**Supplementary Materials – Analyses of Verdict Preference**

Following Tenney et al. ([2007](#_ENREF_4)), the verdict and verdict confidence scores were combined to create a verdict preference variable. We might expect mock-jurors to set a strict criterion for a guilty verdict (e.g., reflecting the commonly-encountered maxim that, to be found guilty, a defendant must be guilty *beyond a reasonable doubt*), thus a binary verdict measure may be relatively insensitive to factors that affect mock-jurors’ perceptions of identification evidence. A verdict preference score provides a potentially more sensitive measure of the effects of our manipulations on mock-juror evaluations. A value of 0.5 was added to each verdict confidence score so that guilty and not guilty verdicts made with 0% confidence would be separated by 1 unit on the verdict preference scale. Thus, possible scores ranged from -100.5 (indicating a not guilty verdict made with 100% confidence) to 100.5 (indicating a guilty verdict made with 100% confidence). For effect sizes, we report partial Cohen’s *f* for omnibus analyses and Cohen’s *d* for differences between two means, with 95% confidence intervals reported in brackets. Confidence intervals for *f* were calculated using the MBESS package in R ([Kelly & Lai, 2012](#_ENREF_2); [R Core Team, 2013](#_ENREF_3))[[1]](#footnote-1). Bootstrapped confidence intervals for *d* were calculated using bootES ([Gerlanc & Kirby, 2013](#_ENREF_1)) based on 2000 replications.

**Experiment 1**

A 2 (confidence) × 3 (discrimination) between-subjects ANOVA on verdict preference scores yielded a significant main effect of confidence, *F*(1, 127)= 13.26, *p* < .001, *f* = 0.32, 95% CI [0.14, 0.50]. The main effect of discrimination, *F*(2, 127)= 1.89, *p* = .155, *f* = 0.17 [0.00, 0.33], and the interaction, *F*(2, 127)= 0.85, *p* = .428, *f* = 0.12 [0.00, 0.26], were non-significant. Thus, consistent with prior research, mock-jurors reported a stronger preference for guilty verdicts when the suspect was implicated with high versus low confidence. However, contrary to expectations, there was no indication that the level of discrimination evidenced affected mock-jurors’ evaluations (see Figure S1).

**Experiment 1A**

Experiment 1A produced results very similar to those of Experiment 1 (see Figure S1). A 2 (confidence) × 3 (discrimination) between-subjects ANOVA on verdict preference scores yielded a significant main effect of confidence, *F*(1, 136)= 10.55, *p* = .001, *f* = 0.28 [0.11, 0.45], while the main effect of discrimination and the interaction were non-significant, *F*(2, 136)= 0.26, *p* =.775, *f* = 0.06 [0.00, 0.19], and *F*(2, 136)= 0.63, *p* =.536, *f* = 0.10 [0.00, 0.24], respectively[[2]](#footnote-2).

**Experiment 2**

 A 2 (confidence) × 3 (discrimination) mixed ANOVA on verdict preference scores yielded significant main effects for confidence, *F*(1, 69)= 14.13, *p* < .001, *f* = 0.35 [0.21, 0.49], and discrimination, *F*(2, 68)= 21.74, *p* < .001, *f* = 0.35 [0.20, 0.48]. These are best interpreted in the context of the modest but significant Confidence × Discrimination interaction, *F*(2, 68)= 3.62, *p* = .029, *f* = 0.14 [0.00, 0.27]. We assessed the effect of discrimination on verdict preference for each of the confidence conditions via simple effects analyses (with a Bonferroni correction for multiple comparisons; alpha = .017). Results are shown in Figure S2.

The simple effects analyses revealed two important findings. First, when confidence for the suspect was moderate (50%), evidence of good discrimination improved the persuasiveness of the identification evidence compared to not only the poor discrimination condition, *t*(36)= 3.24, *p* = .003, *d* = 0.61 [0.16, 1.14], but also the standard identification condition, *t*(36)= 3.09, *p* =.004, *d* = 0.46 [-0.01, 0.94]. Verdict preferences did not differ significantly between the poor discrimination and standard identification conditions, *t*(36)= 0.84, *p* = .406, *d* = 0.15 [-0.33, 0.60]. Second, when confidence for the suspect was high (90%), evidence of poor discrimination undermined the persuasiveness of the evidence compared to good discrimination, *t*(33)= 4.97, *p* < .001, *d* = 1.12 [0.54, 1.75], and standard identification conditions, *t*(33)= 4.53, *p* < .001, *d* = 0.80 [0.26, 1.40]. Verdict preferences did not differ significantly between the good discrimination and standard identification conditions, *t*(33)= 1.38, *p* = 1.78, *d* = 0.26 [-0.25, 0.73], though the small effect indicated that evidence of discrimination may improve the persuasiveness relative to the standard identification condition.

**Experiment 3**

 A 2 (confidence) × 3 (discrimination) between-subjects ANOVA revealed a large main effect of discrimination, *F*(2, 63) = 11.29, *p* < .001, *f* = 0.60 [0.30, 0.83] (see Figure S3). Follow-up simple effects analyses demonstrated that evidence of poor discrimination undermined the persuasiveness of the identification evidence relative to both the good discrimination, *t*(42) = 4.72, *p* < .001, *d* = 1.43 [0.66, 2.37], and standard identification conditions, *t*(47) = 4.10, *p* < .001, *d* = 1.17 [0.47, 1.87]. The difference between the good discrimination and standard identification conditions was small and non-significant, *t*(43) = 0.89, *p* = .381, *d* = 0.27 [-0.36, 0.89]. Although evidence of good discrimination did not significantly increase jurors’ evaluations of the identification evidence relative to the standard identification condition, evidence of poor discrimination reduced the persuasiveness of the evidence compared to both the standard identification and good discrimination conditions. Thus, there is evidence that following instruction, jurors were willing and able to incorporate filler ratings into their assessment of the identification evidence provided by a single witness. The non-significant difference between the good discrimination and standard identification conditions is consistent with the possibility that, if jurors are not presented with explicit evidence regarding the witness’s ability to discriminate, they assume the witness’s ability to discriminate is good.

In contrast to the previous experiments, the main effect of confidence on mock-jurors’ verdict preferences was non-significant, *F*(1, 63) = 1.49, *p* = .226, *f* = .15 [0.00, 0.40. The non-significant interaction, *F*(2, 63) = 0.05, *p* = .954, *f* = .04 [0.00, 0.14], provided no evidence of difference in the effect of discrimination across confidence conditions. Together, these findings may reflect the emphasis (in the information sheet given to participants) placed on considering discrimination information. Mock-jurors may have been more concerned with assessing the extent to which the witness could discriminate between lineup members, than the absolute confidence rating given to the suspect.

**Summary**

 Across four experiments, consistent with analyses of the categorical verdict measures, analysis of verdict preference date provide no evidence that mock-jurors rejected non-categorical identification evidence as uninformative. However, jurors may benefit from some assistance in interpreting the additional information provided by ratings-based evidence.

**References**

Gerlanc, D., & Kirby, K. (2013). bootES: Bootstrap Effect Sizes. Retrieved from [http://CRAN.R-project.org/package=bootES](http://CRAN.R-project.org/package%3DbootES)

Kelly, K., & Lai, K. (2012). MBESS: MBESS. R package version 3.3.3. Retrieved from [http://CRAN.R-project.org/package=MBESS](http://CRAN.R-project.org/package%3DMBESS)

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.



*Figure S1.* Verdict preference ratings according to the confidence (Experiment 1) or match (Experiment 1A) rating assigned to the suspect (or the confidence expressed in the accuracy of the suspect identification), and the level of discrimination evidenced in Experiment 1 (panel A) and Experiment 1A (panel B). Error bars indicate 95% CIs.



*Figure S2.* Verdict preference ratings according to the level of match for the suspect and the level of discrimination evidenced in Experiment 2. Error bars indicate 95% CIs.



*Figure S3.* Verdict preference ratings according to the level of match for the suspect and the level of discrimination evidenced in Experiment 3. Error bars indicate 95% CIs.

1. To calculate confidence intervals for Cohen’s *f*, we first calculated confidence for η2 and then converted these into Cohen’s *f* values using the formula *f* = SQRT(η2 / (1 - η2)). Because η2 values are bounded at zero, our *f* values are also bounded at zero. [↑](#footnote-ref-1)
2. The pattern of results for Experiments 1 and 1A might appear inconsistent, particularly with reference to possible effects of discrimination on the verdict preference measure. We ran Bayesian analyses on our data to determine the strength of evidence for the null and alternative hypotheses. For Experiment 1, a Bayesian ANOVA testing the effects of discrimination on verdict ratings produced a Bayes factor of 3.56 (indicating moderate support) in favor of the null hypotheses. For Experiment 1A, the Bayes factor in favor of the null was 12.56, indicating strong support for the null. Based on these data, there is little or no reason to believe mock-juror responses differed meaningfully according to the discrimination manipulation in either experiment, or that the results from Experiments 1 and 1A are inconsistent [↑](#footnote-ref-2)