, 219–236.

Rademaker, R. L., & Pearson, J. (2012). Training Visual Imagery: Improvements of Metacognition, but not Imagery Strength. *Frontiers in psychology*, *3*, 224. https://doi.org/10.3389/fpsyg.2012.00224

Rhodes, M. (2016). Judgments of Learning: Methods, Data, and Theory. In J. Dunlosky & S. Tauber (Eds.), *The Oxford Handbook of Metamemory (*65-80). Oxford University Press.

Sauer, J., Brewer, N., Zweck, T. & Weber, N. (2010). The Effect of Retention Interval on the Confidence–Accuracy Relationship for Eyewitness Identification. *Law Hum Behav* 34, 337–347. https://doi.org/10.1007/s10979-009-9192-x

Sauerland, M. & Sporer, S. (2009). Fast and Confident: Postdicting Eyewitness Identification Accuracy in a Field Study. *Journal of Experimental Psychology: Applied*, *15*, 46-62.

Semmler, C., Brewer, N., & Douglass, A. B. (2012). *Jurors believe eyewitnesses.* In B. L. Cutler (Ed.), *Conviction of the innocent: Lessons from psychological research* (p. 185–209). American Psychological Association. https://doi.org/10.1037/13085-009

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

Smith, A., Wilford, M., Quigley-McBride, A., & Wells, G. (in press). Mistaken Eyewitness Identification Rates Increase When Either Witnessing or Testing Conditions get Worse.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory & Cognition*, *22*, 63-69.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society*, *30*, 125–128.

Shambaugh, L. J. & Charman, S. D. (in preparation). "The Ability of Pre-Identification Confidence to Predict Subsequent Lineup Accuracy."

Sjöberg, M. P. (2016). The Show-Up Identification Procedure: A Literature Review. *Open Journal of Social Sciences*, *4*, 86-95. http://dx.doi.org/10.4236/jss.2016.41012

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, *78*, 22-33. http://dx.doi.org/10.1037/0021-9010.78.1.22

Sporer, S. L. (1994). Decision times and eyewitness identification accuracy in simultaneous and sequential lineups. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 300-327). Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511759192.015

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

Steblay, N., Dysart, J., Fulero, S. & Lindsay, R. C. L. (2003). Eyewitness Accuracy Rates in Police Showup and Lineup Presentations: A Meta-Analytic Comparison. *Law Hum Behavior*, *27***,** 523–540. https://doi.org/10.1023/A:1025438223608

Stone, N.J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*, 437-473.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201-221.

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *The Quarterly Journal of Experimental Psychology*, *65*, 1376-1396. http://dx.doi.org/10.1080/17470218.2012.656665

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*, 46-50. http://doi.org/10.1111/j.1467-9280.2007.01847.x

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experiment Psychology: Learning, Memory, & Cognition*, *31*, 1267-1280.

Todorov, I., Sundqvist, M. L., Kornell, N., & Jönsson, F.U. (2013). Phrasing Questions in Terms of Current (Not Future) Knowledge Increases Preferences for Cue-Only Judgments of Learning. *Archives of Scientific Psychology*, *1*, 7-13. http://dx.doi.org/10.1037/arc0000002

Townsend, C. & Heit, E. (2011). Judgments of learning and improvement. *Memory & Cognition*, *39*, 204-216.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352–373. https://doi.org/10.1037/h0020071

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22*, 217-237. htpps://dx.doi.org/10.1023/A:1025746220886

Valentine, T., & Mesout, J. (2009). Eyewitness Identification Under Stress in the London Dungeon. *Applied Cognitive Psychology, 23*, 151-161.

Wells, G. L., Lindsay, R. C., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology, 64*, 440-448. http://dx.doi.org/10.1037/0021-9010.64.4.440

Wells, G. L., Kovera, M., Douglass, A., Brewer, N., Meissner, C., & Wixted, J. T. (2020). Policy and Procedure Recommendations for the Collection and Preservation of Eyewitness Identification Evidence. *Law and Human Behavior, 44*, 3-36. https://dx.doi.org/10.1037/lhb0000359

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360-376. http://dx.doi.org/10.1037/0021-9010.83.3.360

Wells, G.L., Rydell, S.M., & Seelau, E.P. (1993). The Selection of Distractors for Eyewitness Lineups. *Applied Cognitive Psychology, 78*, 835-844.

Whittington, J., Carlson, C., Carlson, M., Weatherford, D., Krueger, L., & Jones, A. (2019). Asking an eyewitness to predict their later lineup performance could harm the confidence-accuracy relationship. *Applied Cognitive Psychology, 34*, 782-783.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & Cognition, 26*, 716–730.

Wixted, J.T. & Mickes, L. (2014). A Signal-Detection-Based Diagnostic-Feature-Detection Model of Eyewitness Identification. *Psychological Review, 121*, 262-276.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. III (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*, 515–526. https://doi.org/10.1037/a0039510

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

Table 1.
Mean similarity rating of each filler photograph to the target's photograph (standard deviation in parentheses).

Poor Sample Fillers		Good Sample Fillers		Lineup Fillers	
		Carjacking (*n* =35)			
Filler	Mean Similarity	Filler	Mean Similarity	Filler	Mean Similarity
1	1.60 (0.85)	1	3.60 (1.74)	1	1.57 (0.74)
2	1.86 (1.35)	2	3.51 (1.79)	2	3.60 (1.77)
3	1.20 (0.47)	3	3.34 (1.47)	3	3.80 (1.91)
4	1.46 (0.82)	4	3.06 (1.59)	4	4.14 (1.73)
5	1.20 (0.53)	5	2.83 (1.38)	5	4.26 (1.80)
6	1.40 (0.78)	6	2.54 (1.70)	6	3.83 (1.64)
Poor Sample Fillers		Good Sample Fillers		Lineup Fillers	
		Graffiti (*n* = 40)			
Filler	Mean Similarity	Filler	Mean Similarity	Filler	Mean Similarity
1	2.05 (1.47)	1	3.80 (1.57)	1	2.68 (1.56)
2	1.50 (0.88)	2	3.60 (1.61)	2	3.20 (1.73)
3	1.40 (0.67)	3	3.28 (1.68)	3	2.72 (1.57)
4	2.08 (1.31)	4	3.17 (1.68)	4	4.27 (1.75)
5	1.33 (0.73)	5	3.15 (1.90)	5	2.95 (1.62)
6	1.48 (0.75)	6	3.00 (1.65)	6	3.48 (1.54)

.

Table 2.
Proportion of correct IDs, filler IDs, and misses (TP lineups), and correct rejections and picks (TA lineups) for the carjacking mock crime as a function of encoding distance, duration, and sample filler condition.

	Control			Poor Sample Fillers			Good Sample Fillers		
				Target-Present Lineup					
Encoding	Hit	Filler ID	Miss	Hit	Filler ID	Miss	Hit	Filler ID	Miss
Close									
Short	0.29 (4)	0.36 (5)	0.36 (5)	0.17 (1)	0.50 (3)	0.33 (2)	0.38 (3)	0.38 (3)	0.25 (2)
Long	0 (0)	0.83 (5)	0.17 (1)	0.20 (2)	0.30 (3)	0.50 (5)	0.40 (2)	0.20 (1)	0.40 (2)
Far									
Short	0.33 (2)	0.5 (3)	0.17 (1)	0 (0)	0.50 (4)	0.50 (4)	0.13 (1)	0.38 (3)	0.50 (4)
Long	0 (0)	1.0 (3)	0 (0)	0.22 (2)	0.56 (5)	0.22 (2)	0.26 (5)	0.21 (4)	0.53 (10)
Collapsed	0.26 (6)	0.68 (16)	0.30 (7)	0.15 (5)	0.45 (15)	0.39 (26)	0.28 (11)	0.28 (11)	0.45 (18)

	Control		Poor Sample Fillers		Good Sample Fillers	
			Target-Absent Lineup			
Encoding	Correct Rejection	TA Pick	Correct Rejection	TA Pick	Correct Rejection	TA Pick
Close						
Short	0.11 (1)	0.89 (8)	0.25 (2)	0.75 (6)	0 (0)	1.0 (2)
Long	0.22 (2)	0.78 (7)	0.20 (1)	0.80 (4)	0.17 (1)	0.83 (5)
Far						
Short	0.56 (5)	0.44 (4)	0.22 (2)	0.78 (7)	0.38 (6)	0.63 (10)
Long	0.36 (4)	0.64 (7)	0.44 (4)	0.56 (5)	0.71 (5)	0.29 (3)
Collapsed	0.32 (12)	0.68 (26)	0.29 (9)	0.71 (22)	0.38 (12)	0.63 (20)

Table 3.
Proportion of correct IDs, filler IDs, and misses (TP lineups), and correct rejections and picks (TA lineups) for the graffiti mock crime as a function of encoding distance, duration, and sample filler condition.

	Control			Poor Sample Fillers			Good Sample Fillers		
				Target-Present Lineup					
Encoding	Hit	Filler ID	Miss	Hit	Filler ID	Miss	Hit	Filler ID	Miss
Close									
Short	0.17 (2)	0.42 (5)	0.42 (5)	0.20 (2)	0.60 (6)	0.20 (2)	0.25 (2)	0.25 (2)	0.50 (4)
Long	0.30 (3)	0.30 (3)	0.40 (4)	0.14 (1)	0.43 (3)	0.43 (3)	0.40 (4)	0.10 (1)	0.50 (5)
Far									
Short	0.46 (5)	0.46 (5)	0.09 (1)	0.50 (2)	0.25 (1)	0.25 (1)	0.22 (2)	0.44 (4)	0.33 (3)
Long	0.67 (6)	0.22 (2)	0.11 (1)	0.20 (1)	0.40 (2)	0.40 (2)	0.10 (1)	0.50 (5)	0.40 (4)
Collapsed	0.38 (16)	0.36 (15)	0.26 (11)	0.23 (6)	0.46 (12)	0.31 (8)	0.24 (9)	0.32 (12)	0.43 (16)

	Control		Poor Sample Fillers		Good Sample Fillers	
			Target-Absent Lineup			
Encoding	Correct Rejection	TA Pick	Correct Rejection	TA Pick	Correct Rejection	TA Pick
Close						
Short	0.50 (4)	0.50 (4)	0.38 (3)	0.62 (5)	0.67 (4)	0.33 (2)
Long	0.50 (4)	0.50 (4)	0.70 (7)	0.30 (3)	0.67 (6)	0.33 (3)
Far						
Short	0.11 (1)	0.89 (8)	0.33 (4)	0.67 (8)	0.60 (6)	0.40 (4)
Long	0.71 (5)	0.29 (2)	0.50 (5)	0.50 (5)	0.86 (6)	0.14 (1)
Collapsed	0.44 (14)	0.56 (18)	0.48 (19)	0.53 (21)	0.69 (22)	0.31 (10)

Table 4.
Proportion of correct IDs, filler IDs, and misses (TP lineups), and correct rejections and picks (TA lineups) for the carjacking mock crime as a function of predictive confidence bin and sample filler condition.

	Control			Poor Sample Fillers			Good Sample Fillers		
				Target-Present Lineup					
Predictive Confidence	Hit	Filler ID	Miss	Hit	Filler ID	Miss	Hit	Filler ID	Miss
0-20%	0 (0)	0.17 (5)	0.07 (2)	0.03 (1)	0.09 (3)	0.09 (3)	0.13 (5)	0.10 (4)	0.05 (2)
30-40%	0.03 (1)	0 (0)	0 (0)	0.03 (1)	0.03 (1)	0.12 (4)	0.08 (3)	0.05 (2)	0.20 (8)
50-60%	0.07 (2)	0.21 (6)	0.07 (2)	0.03 (1)	0.12 (4)	0.06 (2)	0.03 (1)	0.13 (5)	0.05 (2)
70-100%	0.10 (3)	0.17 (5)	0.10 (3)	0.06 (2)	0.21 (7)	0.12 (4)	0.06 (2)	0 (0)	0.15 (6)
Collapsed	0.21 (6)	0.55 (16)	0.24 (7)	0.15 (5)	0.46 (15)	0.39 (13)	0.28 (11)	0.28 (11)	0.45 (18)

	Control		Poor Sample Fillers		Good Sample Fillers	
			Target-Absent Lineup			
Predictive Confidence	Correct Rejection	TA Pick	Correct Rejection	TA Pick	Correct Rejection	TA Pick
0-20%	0 (0)	0.13 (5)	0 (0)	0.10 (3)	0.13 (4)	0.23 (7)
30-40%	0.16 (6)	0.16 (6)	0.10 (3)	0.19 (6)	0.13 (4)	0.10 (3)
50-60%	0.11 (4)	0.24 (9)	0.06 (2)	0.23 (7)	0.03 (1)	0.26 (8)
70-100%	0.08 (3)	0.16 (6)	0.12 (4)	0.20 (6)	0.10 (3)	0.03 (1)
Collapsed	0.32 (12)	0.68 (26)	0.29 (9)	0.39 (22)	0.39 (12)	0.61 (19)

Table 5.
Proportion of correct IDs, filler IDs, and misses (TP lineups), and correct rejections and picks (TA lineups) for the graffiti mock crime as a function of predictive confidence bin and sample filler condition.

	Control			Poor Sample Fillers			Good Sample Fillers		
				Target-Present Lineup					
Predictive Confidence	Hit	Filler ID	Miss	Hit	Filler ID	Miss	Hit	Filler ID	Miss
0-20%	0.05 (2)	0.12 (5)	0 (0)	0 (0)	0.15 (4)	0.08 (2)	0.05 (2)	0.12 (4)	0.08 (3)
30-40%	0.17 (7)	0.05 (2)	0.07 (3)	0.12 (3)	0.12 (3)	0.08 (2)	0.12 (4)	0.12 (4)	0.16 (6)
50-60%	0.10 (4)	0.07 (3)	0.10 (4)	0 (0)	0.15 (4)	0.15 (4)	0.08 (3)	0.08 (3)	0.16 (6)
70-100%	0.07 (3)	0.12 (5)	0.09 (4)	0.12 (3)	0.04 (1)	0 (0)	0 (0)	0.03 (1)	0.03 (1)
Collapsed	0.38 (16)	0.36 (15)	0.26 (11)	0.23 (6)	0.46 (12)	0.31 (8)	0.24 (9)	0.32 (12)	0.43 (16)

	Control		Poor Sample Fillers		Good Sample Fillers	
			Target-Absent Lineup			
Predictive Confidence	Correct Rejection	TA Pick	Correct Rejection	TA Pick	Correct Rejection	TA Pick
0-20%	0.03 (1)	0.03 (1)	0.05 (2)	0.13 (5)	0.19 (6)	0.13 (4)
30-40%	0.13 (4)	0.09 (3)	0.15 (6)	0.10 (4)	0.25 (8)	0.13 (4)
50-60%	0.19 (6)	0.22 (7)	0.13 (5)	0.20 (8)	0.16 (5)	0.03 (1)
70-100%	0.09 (3)	0.22 (7)	0.15 (6)	0.11 (4)	0.09 (3)	0.03 (1)
Collapsed	0.44 (14)	0.56 (18)	0.48 (19)	0.53 (21)	0.69 (22)	0.31 (10)

Table 6.
Proportion of correct IDs, filler IDs, and misses (TP lineups), and correct rejections and picks (TA lineups) collapsed across mock crime as a function of predictive confidence bin and sample filler condition.

	Control			Poor Sample Fillers			Good Sample Fillers		
				Target-Present Lineup					
Predictive Confidence	Hit	Filler ID	Miss	Hit	Filler ID	Miss	Hit	Filler ID	Miss
0-20%	0.03 (2)	0.14 (10)	0.03 (2)	0.02 (1)	0.12 (7)	0.08 (5)	0.09 (7)	0.10 (8)	0.06 (5)
30-40%	0.11 (8)	0.03 (2)	0.04 (3)	0.07 (4)	0.07 (4)	0.10 (6)	0.09 (7)	0.08 (6)	0.18 (14)
50-60%	0.09 (6)	0.13 (9)	0.09 (6)	0.02 (1)	0.14 (8)	0.10 (6)	0.05 (4)	0.10 (8)	0.10 (8)
70-100%	0.09 (6)	0.14 (10)	0.10 (7)	0.08 (5)	0.13 (8)	0.07 (4)	0.02 (2)	0.01 (1)	0.09 (7)
Collapsed	0.31 (22)	0.44 (31)	0.26 (18)	0.19 (11)	0.46 (27)	0.36 (21)	0.26 (20)	0.30 (23)	0.44 (34)

	Control		Poor Sample Fillers		Good Sample Fillers	
			Target-Absent Lineup			
Predictive Confidence	Correct Rejection	TA Pick	Correct Rejection	TA Pick	Correct Rejection	TA Pick
0-20%	0 (0)	0.06 (4)	0.06 (4)	0.10 (7)	0.14 (9)	0.14 (9)
30-40%	0.14 (10)	0.10 (7)	0.10 (7)	0.14 (10)	0.13 (8)	0.11 (7)
50-60%	0.10 (7)	0.21 (15)	0.03 (2)	0.18 (13)	0.16 (10)	0.19 (12)
70-100%	0.13 (9)	0.25 (18)	0.21 (15)	0.18 (13)	0.11 (7)	0.02 (1)
Collapsed	0.37 (26)	0.62 (44)	0.39 (28)	0.61 (43)	0.54 (34)	0.46 (29)

Table 7.
Calibration (*C*), Over/Under (*O/U*) and Adjusted Normalized Resolution Index (*ANRI*) values [95% ICIs within square brackets] for choosers and non-choosers, and for pre- and post-identification confidence in each sample filler condition condition.

	C	*O/U*	*ANRI*
	Control		
Pre-ID			
Choosers	0.12 [0.07, 0.17]*	0.25 [0.18, 0.31]**	0.04 [-0.06, 0.14]
Non-Choosers	0.07 [0.01, 0.14]	-0.08 [-0.20, 0.04]	-0.02 [-0.17, 0.13]
Post-ID			
Choosers	0.25 [0.19, 0.32]	0.42 [0.35, 0.49]	-0.02 [-0.08, 0.05]
Non-Choosers	0.14 [0.05, 0.22]	0.01 [-0.12, 0.14]	0.11 [-0.13, 0.15]
	Poor Sample Fillers		
Pre-ID			
Choosers	0.23 [0.17, 0.29]*	0.36 [0.31, 0.44]	0.04 [-0.07, 0.14]
Non-Choosers	0.22 [0.11, 0.32]	0.03 [-0.16, 0.10]	0.09 [-0.09, 0.27]
Post-ID			
Choosers	0.36 [0.29, 0.43]	0.51 [0.44, 0.58]	0.06 [-0.08, 0.20]
Non-choosers	0.23 [0.14, 0.32]	0.01 [-0.13, 0.14]	0.06 [-0.05, 0.18]
	Good Sample Fillers		
Pre-ID			
Choosers	0.18 [0.11, 0.25]	0.37 [0.28, 0.45]	-0.02 [-0.10, 0.07]
Non-Choosers	0.08 [0.03, 0.13]	0.12 [0.02, 0.21]	-0.05 [-0.11, 0.20]
Post-ID			
Choosers	0.23 [0.15, 0.31]	0.43 [0.34, 0.51]	-0.04 [-0.07, -0.01]
Non-Choosers	0.09 [0.04, 0.15]	0.07 [-0.03, 0.17]	-0.05 [-0.10, -0.004]
	Good/Poor Collapsed		
Pre-ID			
Choosers	0.20 [0.16, 0.25]	0.37 [0.31, 0.43]**	0.02 [-0.03, 0.07]
Non-Choosers	0.13 [0.08, 0.17]	0.06 [0.03, 0.14]	0.01 [-0.05, 0.06]
Post-ID			
Choosers	0.29 [0.29, 0.30]	0.47 [0.41, 0.52]	0.02 [-0.03, 0.06]
Non-Choosers	0.14 [0.09, 0.19]	0.04 [0.04, 0.13]	0.01 [-0.05, 0.06]

*Significant difference ($p < .05$) in the Control/Poor Sample Filler comparison
**Significant difference ($p < .05$) in the Control/Collapsed Filler comparison

Table 8.
Hits (TP lineups) and picks (TA lineups) amongst suspect identifiers. Identification outcomes are broken down by CAC confidence bin (predictive and postdictive) and sample filler condition. Low and high confidence bins determined by performing a median split.

	Control		Poor Sample Fillers		Good Sample Fillers	
	Split at 50%		*Split at 50%*		*Split at 40%*	
	Hit	TA Pick	Hit	TA Pick	Hit	TA Pick
Predictive Confidence						
Low	11	23	6	27	14	16
High	11	21	4	16	6	13
	Control		Poor Sample Fillers		Good Sample Fillers	
	Split at 40%		*Split at 40%*		*Split at 30%*	
Postdictive Confidence						
Low	13	29	5	24	14	15
High	9	15	5	19	6	14

Table 9.
Chi-square values (comparing identification accuracy as a function of being above versus equal to or below split) for each predictive and postdictive confidence value amongst suspect identifiers.

	Control	Poor Sample Fillers	Good Sample Fillers
Predictive Confidence			
0%	-	0.74	2.11
10%	1.57	1.88	0.01
20%	0	0.25	0.08
30%	0.19	0.05	0.01
40%	2.83	0.37	1.10
50%	0.03	0.03	1.58
60%	1.18	0.36	0.88
70%	0.18	1.48	0.07
80%	1.28	2.14	0.07
90%	1.57	0.44	1.48
100%	1.57	0.44	1.48
Postdictive Confidence			
0%	0.07	0.87	0.06
10%	0.46	0.42	0.50
20%	0.04	1.07	0.001
30%	0.03	1.16	1.64
40%	0.30	0.11	0.01
50%	1.55	1.82	0.99
60%	0.06	0.50	0.07
70%	1.73	1.32	0.70
80%	2.03	4.38*	-
90%	2.03	-	-
100%	2.03	-	-

*This chi-square comparison was significant but was disregarded because it violated the assumption of an expected cell count of 5

Table 10.
Hits (TP lineups) and picks (TA lineups) amongst suspect identifiers as a function of MSAC memory strength bin and sample filler condition. Memory strength bin determined by performing median split.

	Control		Poor Sample Fillers		Good Sample Fillers	
	Split at 3.83		*Split at 4.00*		*Split at 3.17*	
Memory Strength	Hit	TA Pick	Hit	TA Pick	Hit	TA Pick
Weak	11	18	7	23	9	17
Strong	11	26	3	20	11	12

Table 11.
Chi square values (comparing identification accuracy as a function of being above versus equal to or below split) for each composite memory value amongst suspect identifiers.

Memory Score	Control	Poor Sample Fillers	Good Sample Fillers
1	0.51	-	1.48
1.33	0.51	-	1.48
1.5	0.26	0.48	0.07
1.67	0.26	0.48	0.07
1.83	0.26	0.48	0.07
2	0.53	0.25	0.07
2.17	0.11	0.004	0.99
2.33	0.11	0.004	1.58
2.5	2.32	0.004	0.11
2.67	2.67	0.14	0.66
2.83	2.67	0.14	0.66
3	1.20	0	0
3.17	0.73	0.09	0.88
3.33	0.73	0.09	0.88
3.5	0.32	0.56	0.55
3.67	0.49	0.81	0.16
3.83	0.49	0.81	0.16
4	0.30	0.9	0.03
4.17	0.31	0.31	0.21
4.33	0.31	0.31	0.21
4.5	1.38	0.09	0.99
4.67	0.39	0.02	0.90
4.83	0.39	0.02	0.90
5	0.77	0.01	0.002
5.17	0	0.25	2.20
5.33	0	0.25	2.20
5.5	0.11	0.02	0.70
5.67	0.11	0.01	-
5.83	0.11	0.01	-
6	0.26	0.44	-
6.17	0.26	0.44	-
6.33	2.03	0.44	-
6.5	-	1.32	-
7	-	1.32	-

Table 12.
Hits (TP lineups) and picks (TA lineups) amongst suspect identifiers as a function of witnesses' dichotomous identification prediction and sample filler condition.

	Control		Poor Sample Fillers		Good Sample Fillers	
	Hit	TA Pick	Hit	TA Pick	Hit	TA Pick
Dichotomous Prediction						
No	12	14	5	19	17	22
Yes	10	30	5	24	3	7

Note: Dichotomous prediction made in response to question: "If you were shown a lineup of six people that may or may not contain the criminal who committed the crime you witnessed, do you believe that you could make an accurate identification decision?"

Table 13.
Parameter estimates for binary logistic regression models predicting identification accuracy from witnesses' dichotomous lineup prediction, sample filler condition, and the interaction, as a function of sample filler comparison.

	Control vs. Poor Fillers				Control vs. Good Fillers				Poor Fillers vs. Good Fillers			
	ß	SE	Wald	*p*	ß	SE	Wald	*p*	ß	SE	Wald	*p*
Dichotomous Prediction	0.68	0.41	2.73	.098	0.33	0.37	0.78	.378	-0.60	0.39	2.35	.125
Sample Filler Condition	0.33	0.37	0.77	.378	-0.14	0.36	0.16	.691	-0.82	0.36	5.33	.022*
Interaction	-0.93	-0.54	2.96	.085	-0.05	0.53	-0.01	.930	0.88	0.55	2.61	0.11
Intercept	0.52	0.29	3.24	.072	0.52	0.29	3.24	.072	1.20	0.29	16.73	<.001

Note: Dichotomous prediction made in response to question: "If you were shown a lineup of six people that may or may not contain the criminal who committed the crime you witnessed, do you believe that you could make an accurate identification decision?"

*Significant difference at $p < .05$ level

APPENDICES

Appendix A

Secondary Memory (Metamemory) Questions

1. How good a view of the criminal did you have?
    a. 1 (*Not a good view at all*) – 7 (*Extremely good view*)
2. How long would you estimate the criminal's face was in sight?
    a. 1 (*Not long at all*) – 7 (*Extremely long*)
3. How well were you able to make out specific features of the criminal?
    a. 1 (*Not well at all*) – 7 (*Extremely well*)
4. How much attention did you pay to the criminal?
    a. 1 (*No attention at all*) – 7 (*My complete attention*)
5. How strong is your memory of the criminal?
    a. 1 (*Not at all strong*) – 7 (*Extremely strong*)
6. How clear an image of the criminal do you have in your mind?
    a. 1 (*Not clear at all*) – 7 (*Extremely clear*)

Appendix B

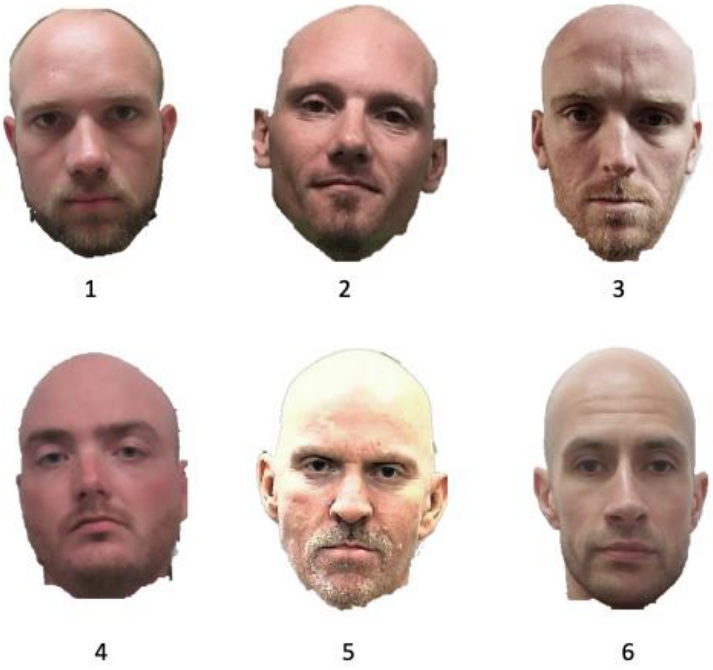
Post-ID Memory Questions

1. How confident were you in your identification decision?
    a. 0% (*Not at all confident*) – 100% (*Extremely confident*)
2. To what extent do you feel that you had a good basis (enough information) to make an identification?
    a. 1 (*No basis at all*) – 7 (*An extremely good basis*)
3. How easy or difficult was the identification task for you?
    a. 1 (*Extremely easy*) – 7 (*Extremely difficult*)
4. After you were first presented with the photos, how long do you estimate it took you to make an identification?
    a. 1 (*Not long at all*) – 7 (*Extremely long*)
5. How willing would you be to testify that you made the correct identification decision?
    a. 1 (*Not at all willing*) – 7 (*Extremely willing*)

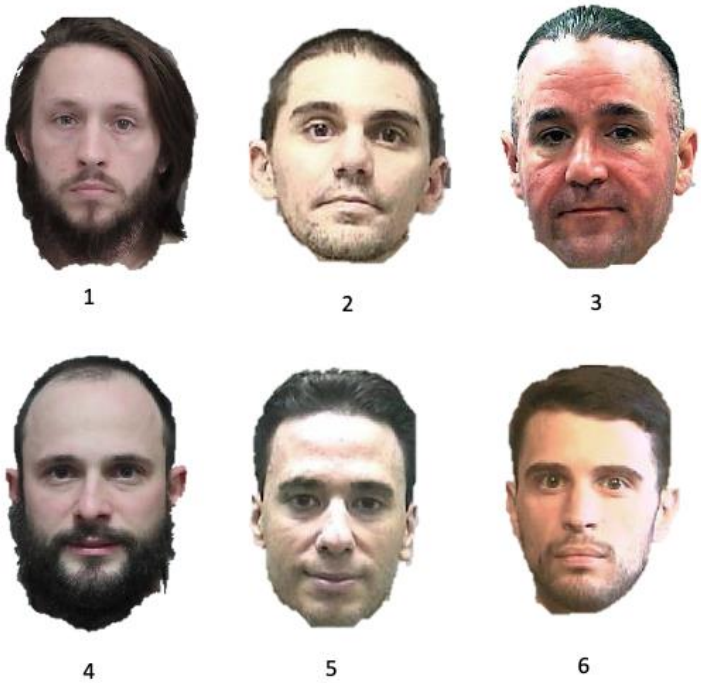
Appendix C

Graffiti Sample Fillers

"Good" Sample Fillers



"Poor" Sample Filers

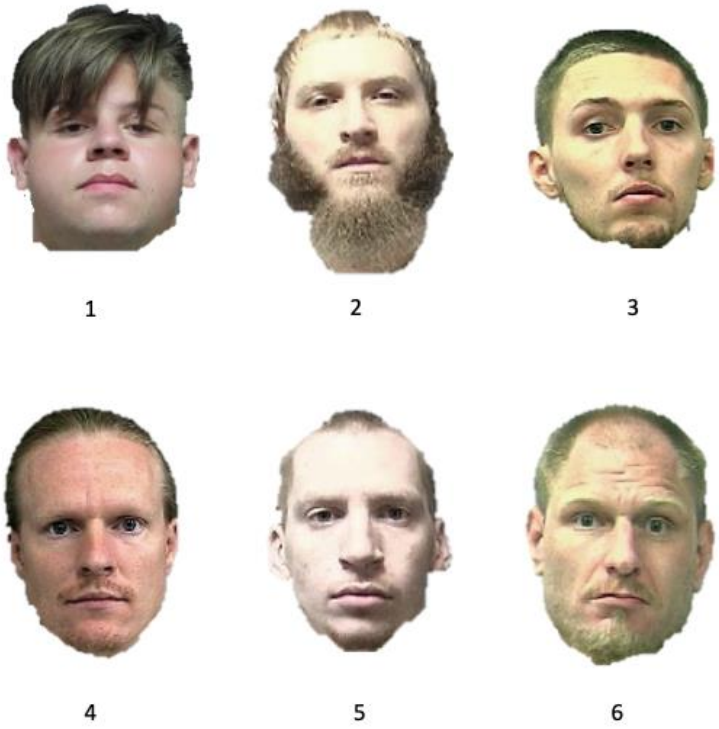


Appendix D

Carjacking Sample Fillers

"Good" Sample Fillers

"Poor" Sample Filers



Appendix E

Carjacking Lineups

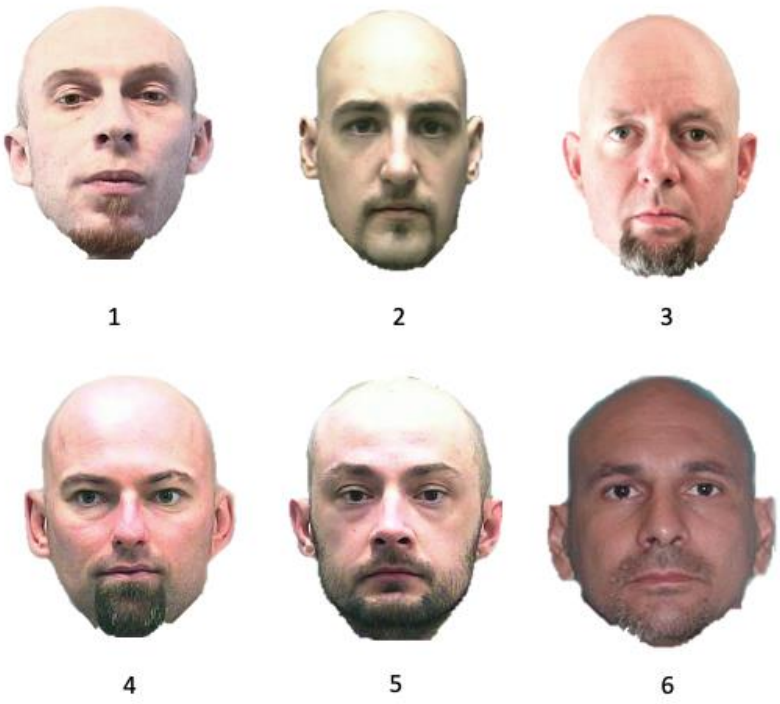
Target Present Lineup (Example)

Target Absent Lineup (Example)

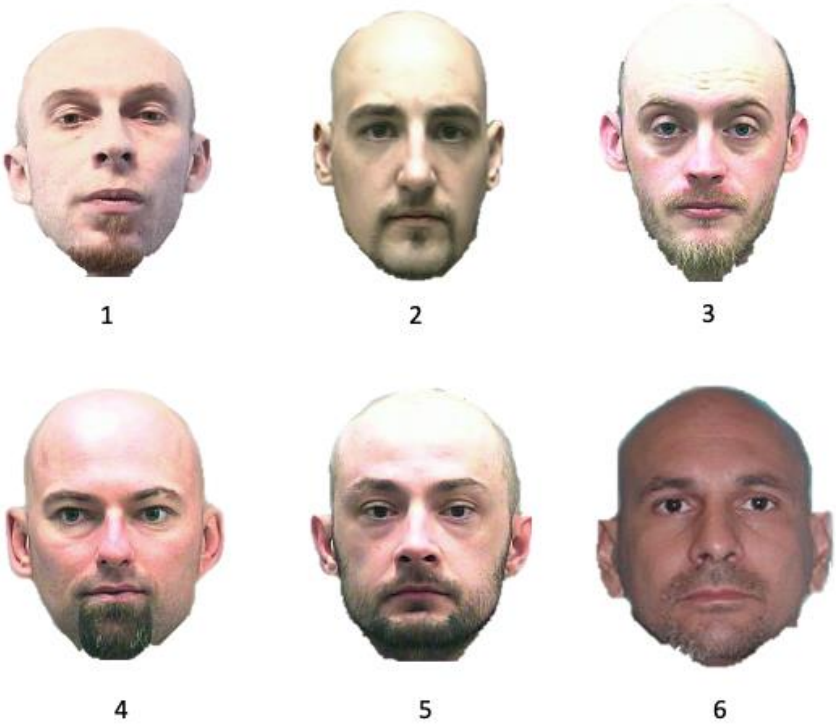
Appendix F

Graffiti Lineups

Target Present Lineup (Example)



Target Absent Lineup (Example)



VITA

LAURA J. SHAMBAUGH

Born, Cedar Rapids, Iowa

2013-2017	B.S., Psychology Iowa State University Ames, Iowa
	B.S., World Languages and Cultures (Spanish) Iowa State University Ames, Iowa
2017-2020	M.S., Experimental Psychology Florida International University Miami, Florida
2020-2022	Doctoral Candidate, Psychology Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Charman, S.D., Shambaugh, L.J., Cahill, B., & Molinaro, P. (in press). "The accuracy of high-confidence lineup identifications is undermined by the appearance-change instruction and target appearance change."

Charman, S.D. & Shambaugh, L.J. (2022). "Strategies for Improving the Quality of Alibi Evidence: A System Variable Approach". In J. Behl & M. Kienzle (Eds.), *Alibis and Corroborators: Psychological, Criminological, and Legal Perspectives* (55-73). Springer Cham.

Goldfarb, D., Chae H., & Shambaugh, L. (2021). Navigating Tricky Waters: Understanding and Supporting Children's Testimony About Experiencing and Witnessing Violence. In G. Calloway & M. Lee (Eds.), *Handbook of Children in the Legal System: A Guide for Forensic and Mental Health Practitioners* (84-109). Routledge.

Shambaugh, L.J. & Charman, S.D. (March 2021). Delay moderates the relationship between pre-identification confidence and subsequent lineup accuracy. Poster presented at the 2021 American Psychology-Law Society Annual Conference, virtual.

Shambaugh, L.J., Charman, S.D., & Bradfield Douglass, A. (July 2021). Prior Knowledge of Case Information Biases Evaluators' Impressions of Videotaped

Eyewitness Identification Evidence. Paper talk presented at the 2021 Society for Applied Research in Memory and Cognition Conference, virtual.

Shambaugh, L., Mansour, J.K., & Vallano, J.P. (September 2021). "Much Ado About Tattoos", APA Monitor on Psychology, *52*(6).

Shambaugh, L.J., Vallano, J.P., & Charman, S.D. (March 2019). Informant testimony and witness confidence: An extension of the Selective Cue Integration Framework. Paper talk presented at the 2019 American Psychology-Law Society Annual Conference, Portland, OR.

Shambaugh, L.J., Vallano, J.P., & Charman, S.D. (March 2020). Informant testimony and witness confidence: Testing the "search" and "evaluation" stages of the SCIF. Paper talk presented at the 2020 American Psychology-Law Society Annual Conference, New Orleans, LA.

Vallano, J.P., Pickell, K.L., & Shambaugh, L.J. (2020). Legal Perspectives on Historical Misconduct Cases: Issues with Civil and Criminal Cases. In J. Pozzulo, E. Pica, & C. Sheahan (Eds.), *Memory and Sexual Misconduct* (175-197). Routledge.

Vallano, J.P. & Shambaugh, L.J. (July 2019). "I'll Take the Deal, if You Think it's Best", APA Monitor on Psychology, *50*(7).