Mock-juror evaluations of traditional and ratings-based eyewitness identification evidence.

James D. Sauer & Matthew A. Palmer

University of Tasmania

Neil Brewer

Flinders University

James D. Sauer**,** School of Medicine (Psychology), University of Tasmania, Hobart, Tasmania 7001, Australia; Matthew A. Palmer, School of Medicine (Psychology), University of Tasmania, Launceston, Tasmania, Australia; Neil Brewer, School of Psychology, Flinders University, Adelaide, South Australia, Australia. This research was supported by funding from the Criminology, Law, and Police Studies Research Group and the Institute for the Study of Social Change at the University of Tasmania, Australian Research Council grants DP140103746 to M. Palmer et al. and DP150101905 to Neil Brewer et al., and a Flinders Research Grant. We are grateful to the class members of our undergraduate Psychology & Law class for their assistance with data collection for Experiment 4.

Corresponding Author:

James D. Sauer**,** School of Medicine (Psychology), University of Tasmania, Locked Bag 30, Hobart, Tasmania 7001, Australia

Email: Jim.Sauer@Utas.edu.au

Abstract

Compared to categorical identifications, culprit likelihood ratings (having the witness rate, for each lineup member, the likelihood that the individual is the culprit) provide a promising alternative for assessing a suspect's likely guilt. Four experiments addressed two broad questions about the use of culprit likelihood ratings evidence by mock-jurors. First, are mock-jurors receptive to non-categorical forms of identification evidence? Second, does the additional information provided by ratings (relating to discrimination) affect jurors' evaluations of the identification evidence? Experiments 1 and 1A manipulated confidence (90% vs. 50%) and discrimination (good, poor, no information) between-participants. Evaluations were influenced by confidence, but not discrimination. However, a within-participant manipulation of discrimination (Experiment 2) demonstrated that evidence of good discrimination enhanced the persuasiveness of moderate levels of confidence, while poor discrimination reduced the persuasiveness of high levels of confidence. Thus, participants can interpret ratings-based evidence, but may not intuit the discrimination information when evaluating ratings for a single identification procedure. Providing detailed instructions about interpreting ratings produced clear discrimination effects when evaluating a single identification procedure (Experiment 3). Across four experiments, we found no evidence that mock-jurors perceived non-categorical identification evidence to be less informative than categorical evidence. However, jurors will likely benefit from instruction when interpreting ratings provided by a single witness.

Keywords: eyewitness identification; confidence; ratings; juror evaluations of evidence

Mock-juror evaluations of traditional and ratings-based eyewitness identification evidence.

Jurors perform an important task under difficult conditions: Non-experts must often assess complex or ambiguous information to reach a decision of significant consequence. This may underlie the persuasiveness of eyewitness identification evidence for jurors (Semmler, Brewer, & Douglass, 2011). Jurors' reliance on identification evidence might reflect their desire for an apparently clear-cut indication of guilt in a setting often characterized by ambiguity. However, current identification practice confers two major limitations on identification evidence. First, eyewitness identification decisions are prone to error (e.g., Steblay, Dysart, & Wells, 2011; Wells, 1993). Second, a categorical identification is not necessarily as informative as it may appear at face value. An identification of the suspect indicates that, of the lineup members presented, the suspect probably provided the best match to the witness's memory of the culprit. However, it does not speak to how closely the suspect matched the witness's memory in an absolute sense, or the extent to which the suspect was favored over the other lineup members (Sauer & Brewer, 2015). This information is important when assessing the diagnostic value of a suspect identification. Thus, although jurors find categorical identification evidence compelling, this relatively coarse index obscures important information relating to the witness's recognition of the suspect.

Further, jurors are not particularly adept at discriminating between identifications that are likely to be correct and those that are likely to be incorrect. Jurors seem relatively insensitive to the effects of viewing conditions, retention interval, and lineup fairness on the reliability of identification decisions (Cutler, Penrod, & Stuve, 1988; though see Reardon & Fisher, 2011, for evidence of discrimination based on witness behaviour during the identification task). Thus, current identification practices produce information that is prone to error and quite possibly of limited informational value. A relatively recent procedural innovation may help address these limitations. Ratings-based identification protocols avoid categorical identification responses and, instead, have witnesses rate the degree to which each lineup member matches their memory of the culprit (Brewer, Weber, Wootton, & Lindsay, 2012; Sauer, Brewer, & Weber, 2008, 2012a). We examined mock-jurors' evaluations of this non-categorical form of identification evidence, and whether mock-jurors' evaluations of identification evidence could benefit from the additional information provided by ratings-based protocols. This research represents an important step in understanding how jurors will respond to ratings-based evidence, and provides further insight about the types of information that shape jurors' interpretations of identification evidence.

Ratings-based Identification Evidence

An identification test probes whether or not the witness recognizes the suspect, thereby implicating the suspect in the matter under investigation. Thus, the critical constructs at play are (a) the quality of the witness's memory for the culprit and (b) the extent to which the suspect matches this memory. However, various non-memorial factors can influence a witness's decision criterion, affecting the likelihood that a suspect will be identified (or that a lineup will be rejected) and compromising the diagnostic value of the identification decision. An appropriately-collected confidence rating may augment the information provided by a categorical identification (i.e., by providing an index of the strength of recognition), but systematic protocols for collecting retrospective confidence ratings are rare in applied settings (Sauer & Brewer, 2015). Collecting ratings for each lineup member, rather than a single categorical decision, was proposed as a method for attenuating non-memorial influences on witness's decision criteria, providing a more informative index of recognition, and allowing a more valid assessment of the witness's memory for the suspect (Sauer et al., 2008).

Early work has supported the contribution of ratings-based protocols to improving the reliability and informational value of identification evidence. Although only a few experiments have examined the issue, results to date consistently show that, compared to standard categorical identification decisions, ratings-based approaches have provided a more sensitive approach to assessing the likely guilt of a suspect (Brewer et al., 2012; Sauer et al., 2008, 2012a). Initial analytical approaches relied on the application of classification criteria to determine when patterns of ratings could be taken as an indication of suspect guilt. Briefly, this analysis involved first identifying whether a pattern of ratings included a single-highest, *max* rating and whether this rating implicated the suspect. Classification criteria were then applied to determine whether or not a pattern of ratings indicated suspect guilt based on the magnitude of the *max* rating and the difference between the *max* and non-*max* ratings. This approach was useful for comparing the diagnostic utility of ratings-based and categorical forms of identification evidence, but reduced the richness of the available identification evidence. Addressing this shortcoming, Brewer et al.’s (2012) profile analyses – examining how the likely guilt of a suspect varied as a function of the discrepancy between the *max* and next highest rated lineup members – demonstrated that the guilt of the suspect varied almost monotonically according to the degree to which the suspect was favored over the alternative lineup members.

Brewer et al.’s (2012) analyses revealed that the potential benefits of ratings-based identification evidence extend beyond improving the reliability of categorical classifications of suspect guilt. In this context, the confidence rating given to the suspect can be conceptualized as an index of recognition for the suspect (ranging from weak to strong), while ratings given to non-suspect lineup members can be conceptualized as indices of discrimination. Admittedly, this is a coarse distinction. However, it serves as a conceptual framework for the issue under investigation in the present research. Specifically, does information pertaining to the witness's ability to discriminate between alternatives affect mock-jurors' perceptions of identification evidence? Thus, compared to categorical identification decisions, ratings-based identification evidence provides a richer source of information about the quality of the witness's memory for the culprit, the degree of their recognition for the suspect, and their ability to discriminate between previously seen and unseen lineup members (Sauer & Brewer, 2015). They may, therefore, benefit jurors' assessments of (a) identification reliability and (b) defendant guilt.

In a broader sense, Brewer et al.'s (2012) findings highlight the potential to eschew traditional, binary conceptualizations of identification evidence (i.e., the witness either identified the suspect or they did not) in favor of more probabilistic interpretations reflecting the extent to which the witness (a) recognized the suspect and (b) was able to discriminate between lineup members (see Brewer & Wells, 2009). Whereas current identification practice assigns a witness's recognition memory 'output' to categories such as "It's number 2" or "She's not there", triers-of-fact may benefit from an approach that increases the amount of information available from the witness's memory, and from considering what this information says about the likely guilt of the defendant (Sauer & Brewer, 2015). However, before this idea can be entertained, it is important to understand how triers-of-fact respond to non-categorical identification evidence. As Brewer and Wells (2011) suggest, a procedure that does not involve a witness actually picking or rejecting a suspect may encounter strong resistance in criminal justice settings. Moreover, jurors may have trouble interpreting this form of evidence.

Jurors' Interpretations of Ratings-Based Evidence

Jurors have difficulty making effective use of complex (e.g., Heuer & Penrod, 1994; Horowitz, ForsterLee, & Brolly, 1996) or probabilistic evidence (e.g., Goodman, 1992; Thompson & Schumann, 1987). Therefore, jurors may be resistant to, or experience difficulty interpreting, ratings-based (i.e., non-categorical) identification evidence. Further, as stated earlier, jurors often fail to consider information that adds nuance when evaluating eyewitness identification evidence (e.g., Cutler, Penrod, & Stuve, 1988). Although the additional information provided by ratings-based identification evidence – relating to the relationship between memory quality, discrimination, and the ratings given to a target versus fillers in a lineup – may be readily understood by memory researchers, this may not be true for lay-people. Given the problematic views of memory commonly espoused by jury-eligible samples and decision-makers in the criminal justice system and (e.g., Simons & Chabris, 2011; Wise & Safer, 2004, 2010), it may be unrealistic to expect jurors to intuitively recognize the value of the additional information provided by ratings-based evidence (i.e., relating to the witness's ability to discriminate between previous seen and unseen lineup members). Lay-people may simply lack the cognitive framework necessary to interpret this information. Thus, jurors may be insensitive to the additional information provided by ratings given to fillers in the lineup.

However, other work paints a more promising picture. Tenney, MacCoun, Spellman, and Hastie (2007) presented mock-jurors with testimony provided by a prosecution witness and pointing to the guilt of a defendant. Mock-jurors found the incriminating testimony more persuasive when the witness expressed high (cf. low) confidence, unless cross-examination revealed that the witness made an error regarding one detail of their testimony (the time that the defendant was allegedly seen leaving the crime scene). If the witness made an error regarding this detail, confident witnesses were deemed less credible than unconfident ones. These results suggest that, under some circumstances, jurors might be sensitive to information that relates to a witness's ability to discriminate between correct and incorrect responses.

Tenney et al. (2007) proposed two mechanisms that could account for the observed effects of this additional information on juror evaluations. First, jurors might draw inferences about a witness's metacognitive knowledge based on the accuracy and confidence of their testimony. If a witness makes a highly confident error, jurors may conclude that the witness shows poor discrimination (i.e., the witness cannot use confidence to distinguish accurate from inaccurate information). Hence, other details of the witness's testimony are unreliable, even if accompanied by high confidence. In contrast, if the witness makes an error with low confidence, this suggests that the witness can discriminate between parts of their testimony that are more or less likely to be correct. Thus, details of the witness's testimony that are made with high confidence are more likely to be accurate. Second, jurors might use confidence and accuracy information to draw inferences about a witness's motivation. For example, if a witness expresses high confidence in all parts of their testimony (accurate and inaccurate), jurors might conclude that the witness's use of confidence is motivated by a desire to be persuasive, rather than an attempt to index the reliability of the information provided.

In the context of ratings-based identification evidence, we are concerned with the first of these two mechanisms: The extent to which evidence of discrimination affects jurors' evaluation of the identification evidence against the suspect. We expect jurors to find identification evidence more persuasive when a witness is highly (e.g., 90%) rather than moderately (e.g., 50%) confident in their recognition of the suspect (e.g., Bradfield & Wells, 2000; Brewer & Burke, 2002; Cutler, Penrod, & Dexter, 1990; Lindsay, Wells, & Rumpel, 1981). However, if the witness also gives relatively high ratings to each other lineup member (e.g., 50-70%), jurors might conclude that the witness shows a poor ability to discriminate between lineup members. Thus, information about the high ratings given to other lineup members should reduce the impact of the high rating for the suspect on juror judgments about the likely guilt of the defendant.

In contrast, and assuming the lineup was unbiased, if a high rating for the suspect was accompanied by low ratings for the other lineup members, this would not suggest poor discrimination. Hence, information about the ratings given to other lineup members should not reduce the impact that a high rating for the suspect has on juror judgments of likely guilt. In fact, a high suspect rating combined with low filler ratings could be interpreted as evidence that (a) the suspect provided a high degree of match to the witness's memory for the culprit and (b) all fillers provided only a low degree of match. This, in turn, might indicate effective discrimination on the part of the witness, and potentially increase the impact of a high rating for the suspect on juror assessments of the defendant's guilt. In sum, Tenney et al.'s (2007) results suggest that the extent to which juror judgments about the guilt of the defendant are influenced by the rating given to the suspect may depend upon the ratings given to other lineup members.

From an applied perspective, this research is necessary to understand how jurors evaluate ratings-based identification evidence because this approach represents a promising alternative to current identification practice. If empirical support for a ratings-based approach continues to grow, thereby opening up the possibility of it being used in police investigations, we must consider how jurors will interpret the resulting evidence. From a theoretical viewpoint, this research will offer some insight about the types of information that shape jurors' interpretations of identification evidence. Prior research has demonstrated that jurors' evaluations of eyewitness evidence are influenced by information relating to memory quality in some situations (e.g., Tenney et al., 2007) but not others (e.g., Cutler et al., 1988). It may be that jurors are simply more receptive to such information when evaluating the credibility of a witness's account of the events surrounding a crime (as per Tenney et al., 2007) than when evaluating the credibility of an identification decision. Alternatively, and perhaps more likely, it may be that whereas Tenney et al.'s manipulation spoke directly to the reliability of the witness in question (i.e., pointing to the witness's ability to monitor the accuracy of their memory reports), information relating viewing or testing conditions to identification reliability is more general in nature. Jurors may be more likely, or better able, to use witness-specific (cf. general) indicators of reliability. If this is true, jurors may consider ratings that index a witness's ability to discriminate between lineup members when assessing the witness's recognition of the defendant.

The Present Research

Given (a) the limitations of categorical identification evidence, (b) the promising findings relating to the diagnostic utility of ratings-based evidence, and (c) the capacity for ratings to provide a richer source of information about the witness's memory for the suspect, we investigated mock-jurors' evaluations of this novel, ratings-based (cf. categorical) form of identification evidence. Four experiments addressed two broad questions. First, are mock-jurors receptive to non-categorical forms of identification evidence, or do they generally find the non-categorical evidence less compelling than a categorical identification? Second, does the additional information provided by ratings-based protocols (relating to the witness's ability to discriminate between lineup members) affect jurors' evaluations of the evidence against a defendant?

In each experiment we presented mock-jurors with a trial transcript containing incriminating identification evidence obtained using either a standard identification task (providing a categorical identification response and associated confidence judgment) or a rating procedure where, for each lineup member, the witness provided a rating indicating the likelihood that that person was the culprit. Thus, in ratings conditions, instead of reading that the witness identified the suspect with a particular level of confidence (e.g., 90%), jurors read that the witness provided a rating for the suspect (e.g., 90%) and for each of the other lineup members (e.g., between 0% and 10%). Our focus was on whether information about the ratings given to other lineup members would shape jurors' interpretations of the rating given to the suspect. As stated earlier, we know that jurors find identification decisions more persuasive of guilt if they are made with relatively high (e.g., 90%) versus low confidence (e.g., 50%). But what if jurors learn that in addition to being 90% confident that the suspect was the culprit, the witness was also 70% confident that each of the other five lineup members was the culprit? Will the high rating given to the suspect trump the moderately high ratings given to other lineup members, or will information about the ratings given to other lineup members color jurors' interpretations of the high rating given to the suspect? Based on Tenney et al.'s (2007) findings, we expected that a high rating for the suspect would have less effect on juror judgments when other lineup members also receive high (indicating poor discrimination), rather than low (indicating good discrimination) ratings. Further to this interaction, and based on the well-documented effects of witness confidence on mock-juror decision-making (e.g., Bradfield & Wells, 2000; Brewer & Burke, 2002; Cutler et al., 1990; Lindsay et al., 1981), we predicted a main effect of confidence, with guilty verdicts occurring more frequently in the high (cf. moderate) confidence conditions. We had no *a priori* basis for predicting any main effects of discrimination.

Experiment 1

Method

Design and participants. Experiment 1 used a 2 (confidence: high, moderate) × 3 (discrimination: good, poor, none) between-subjects design (see Table 1). One hundred and thirty-three undergraduate student participants (76 female; mean age = 21 years, range = 17-47 years) were randomly allocated across the six experimental conditions.

Materials. A transcript from a fictional criminal trial described a police officer giving evidence-in-chief. The officer described an act of arson committed in a domestic property, an account given by a witness at the scene, and the procedures used to gather identification evidence from that witness on a later date. The transcripts differed only in the nature of the identification evidence reported. In all conditions, the identification evidence pointed to the guilt of the defendant. Participants in the standard identification conditions were told that the witness identified the police suspect with either 90% (high) or 50% (moderate) confidence. In the other conditions, the transcript reported that a witness had viewed a lineup containing the police suspect and been asked to rate (out of 100) how confident they were that each lineup member was the person they had seen committing the crime in question. The witness gave the suspect a rating of 90% (high confidence condition) or 50% (moderate confidence condition). In the good discrimination conditions, the witness gave confidence ratings of 0-10% to other lineup members. In the poor discrimination condition, other lineup members were given ratings of either 50-70% (high confidence condition) or 20-40% (moderate confidence condition). No information was provided about whether the lineup members were presented simultaneously or sequentially. However, in all conditions (and in all subsequent experiments) participants were informed that the lineup contained one suspect accompanied by known-innocent fillers.

Procedure. Participants were tested in small groups in a quiet classroom setting. Each participant received a test booklet containing a transcript and a series of questions. After reading the transcript, participants provided demographic information (age and sex), a verdict based on the evidence presented (guilty or not guilty), a confidence rating in their verdict (from 0% - *completely uncertain* to 100% - *completely certain*), and an assessment of likely guilt. To allow clearer comparisons with Tenney et al.'s (2007) results, we combined the verdict and verdict confidence measures to create a verdict preference score (see

Supplementary materials). After completing these measures, participants also completed a manipulation check requiring them indicate the level of confidence given to the suspect and (in the discrimination conditions) the fillers.

Ethics approval. All experiments were approved by the Flinders University Social and Behavioral Research Ethics Committee.

Results and Discussion

We present analyses based on the binary verdict measure, with odds ratio (*OR*) presented as indices of effect size, with 95% CIs reported in brackets. For comparison with Tenney et al. (2007), we direct interested readers to our analyses of verdict preference scores in the supplementary materials.

A 2 (confidence) × 3 (discrimination) × 2 (verdict) hierarchical log linear analysis retained only the confidence × verdict association, $\chi^2(1, N = 133) = 13.81$, $p < .001$, with follow-up cross-tabulation analysis revealing a higher rate of guilty verdicts in the high confidence (68.2% 95% CI [56.9, 79.4]) than low confidence (37.3% [25.7, 48.9]) conditions, $\chi^2(1, N = 133) = 12.71$, $p < .001$, $OR = 3.60$ [1.76, 7.37]. Thus, consistent with the hypothesized main effect of confidence, a guilty verdict was approximately three and a half times more likely in the high, compared to low, confidence conditions. Including the Confidence × Discrimination × Verdict association (i.e., testing the interactive effect of confidence and discrimination on verdict) did not improve the fit of the model, $\chi^2(2, N = 133) = 2.04$, $p =.361$.

The anticipated interaction between confidence and discrimination did not emerge. Perhaps participants in the discrimination conditions misinterpreted the meaning of the ratings given to lineup members other than the suspect. Rather than viewing ratings as an index of the degree of match between each lineup member and the witness’s memory of the perpetrator, participants may have interpreted each rating as an index of certainty in the accuracy of an implicit decision for each lineup member (i.e., selection of the suspect or rejection of each filler). A misinterpretation along these lines might have led some mock-jurors to think that the witness in the good discrimination condition was actually expressing low confidence in their decision overall (e.g., "This witness was 90% sure about one lineup member, but the low ratings for the others indicate that they were not really certain"). Conversely, in the poor discrimination condition, some mock-jurors may have thought the witness was expressing relatively high certainty in their decision overall (e.g., "This witness is pretty sure about each of the lineup members, so their testimony is probably accurate"). These misinterpretations would lower the persuasiveness of evidence in the good discrimination condition and enhance it in the poor discrimination condition, diluting any effects of the discrimination manipulation. Experiment 1A tested this explanation by presenting the identification evidence in the discrimination conditions as the witness's ratings of the degree of match between each lineup member and the culprit (i.e., more clearly indicating the theoretical basis for these ratings, Sauer et al., 2008; Sauer, Weber, & Brewer, 2012b), rather than the witness's *confidence* that each lineup members was the culprit. However, given the conceptual similarity of these two constructs, we retain the "confidence" label to enhance clarity for the reader[1].

Experiment 1A

We expected a main effect of confidence, with guilty verdicts being more likely to occur in the high (cf. moderate) confidence conditions. However, again our primary interest was in testing the hypothesized interaction between confidence and discrimination. Based on Tenney et al.'s (2007) findings, we expected that the effect of confidence on juror judgments would be reduced in the poor (cf. good) discrimination condition.

Method

[1] This applies for Experiments 1A, 2, and 3. We thank a reviewer for this suggestion.

One hundred and forty-two undergraduate students and community members (106 female; mean age = 26 years, range = 16-74 years) participated. The design, materials, and procedures for Experiment 1A were identical to those used in Experiment 1 with one exception: Participants read that the witness was asked to rate the *degree of match* between each lineup member and their memory of the culprit, rather than their *confidence* that each lineup member person was the culprit. In all conditions, the numerical values used to operationalize discrimination were identical to those in Experiment 1.

Results and Discussion

Experiment 1A produced results very similar to those of Experiment 1. The 2 (confidence) × 3 (discrimination) × 2 (verdict) hierarchical log linear analysis retained only the confidence × verdict association, $\chi^2(1, N = 142) = 10.32$, $p = .001$. Again, consistent with the hypothesized main effect of confidence, follow-up cross-tabulation analysis revealed a higher rate of guilty verdicts in the high confidence (56.9% [45.5, 68.4]) than low confidence (30% [19.3, 40.7]) condition, $\chi^2(1, N = 142) = 10.48$, $p = .001$, *OR* = 3.09 [1.54, 6.17]. Consistent with Experiment 1, including the Confidence × Discrimination × Verdict association did not improve the fit of the model, $\chi^2(2, N = 142) = 2.59$, $p = .274$.

These results rule out jurors' misinterpretations of the ratings as an explanation for the absence of a significant interaction between confidence and discrimination in Experiment 1. Nonetheless, in the interest of minimizing ambiguity for participants, materials for the remaining experiments referred to ratings of match (cf. confidence).

Experiment 2

In Experiments 1 and 1A, identification ratings given to the suspect influenced mock-jurors' evaluations, but this effect was not moderated by ratings given to other lineup members (i.e., evidence relating to extent to which the witness could discriminate between the suspect and fillers). Two explanations come to mind. First, mock-jurors in the previous experiments understood the information provided by the additional ratings, but ignored it in favor of the simplicity afforded by interpreting only the magnitude of the rating given to the suspect. Second, mock-jurors are not particularly sensitive to factors that moderate the reliability of identification evidence (Cutler et al., 1988), and the ratings given to non-suspect lineup members (i.e., fillers) may represent a type of information that is particularly difficult for mock-jurors to interpret. This is a novel form of identification evidence, and mock-jurors may lack the cognitive framework required to interpret ratings given to non-suspect lineup members when evaluating ratings-based identification evidence. For example, if the witness gives the fillers ratings ranging between 10 and 20, is this evidence of a reliable witness (because these ratings are relatively low) or an unreliable witness (because there is at least some evidence the witness found known-innocent lineup members to be plausible candidates)? Cognitive psychologists may readily dismiss this second position as placing unreasonable expectations on human memory performance. However, the belief that recognition is an all or nothing process – the witness both recognizes and identifies the suspect, or rejects the lineup – is not uncommon in legal settings (e.g., Weber & Perfect, 2013). An all-or-nothing conceptualization of recognition leaves little room for the ambiguities inherent in ratings-based indices of identification.

In Experiment 2, we tested these explanations by manipulating discrimination information within-subjects and across defendants. We modified the previously used transcript to include three culprits fleeing the scene in different directions, each observed by a different eyewitness. Each eyewitness completed an identification procedure for the culprit they saw. One witness provided a categorical identification, the other two witnesses provided ratings that demonstrated good and poor discrimination, respectively. Thus, rather than expecting participants to grasp the meaning of a pattern of ratings in isolation, we tested whether participants could recognize the value added by this discrimination information in a relative context, and apply that information to their evaluation. Clearly, jurors in a trial are unlikely to be presented with identification evidence resembling that used here. However, manipulating this critical factor within-subjects permits a test of the two explanations above. Thus, Experiment 2 investigated whether, when presented with contrasting examples of witnesses who demonstrated good and poor levels of discrimination, mock-jurors could (and would) make use of this information to evaluate the reliability of identification evidence. In turn, this would tell us if the absence of any discrimination effects in Experiments 1 and 1A reflected participants' inability or unwillingness to use this information. Based on Tenney et al.'s (2007) work, we hypothesized that the effect of confidence jurors' verdicts would vary according to the level of discrimination evidenced. Specifically, we expected evidence of good discrimination (cf. no discrimination) to increase guilty verdicts for the moderate confidence condition, and evidence of poor discrimination (cf. no discrimination) to decrease guilty verdicts in the high confidence condition. However, given the absence of this effect in our previous experiments, we advanced this hypothesis rather tentatively.

Method

Design and participants. Seventy-one undergraduate students and community members (42 female, one missing data; mean age = 30 years, range = 18-83 years) were randomly allocated to one of the two suspect confidence (match) conditions (high or moderate).

Materials and procedure. The transcript used in Experiment 2 was modified so that the crime involved three culprits who fled the crime scene in different directions, each passing by a different witness. Each witness completed an identification test, conducted by different police officers using different procedures. Each witness's response pointed to the guilt of the suspect they saw. One witness made a standard categorical identification decision, identifying the suspect from a lineup and providing a rating of how well the suspect matched their memory of the perpetrator (e.g., 90% or 50% for the high and moderate match conditions, respectively). The other two witnesses provided match ratings and gave a higher rating to the suspect (90% or 50%) than other lineup members. In the good discrimination condition, the witness gave ratings of 0-10% to the other lineup members. In the poor discrimination condition, the witness gave relatively high ratings to other lineup members (50-70% in the high match condition and 20-40% in the moderate match condition). The order in which the three pieces of identification evidence were presented, and the suspect that each piece of evidence referred to, was counterbalanced. As per Experiments 1 and 1A, suspect match was manipulated between-subjects. Thus, in each version of the transcript, the identification ratings for the three suspects were the same (all 90% or all 50%).

Results and Discussion

Given the difficulties associated with analyzing categorical data from factorial, repeated-measures designs, we ran planned comparisons to test the effect of discrimination information on mock-jurors' verdicts. We split the data file according to confidence, and compared the proportion of guilty verdicts in the three discrimination conditions using Related-Samples Friedman's Two-Way ANOVA by Ranks. Both tests indicated an effect of discrimination information: *F* (2, *N* = 37) = 10.50, *p* = .005, and *F* (2, *N* = 34) = 25.90, *p* < .001, for the moderate and high confidence conditions, respectively.

Follow-up Wilcoxon Signed Rank Tests compared the standard identification condition to the good and poor discrimination conditions, for both high and moderate levels of confidence (Figure 1, Panel A provides descriptive statistics). Results were consistent with the hypothesized interaction between confidence and discrimination. When ratings for the suspect were moderate, guilty verdicts were more likely in the good discrimination condition than in the standard identification condition, *Z* (*N* = 37) = 2.53, *p* = .011, *OR* = 2.42 [0.95, 6.19]. However, there was no significant difference between the good discrimination and standard identification conditions when ratings for the suspect were high, *Z* (*N* = 34) = 0.816, *p* = .414, *OR* = 1.5 [0.43, 5.31]. Further, when ratings for the suspect were high, evidence of poor discrimination led to fewer guilty verdicts compared to the standard identification condition, *Z* (*N* = 34) = 3.87, *p* < .001, *OR* = 0.14 [0.05, 0.42]. There was no significant difference between the poor discrimination and standard identification conditions when ratings for the suspect were moderate, *Z* (*N* = 37) = 0.71, *p* = .480, *OR* = 0.78 [0.29, 2.07].

Consistent with Tenney et al.'s (2007) findings, evidence of the witness's ability to discriminate affected mock-jurors' evaluations. A high rating for the suspect was persuasive of guilt unless the witness showed poor capacity to discriminate between lineup members, and a moderate rating for the suspect was more persuasive when it was accompanied by evidence that the witness could discriminate between lineup members. Thus, mock-jurors' evaluations of identification evidence were shaped by the ratings given to other lineup members. These results indicate that, at least under some circumstances, mock-jurors were able to make sensible use of the identification ratings given to other lineup members when evaluating ratings-based identification evidence against the suspect. Specifically, these results suggest that mock-jurors valued information about not only the extent to which the suspect matched the witness's memory of the culprit, but also the similarity of the suspect to the witness's memory, relative to other lineup members (Sauer & Brewer, 2015).

As an aside, it is interesting to note that in the moderate confidence condition, mock-jurors tended to favor 'not guilty' verdicts in the standard identification and poor discrimination conditions but 'guilty' verdicts in the good discrimination condition. Thus, even when suspect ratings were relatively low (i.e., 50%), jurors still tended to favor 'guilty' (cf. 'not guilty') verdicts provided the witness demonstrated the ability to discriminate between lineup members. Sauer and colleagues (Sauer & Brewer, 2015; Sauer et al., 2008, 2012a; Sauer et al., 2012b) have previously argued that one potential benefit of a ratings-based procedure is that ratings may indicate that a witness recognizes the suspect even in cases where a categorical response might have resulted in a lineup rejection. For example, a witness who is particularly concerned about the prospect of identifying an innocent person may reject a lineup even if they are reasonably sure that they recognize someone in the lineup. In a rating procedure, such a witness would likely give a moderate rating to the person they think they recognize and lower ratings to other lineup members (who seem less familiar). Previous research has shown that when this pattern of confidence ratings occurs, the lineup member who receives the highest rating is often the culprit (Brewer et al., 2012; Sauer et al., 2012a). The results from Experiment 2 suggest that, in a relative comparison between different sets of ratings, jurors have some intuitive appreciation of this. This is consistent with the idea that lay-people may lack the cognitive framework required, rather than willingness, to interpret ratings-based evidence when only exposed to a single lineup.

Experiment 3

Experiments 1 and 1A suggested that when evaluating identification evidence from a single witness, overall, ratings given to non-suspect lineup members had little effect on mock-jurors' verdict preferences. However, Experiment 2 demonstrated that when discrimination information was manipulated within-subjects, mock-jurors considered ratings given to other lineup members when evaluating identification evidence. Thus, the results from Experiments 1 and 1A seem to indicate an inability to interpret, rather than unwillingness to consider, ratings for lineup fillers when evaluating identification evidence against a suspect. In Experiment 3, we investigated whether providing mock-jurors with information about how to interpret ratings-based identification evidence would help them make use of filler ratings when evaluating the likely guilt of a suspect based on evidence provided by a single witness. Again, we expected a main effect of confidence, and an interaction between confidence and discrimination, such that the effect of confidence (high vs. moderate) on verdict would be reduced in the poor compared to good discrimination condition.

Method

Participants and design. Sixty-nine undergraduate student participants (55 female; mean age = 27, range = 18-62) were randomly allocated to one of the 6 cells created by the 2 (confidence: 90%, 50%) × 3 (discrimination: good, poor, none) between-subjects design.

Materials and procedure. The materials and procedure used were identical to those used in Experiment 1A, except that the evidence read by mock-jurors in the discrimination conditions included an information sheet providing (a) a conceptual understanding of what patterns of confidence ratings might indicate, and (b) a schematic illustration of what strong and weak evidence of suspect guilt might look like (see Figure 2). These schematics deliberately depicted extreme patterns of confidence ratings. The intention was to test whether participants could apply this information to their evaluations of identification ratings, rather than to test their ability to map actual ratings on to provided 'templates' for strong and weak evidence. Participants were instructed to consider three things when evaluating ratings-based evidence. First, did the suspect get the highest rating? Second, how high was this rating? Third, how high were the ratings given to the fillers? The logic of this process is that, if the suspect was given the highest confidence rating, we can assume they represented the best match to the witness's memory. If this is true, then (a) higher suspect ratings indicate a greater degree of match to the witness's memory and (b) lower filler ratings indicate a greater ability to discriminate between the suspect and the fillers.

Participants in the no-discrimination condition were given an equivalent information sheet outlining the basic principles and procedures for a standard simultaneous lineup. The full information sheets can be found in the supplementary materials.

Results and Discussion

The 2 (confidence) × 3 (discrimination) × 2 (verdict) hierarchical log linear analysis retained only the discrimination × verdict association, $\chi^2(1, N = 69) = 15.89$, $p < .001$ (see Figure 1, Panel B for descriptive statistics). Follow-up cross-tabulation analysis revealed that evidence of poor discrimination reduced the rate of guilty verdicts compared to the good discrimination, $\chi^2(1, N = 44) = 13.20$, $p < .001$, *OR* = 0.08 [0.02, 0.35], and standard identification conditions, $\chi^2(1, N = 49) = 10.82$, $p = .001$, *OR* = 0.13 [0.04, 0.46]. The difference between the good discrimination and standard identification conditions was non-significant, $\chi^2(1, N = 45) = 0.39$, $p = .535$, *OR* = 1.56 [0.38, 6.31]. Including the Confidence × Discrimination × Verdict association did not improve the fit of the model, $\chi^2(2, N = 69) = 0.15$, $p = .928$.

In sum, the results from Experiment 3 suggest that, following instruction, mock-jurors did consider information relating to a witness's ability to discriminate, and applied this informative adaptively when evaluating identification ratings. However, we did not obtain the hypothesized main effect of confidence, or the interaction. This may reflect the emphasis placed by the instructions on considering information relating to discrimination.

General Discussion

Compared to categorical identification decisions, ratings-based identification evidence provides a promising alternative method of assessing suspect guilt, and a richer source of information for triers-of-fact assessing the reliability of the identification evidence and the likely guilt of a defendant (Brewer & Wells, 2011; Sauer & Brewer, 2015; Sauer et al., 2008, 2012a). Although more empirical work is needed to establish the effectiveness of this technique in police investigations, we believe that a richer type of identification evidence has a variety of potential benefits for the investigative process, prosecutors' decisions to prosecute suspects, and judges' summaries of presented identification evidence. However, jurors' ability to process this information effectively is central to the applied utility of ratings-based identification evidence if presented in court. Thus, we sought to answer two, broad questions. First, are mock-jurors resistant to non-categorical identification evidence? Second, can mock-jurors interpret and apply ratings-based evidence in an adaptive way?

Two key findings emerged. First, across experiments, when confidence for the suspect was high, there was no evidence that the absence of a categorical identification decision undermined the persuasiveness of the evidence against the defendant. Thus, mock-jurors did not routinely dismiss non-categorical identification evidence. Second, although mock-jurors did not intuitively grasp the value of the additional information provided in the ratings-based evidence conditions when considering evidence provided by a single witness (Experiments 1 and 1A), information relating to the witness's ability to discriminate did affect mock-jurors' evaluations when they were able to compare discrimination across witnesses (Experiment 2). Further, when provided with instructions on interpreting patterns of ratings, mock-jurors were able to apply this information in an adaptive way when evaluating evidence provided by a single witness (Experiment 3).

Our results are partially consistent with Tenney et al.'s (2007) findings. It appears that mock-jurors did not initially appreciate the additional information provided. However, the interaction in Experiment 2 demonstrates that evidence of good discrimination enhanced the persuasiveness of moderate confidence and that evidence of poor discrimination reduced the persuasiveness of high confidence. Together with the effects of discrimination information in Experiment 3, this demonstrates that when mock-jurors were assisted with interpreting this information (either implicitly through relative comparison or through explicit instruction) evidence of a witness's ability to discriminate affected evaluations of the witness's assessment of the suspect. Together with Tenney et al.'s findings, our findings may shed light on the conditions under which mock-jurors are likely to make use of information that adds nuance to evaluations of identification evidence. First, mock-jurors may be more likely to consider such information if that information pertains to the witness's ability to discriminate between memory outputs that are more or less likely to be correct in the present context (cf. information relating to factors that generally increase or decrease the reliability of memory-based evidence). Second, and related to the above point, mock-jurors may lack the cognitive framework required to readily intuit the implications of such information for the reliability of the witness's memory in a given context. It may be necessary to provide instruction to help mock-jurors interpret and apply pertinent information (a point implied by existing judicial instructions provided to help jurors assess eyewitness identification evidence; see *State v. Henderson*, 2011; *United States v. Telfaire*, 1972) .

The immediate applied significance of this research is clearly constrained by several factors. First, we used simple materials that do not faithfully reproduce the complex and noisy decision-making environment that confronts jurors in genuine cases. Given the absence of trial features like cross-examination and detailed judicial instructions, it is not clear that the obtained effects would generalize to real trial settings. However, this research was intended to be a first pass at investigating how potential jurors would respond to non-categorical identification evidence.

Second, the instructions we provided to help mock-jurors evaluate the ratings-based evidence were simplistic. If an approach similar to the ratings-based procedure were adopted in the criminal justice system, careful consideration would need to be given to the instructions provided to jurors (e.g., to what extent must instructions be modified to accommodate the possibility of procedural bias?), and the manner in which they were provided (e.g., are they best provided by a judge or expert witness?). Determining how instructions could be conveyed in an effective and legally permissible manner would require a collaborative effort involving legal professionals and psycho-legal scholars. We certainly do not claim that the protocols used here are ready for applied settings. Rather, we aimed to investigate whether, with some guidance, mock-jurors were able to interpret varied patterns of ratings-based identification evidence.

Third, it could be argued that the additional information provided by the ratings-based evidence did not assist mock-jurors in evaluating the reliability of the witness's evidence; rather, mock-jurors simply responded to instructions telling them how to evaluate the ratings provided. Thus, one interpretation of our findings might be that Experiment 3 simply demonstrates that mock-jurors were doing what they were told to do, rather than evaluating the identification evidence provided more effectively. Two aspects of the current research speak against this interpretation. First, the interaction in Experiment 2 indicates that mock-jurors valued evidence of discrimination even in the absence of explicit instruction. Second, although the instructions provided in Experiment 3 were explicit relating to the intended meaning of ratings in general (providing schematic examples of how clear-cut patterns of ratings should be interpreted), mock-jurors still needed to apply this information sensibly to interpret the specific evidence provided. It is not the case that mock-jurors could simply match the ratings provided to a schematic example in order to evaluate the evidence. Further, even if jurors were reliant on instructions to interpret ratings-based identification evidence, this is not inherently problematic. As discussed previously, current practice acknowledges that jurors may benefit from instruction when evaluating the reliability of identification evidence, and our data are consistent with the notion that lay-people may not possess the cognitive frameworks required to intuitively process ratings-based evidence.

Finally, it might be argued that because this procedure is not being used in criminal justice settings this research is of minimal applied value. However, assessing potential jurors' responses to this non-traditional, non-categorical form of identification evidence is an important step in determining its appropriateness for applied settings. There is accumulating evidence that ratings-based approaches provide a more informative source of identification evidence for mock-jurors to evaluate, and a more valid assessment of the witness's memory for the suspect/defendant. This approach may encourage the criminal justice system to abandon the problematic conceptualization of identification evidence as an absolute indication of guilt, in favor of a more scientific consideration of identification evidence as yet another piece of probabilistic evidence (Sauer & Brewer, 2015).

In sum, although our results suggest that mock-jurors did not readily intuit the value of the discrimination information provided for individual witnesses, they also demonstrate that mock-jurors did not immediately dismiss non-categorical identification evidence as uninformative. More importantly, the results show that following instruction mock-jurors applied this additional information sensibly. Given that mock-jurors often experience problems assessing the reliability of identification evidence (Cutler et al., 1988), we take participants' responses to the relatively minimalistic instructions used in these experiments (cf. *State v. Henderson*, 2011) as a sign of mock-jurors' willingness and ability to consider the information provided, and the potential applied utility of ratings-based identification evidence.

References

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

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgements. *Law and Human Behavior, 26*, 353-364.

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*(10), 1208-1214. doi: 10.1177/0956797612441217

Brewer, N., & Wells, G. L. (2009). Obtaining and interpreting eyewitness identification test evidence: The influence of police-witness interactions. In T. Williamson, R. Bull & T. Valentine (Eds.), *Handbook of psychology of invetigative interviewing: Current developments and future directions* (pp. 205-220). Chichester: Wiley-Blackwell.

Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science, 20*, 24-27.

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior, 14*(2), 185-191. doi: 10.1007/bf01062972

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Jury decision making in eyewitness identification cases. *Law and Human Behavior, 12*, 41-56.

Goodman, J. (1992). Jurors' comprehension and assessment of probabilistic evidence. *American Journal of Trial Advocacy, 16*, 361-389.

Heuer, L., & Penrod, S. (1994). Trial complexity: A field investigation of its meaning and its effects. *Law and Human Behavior, 18*(1), 29-51. doi: 10.1007/bf01499142

Horowitz, I. A., ForsterLee, L., & Brolly, I. (1996). Effects of trial complexity on decision making. *Journal of Applied Psychology, 81*(6), 757-768. doi: 10.1037/0021-9010.81.6.757

Lindsay, R. C. L., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*, 79-89.

Reardon, M. C., & Fisher, R. P. (2011). Effect of viewing the interview and identification process on juror perceptions of eyewitness accuracy. *Applied Cognitive Psychology, 25*(1), 68-77. doi: 10.1002/acp.1643

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 185-208). Chichester: Wiley Blackwell.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137*, 528-547.

Sauer, J. D., Brewer, N., & Weber, N. (2012a). Using confidence ratings to identify a target among foils. *Journal of Applied Research in Memory and Cognition, 1*(2), 80-88. doi: 10.1016/j.jarmac.2012.03.003

Sauer, J. D., Weber, N., & Brewer, N. (2012b). Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness. *Psychonomic Bulletin & Review, 19*(3), 490-498. doi: 10.3758/s13423-012-0239-5

Semmler, C., Brewer, N., & Douglass, A. B. (2011). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), *Conviction of the innocent: Lessons from psychological research* (pp. 185 - 209). Washington, D.C.: APA Books.

Simons, D. J., & Chabris, C. F. (2011). What People Believe about How Memory Works: A Representative Survey of the U.S. Population. *PloS One, 6*(8), e22757. doi: 10.1371/journal.pone.0022757

State v. Henderson, WL 3715028 (2011).

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology Public Policy and Law, 17*(1), 99-139. doi: 10.1037/a0021650

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.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior, 11*(3), 167-187.

Weber, N., & Perfect, T. J. (2013). Why telling a witness that it’s OK to say they don’t know is good for justice. *The Jury Expert: The Art and Science of Litigation and Advocacy, 25*(3), 1-7.

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*, 553-571.

Wise, R. A., & Safer, M. A. (2004). What US judges know and believe about eyewitness testimony. *Applied Cognitive Psychology, 18*(4), 427-443. doi: 10.1002/acp.993

Wise, R. A., & Safer, M. A. (2010). A Comparison of What US Judges and Students Know and Believe About Eyewitness Testimony. *Journal of Applied Social Psychology, 40*(6), 1400-1422.

Table 1

Ratings[a] given to the suspect and fillers in each condition for Experiments 1 to 3.

	Confidence	
Filler Ratings	High (90%)	Moderate (50%)
NA (Standard ID condition)	-	-
Good Discrimination	0-10%	0-10%
Poor Discrimination	50-70%	20-40%

[a] In Experiment 1, these were confidence ratings. In all other experiments, these were ratings of "match". In the Standard ID condition, mock-jurors were told that the witness identified the suspect with either High or Moderate confidence (Experiment 1) or that the witness identified the suspect and indicated the degree to which the suspect matched their memory of the culprit as either High or Moderate (Experiments 1A-3).

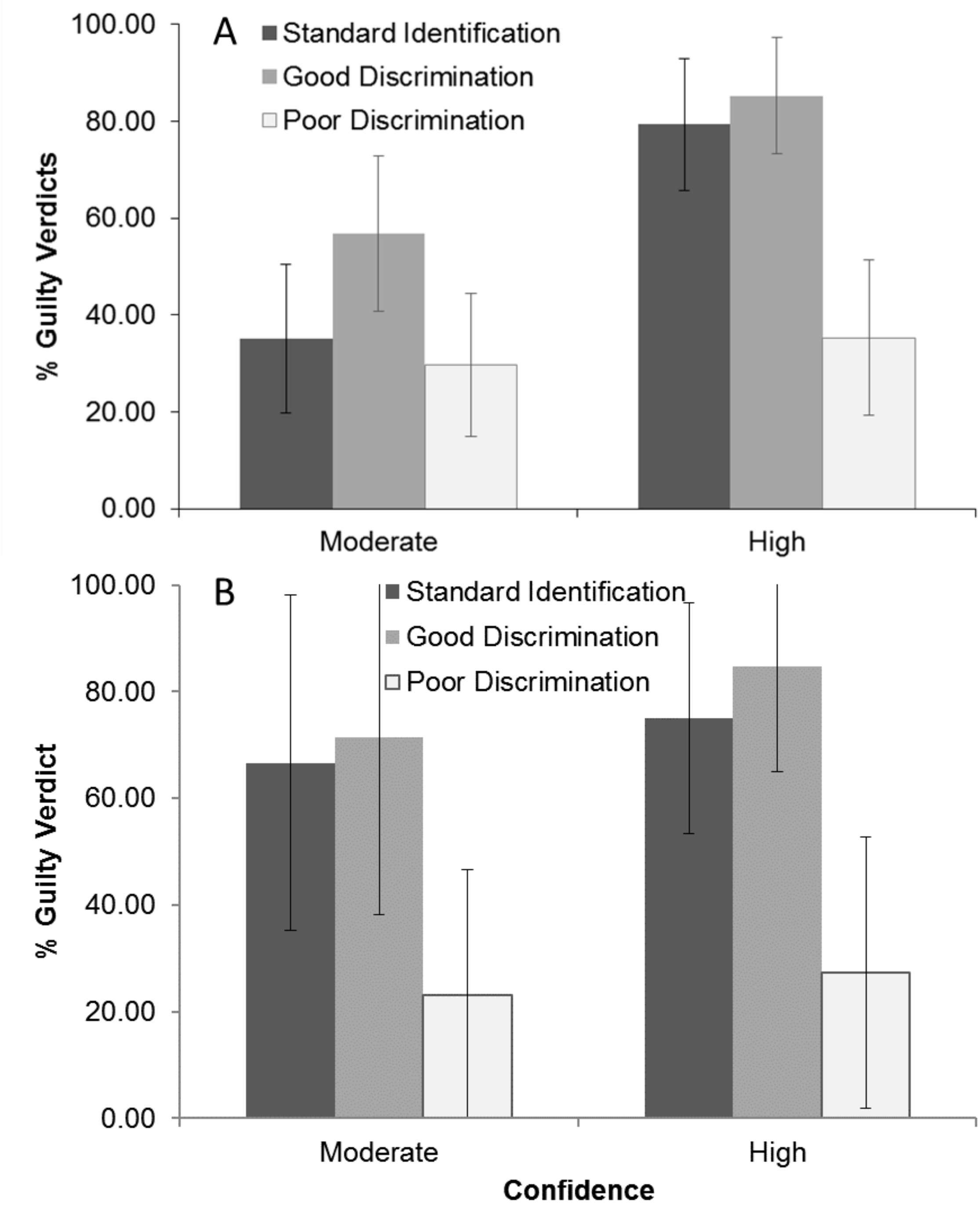

Figure 1. The percentage of guilty verdicts in Experiment 2 (Panel A) and Experiment 3 (Panel B) according to the rating given to the suspect (or the confidence expressed in the accuracy of the suspect identification) and the level of discrimination evidenced. Error bars indicate 95% CIs. **Note**: the data in Panel A were obtained from a within-subjects design. Thus, the 95% CIs are not informative as an indicator of statistical significance.

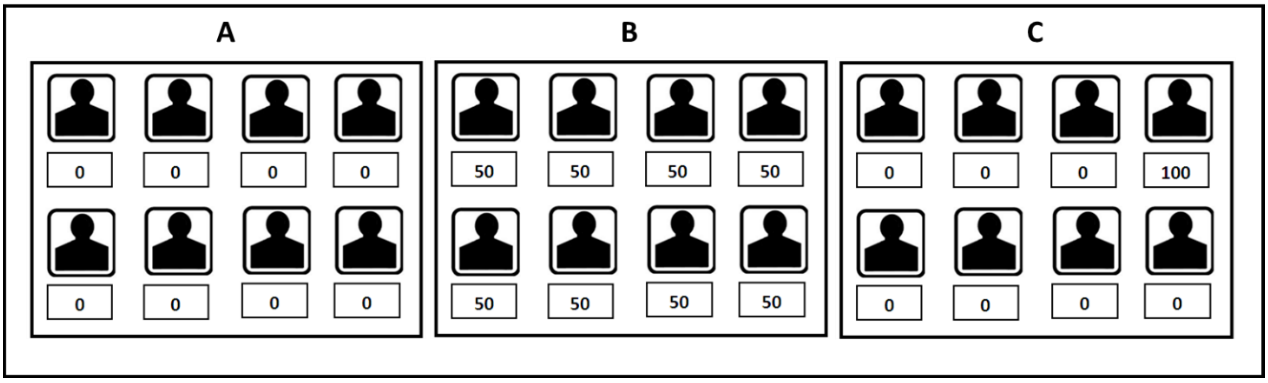

Figure 2. Schematic representations of weak (panels A and B) and strong (panel C) evidence against the suspect in the ratings-based identification conditions in Experiment 3. Note the suspect is positioned in the upper right corner of the lineup.