Filler Siphoning Theory Does Not Predict the Effect of Lineup Fairness on the Ability to Discriminate Innocent from Guilty Suspects: Reply to Smith, Wells, Smalarz, and Lampinen (2017)

Melissa F. Colloff[1], Kimberley A. Wade[2], Deryn Strange[3], and John T. Wixted[4]

Author Note

[1] Centre for Applied Psychology, School of Psychology, University of Birmingham, Birmingham, B15 2TT, UK.

[2] Department of Psychology, University of Warwick, Coventry, CV4 7AL, UK.

[3] Department of Psychology John Jay College of Criminal Justice, City University of New York, New York, USA

[4] Department of Psychology, University of California, San Diego, La Jolla, California 92093, USA.

We thank Heather Flowe and Matthew Palmer for insightful discussions.

Correspondence concerning this article should be addressed to Melissa F. Colloff, School of Psychology, University of Birmingham, Birmingham, UK, B15 2TT. Email: M.Colloff@bham.ac.uk

Abstract

Smith, Wells, Smalarz, and Lampinen (2017) claim that we (Colloff, Wade, & Strange, 2016) were wrong to conclude that fair lineups enhanced people's ability to discriminate between innocent and guilty suspects compared to unfair lineups. They argue our results reflect differential-filler-siphoning, not diagnostic-feature-detection. But a manipulation that decreases identifications of innocent suspects more than guilty suspects (i.e., that increases filler-siphoning or conservative responding) does not necessarily increase people's ability to discriminate between innocent and guilty suspects. Unlike diagnostic-feature-detection, filler-siphoning does not make a prediction about people's ability to discriminate between innocent and guilty suspects. Moreover, we replicated Colloff et al.'s results in the absence of filler-siphoning (N=2,078). Finally, a model is needed to measure ability to discriminate between innocent and guilty suspects. Smith et al.'s model-based analysis contained several errors. Correcting those errors shows that our model was not faulty, and Smith et al.'s model supports our original conclusions.

Keywords: Signal detection theory, diagnostic-feature-detection, filler siphoning, eyewitness identification, decision-making

In Colloff, Wade, and Strange (2016) we set out to test a prediction made by the diagnostic-feature-detection theory (Wixted & Mickes, 2014). That theory posits that the presence of similar-looking lineup members (i.e., "foils" or "fillers") in fair lineups allows shared facial features which are non-diagnostic of guilt to be noticed and discounted. As a result, the theory predicts that witnesses' ability to discriminate between innocent and guilty suspects (i.e., $d'_{\text{Innocent-Guilty}}$) should be—and, as we found, is—better in fair lineups than unfair lineups.

Smith, Wells, Smalarz, and Lampinen (2017) argue instead that (1) fair lineups do not improve but instead worsen people's memory performance and (2) a different theoretical account better explains our results. With regard to the first point, Smith et al. argue that we reached the wrong conclusion because we fit the wrong signal-detection model to the data. With regard to the second point, Smith et al. proposed filler-siphoning theory, which posits that the presence of similar-looking foils in fair lineups make it less likely that witnesses will pick the suspect. The process is hypothesised to be *differential*, with similar-looking foils attracting more identifications when the suspect in the lineup is innocent than when he is guilty. Thus, filler-siphoning predicts that the false alarm rate to innocent suspects will decrease *more than* the hit rate to guilty suspects as lineups become increasingly fair.

We welcome the opportunity to explain in greater detail why filler-siphoning is not a sufficient account of the Colloff et al. (2016) results and how we modelled our data. In what follows, we (1) explain how the two theories speak to different aspects of memory performance and why diagnostic-feature-detection—but not filler-siphoning—predicts the increase in $d'_{\text{Innocent-Guilty}}$ that we observed; (2) present new data from an experiment that tested the same prediction that was tested in Colloff et al., but this time with no foils involved (eliminating the possibility of filler-siphoning) and (3) illustrate that not only was the signal-detection model we fit to the data appropriate, the model preferred by Smith et al., when fit to the data as it should be, confirms that $d'_{\text{Innocent-Guilty}}$ was higher in the fair lineup condition, as predicted by diagnostic-feature-detection theory.

1. Filler-siphoning does not make a prediction about $d'_{\text{Innocent-Guilty}}$

Signal-detection theory holds that there are two distinct elements to performance—discrimination and response bias. A manipulation that influences response bias does not necessarily influence discrimination, and vice versa (Green & Swets, 1966). The notion of filler-siphoning speaks to how *likely people are to choose the suspect* as lineups become increasingly fair. In that sense, it is analogous to a theory of response bias that speaks to how

likely people are to choose the suspect (and foils) as responding becomes increasingly conservative. In both cases, responses that would have been made to innocent or guilty suspects (i.e., responses that would have ended up in the suspect ID category) end up in a different response category. The only difference is that filler-siphoning theory predicts that responses will end up in the foil ID category as lineups become increasingly fair, whereas responses end up in the "not present" category when responding becomes more conservative (e.g., Mickes, Flowe, & Wixted, 2012). In both cases, the hit rate to guilty suspects and false alarm rate to innocent suspects decrease *differentially*. That is, in both cases, the false alarm rate decreases more than the hit rate (e.g., Rotello & Chen, 2016; Rotello, Heit & Dubé, 2015; Wixted & Mickes, 2018). We agree that our Colloff et al. data are fully consistent with differential filler-siphoning theory in this respect. Indeed, we said so (Colloff et al., 2016, Supplemental Materials, p.6).

Critically, however, a manipulation that decreases innocent suspect identifications more than it decreases guilty suspect identifications (i.e., a manipulation that increases filler-siphoning or induces conservative responding) does not necessarily increase people's ability to discriminate between innocent and guilty suspects (e.g., Mickes et al., 2017). Indeed, because the notion of filler-siphoning speaks to how *likely people are to choose the suspect* (analogous to a theory of response bias), it makes no a priori prediction about how $d'_{\text{Innocent-Guilty}}$ (the ability to *discriminate* between innocent and guilty suspects) will change across conditions. Increased filler-siphoning is compatible with an increase in $d'_{\text{Innocent-Guilty}}$, a decrease in $d'_{\text{Innocent-Guilty}}$, or no change in $d'_{\text{Innocent-Guilty}}$. As such, filler-siphoning theory does not make a prediction about the specific change in $d'_{\text{Innocent-Guilty}}$ that we observed in Colloff et al. Diagnostic-feature-detection theory, however, specifically predicts the $d'_{\text{Innocent-Guilty}}$ effect that we observed.

To illustrate this argument, we need a model to understand the mechanism underlying filler-siphoning and the prediction made by diagnostic-feature-detection theory. Figure 1 illustrates a signal-detection interpretation of an unfair lineup, and three possible ways $d'_{\text{Innocent-Guilty}}$ can change, independently of filler-siphoning, as lineups become fairer. In the very unfair lineup in Figure 1A, approximately 20% of foils fall above the decision criterion (shaded grey), and only these foils compete for IDs with the much higher proportion of innocent and guilty suspects who fall above the criterion. When lineups become fairer, the foils in the lineup become more similar to the guilty suspect (i.e., they better match the description of the perpetrator), so the distance between the foil distribution and guilty suspect distribution becomes smaller. In each plot (i,ii,iii) in Figure 1B, the distance between the foil

and guilty suspect distributions has become smaller by the same amount. All that differs is the distance between the innocent suspect and guilty suspect distributions ($d'_{\text{Innocent-Guilty}}$), which is what diagnostic-feature-detection theory makes a prediction about. In plot (i) of Figure 1B, $d'_{\text{Innocent-Guilty}}$ remains *unchanged* as lineups become fairer (contrary to diagnostic-feature-detection theory); in plot (ii), $d'_{\text{Innocent-Guilty}}$ *decreases* as lineups become fairer (again, contrary to diagnostic-feature-detection theory); and in plot (iii), $d'_{\text{Innocent-Guilty}}$ *increases* as lineups become fairer (consistent with diagnostic-feature-detection theory). Crucially, differential filler-siphoning is observed in all three scenarios involving fairer lineups: In each case, approximately 50% of the foils now exceed the decision criterion (shaded grey), and those additional foils compete for IDs with the innocent and guilty suspects who exceed the criterion. Thus, in fairer lineups, the foil ID rate increases while the innocent and guilty suspect ID rates both decrease. Because the foil distribution overtakes a greater proportion of the innocent suspect distribution than the guilty suspect distribution, in all three scenarios the innocent suspect ID rate decreases more than the guilty suspect ID rate. Hence, differential filler-siphoning occurs no matter what the effect of changing to fairer lineups might be on $d'_{\text{Innocent-Guilty}}$. Simply put, the $d'_{\text{Innocent-Guilty}}$ finding in our Colloff et al. (2016) data is compatible with—but not predicted by—the filler-siphoning hypothesis. Conversely, diagnostic-feature-detection theory specifically predicts, and is therefore able to explain a priori, why $d'_{\text{Innocent-Guilty}}$ was larger in the fair lineup conditions.

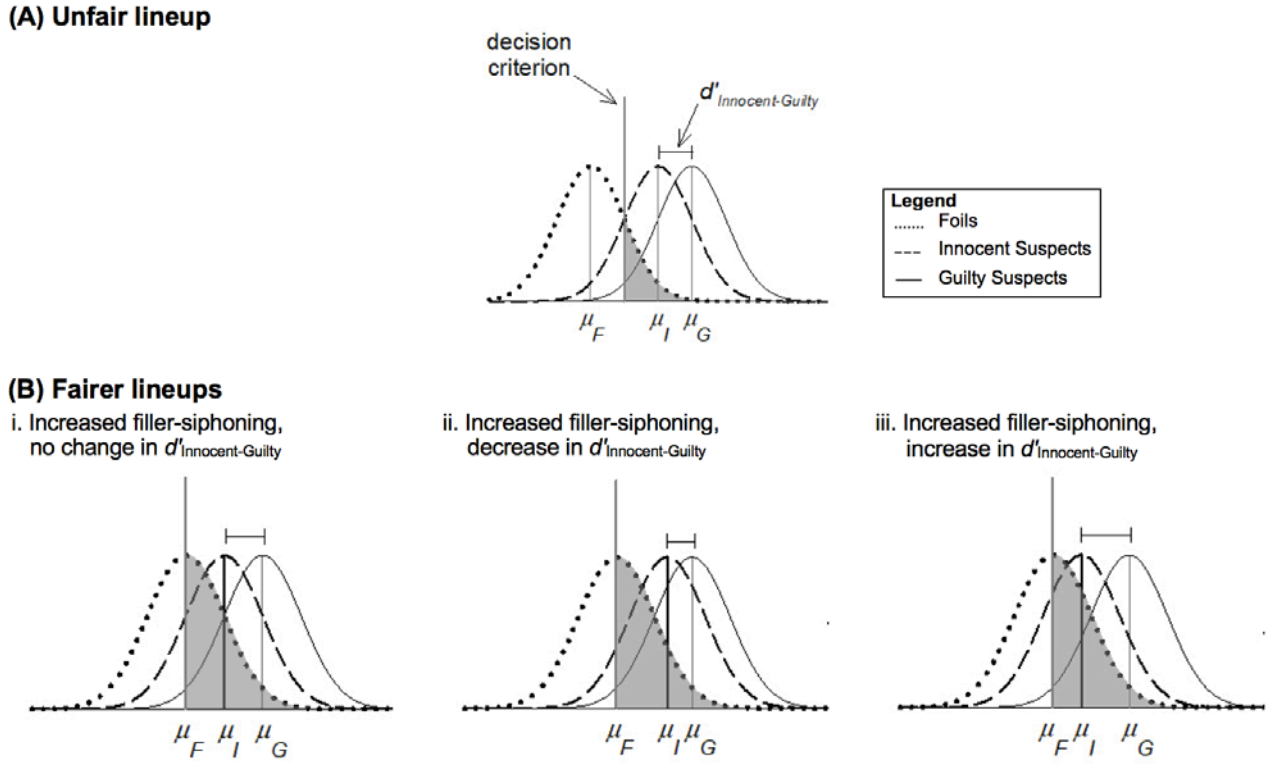


Figure 1. Signal-detection interpretation of (A) an unfair lineup, and (B) three different ways in which $d'_{\text{Innocent-Guilty}}$ can change, independently of filler-siphoning, when a fairer lineup is used. In panel B, we depict lineups that are less unfair (not perfectly fair) compared to panel A to show clearly the predictions made by the filler-siphoning and diagnostic-feature-detection accounts as lineups become increasingly fair. However, the point illustrated in panel B (namely, that differential-filler-siphoning is compatible with any outcome with respect to $d'_{\text{Innocent-Guilty}}$) applies to every degree of increased fairness relative to panel A, including to perfectly fair lineups (as depicted in Figure 2C). Diagnostic-feature-detection theory predicts the outcome illustrated in panel B plot iii (namely, $d'_{\text{Innocent-Guilty}}$ should increase when a fairer lineup is used) for reasons briefly described in this reply and in more detail in Colloff et al. (2016).

2. The predicted effect on $d'_{\text{Innocent-Guilty}}$ occurs even in the absence of fillers

To further test the diagnostic-feature-detection mechanism, and to further underscore its independence from filler-siphoning, we conducted a showup experiment (N=2,078), which removed the possibility of filler-siphoning because there were no foils. Except for the elimination of foils, the "fair" and "unfair" showup conditions were identical to the fair block and unfair do-nothing lineup conditions in Colloff et al. (2016). In the "unfair" showup condition, the innocent and guilty suspects shared a distinctive feature (e.g., a black eye) that was present on the perpetrator at the time of the simulated crime. In the "fair" showup condition, neither suspect had the distinctive feature. Differential filler-siphoning theory makes no prediction about the outcome of this study, but diagnostic-feature-detection theory makes the same prediction as in our original study (i.e., $d_{\text{Innocent-Guilty}}$ should be higher in "fair" showups). Theoretically, a "fair" showup prevents witnesses from relying on a non-diagnostic feature by removing it altogether (enhancing the ability of witnesses to

discriminate between innocent and guilty suspects) just as, in fair lineups, similar foils who share the distinctive feature effectively remove it by causing that feature to be discounted (again, enhancing the ability of witnesses to discriminate between innocent and guilty suspects). We analysed the showup data in the same way that we analysed the lineup data in our original paper and found the same result. ROC analysis showed that people were better able to discriminate between innocent and guilty suspects in "fair" showups that prevented reliance on the distinctive feature (*p*AUC=.102, 95% CI=[.091, .112]) than in "unfair" showups that allowed people to rely on the non-diagnostic distinctive feature (*p*AUC=.075, 95% CI=[.065, .084]), D=3.75, $p < .001$; see Figure 2A. Fitting a model corroborated these findings: $d_{\text{Innocent-Guilty}}$ was significantly larger in "fair" ($d_{\text{Innocent-Guilty}}$=1.13) than "unfair" showups ($d_{\text{Innocent-Guilty}}$=0.92). Note that these analyses are based on participants who identified innocent or guilty suspects in accordance with our pre-registered plans; when full ROC curves are plotted and modelled the conclusions remain the same (see Supplemental Materials). Critically, these findings cannot be explained by filler-siphoning.

3. An empirical comparison of Smith et al.'s model versus our model

Smith et al. also argued that the model we used to estimate $d'_{\text{Innocent-Guilty}}$ was inappropriate because it (1) "misclassified large proportions of false-positive responses as rejections" (p.2) and (2) was a "simple-detection model" which did not "have both detection and identification components" (p.7). Smith et al. fit a different signal-detection model to the data that they argued was more appropriate. Based on the fit of that model, they concluded that discriminability is actually higher for *unfair* lineups (the opposite of the prediction made by diagnostic-feature-detection theory).

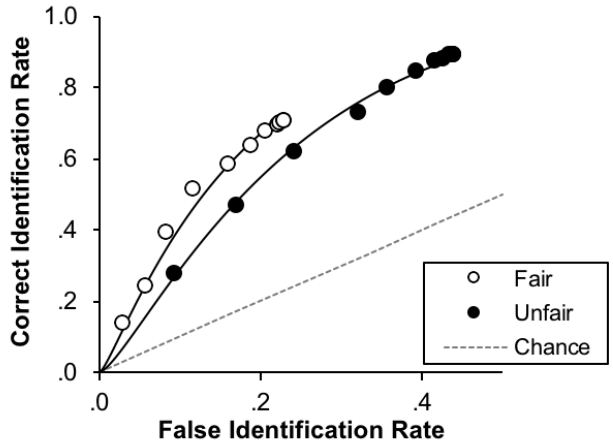
To clarify, our model classified—and, thus, analyzed—false-positives to foils as foil identifications, not as rejections (see Table S2 Colloff et al., 2016). Also, our model is a compound signal-detection model (Duncan, 2006) because it assumes a two-step decision-making process: first, detect the most familiar lineup member, and second, identify that individual if the relevant memory strength variable is strong enough. The only difference between our model and Smith et al.'s is the decision rule: ours uses the independent-observation *best-above-criterion* rule (hereafter, the BEST model; Clark, Erickson, & Breneman, 2011; Macmillan & Creelman, 2005) whereas Smith et al.'s uses the *integration* rule (hereafter, the INTEGRATION model; Palmer, Brewer, & Weber, 2010). When we empirically compared the two models we found that the BEST model offered a noticeably better fit (see Figure 2B). Nonetheless, even Smith et al.'s INTEGRATION model supports our

original conclusion: $d'_{\text{Innocent-Guilty}}$ is higher in fair lineups compared to unfair lineups according to both models (see Figure 2B and Supplemental Materials).

If the BEST model isn't faulty, and the INTEGRATION model supports our original conclusion, why did Smith et al. conclude that fair lineups *impair* discriminability? They came to this conclusion because, when fitting the model to the unfair lineups, they treated foils and innocent suspects as being drawn from the same Gaussian distribution. From a signal-detection perspective, doing so makes sense when the lineup is fair (Figure 2C) but not when the lineup is unfair (Figure 2D). The ability to discriminate between innocent and guilty suspects is represented by $d'_{\text{Innocent-Guilty}}$ in both fair and unfair lineups. In fair lineups, $d'_{\text{Innocent-Guilty}}$ is equal to $d'_{\text{Foil-Guilty}}$ (Figure 2C) because the innocent suspect and the foils are equally similar to the culprit—in Colloff et al. (2016) the innocent suspect and foils had the same distinctive feature as the culprit. But in unfair lineups the innocent suspect looked more like the culprit than did the other foils—only the innocent suspect, not the foils, had the same distinctive feature as the culprit. Thus, from a signal-detection perspective, unfair lineups require two separate d' estimates: $d'_{\text{Innocent-Guilty}}$ and $d'_{\text{Foil-Guilty}}$ (Figure 2D). Even when analyzing the unfair lineup data, Smith et al. combined innocent suspect and foil identifications from target-absent lineups, as if they were drawn from the same memory-strength distribution (reducing a 3-distribution model to a 2-distribution model). Although creating an "omnibus" summary measure of discriminability in unfair lineups seems intuitive, it confounds our measure of interest ($d'_{\text{Innocent-Guilty}}$) with the experimental manipulation ($d'_{\text{Foil-Guilty}}$; see Supplemental Materials).

To summarize, we agree filler-siphoning occurs to a greater extent in fair than unfair lineups, reducing innocent suspect IDs more than guilty suspect IDs. But the diagnostic-feature-detection theory makes a qualitatively different a priori prediction that the filler-siphoning account does not make—$d'_{\text{Innocent-Guilty}}$ should increase in fair lineups. Of course, the findings reported by Colloff et al. (2016) do not prove diagnostic-feature-detection theory is necessarily correct—like any new theory, it needs testing and refining, but the available evidence suggests a diagnostic-feature-detection mechanism is a compelling possibility (e.g., Flowe, Klatt, & Colloff, 2014). Moreover, in our view, a theory such as the diagnostic-feature-detection model is more likely to advance our understanding than filler-siphoning because it is a well-specified, quantitatively-defined theory which makes specific, testable predictions about $d'_{\text{Innocent-Guilty}}$, whereas the filler-siphoning account does not.

A

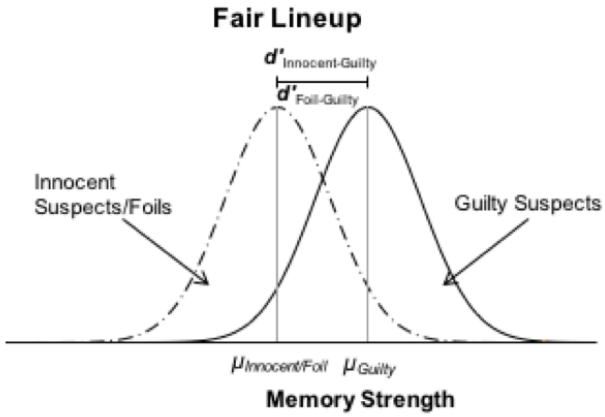


B

	BEST		INTEGRATION	
Estimate	Fair	Unfair	Fair	Unfair
$d'_{Innocent\text{-}Guilty}$	0.86	0.54	0.99	0.66
$d'_{Foil\text{-}Guilty}$	0.86	1.79	0.99	2.00
χ^2	13.31	36.10	32.91	191.61
df	9	13	9	13
p	.149	<.001	<.001	<.001

Note: A lower chi-square indicates a better fit.

C



D

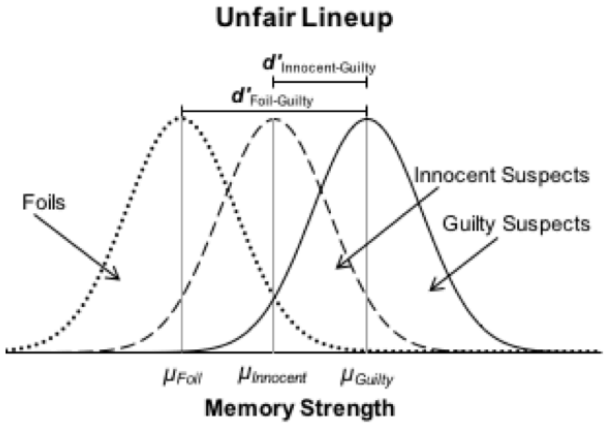


Figure 2. (A) Partial Receiver Operating Characteristic curves for the fair (block) and unfair (do-nothing) showup conditions ($p < .001$), with lines of best fit drawn using the best-fitting parameters from a signal-detection model; (B) Discriminability estimates and goodness-of-fit statistics for the best-fitting versions of the BEST and INTEGRATION models to data from the fair (replication) and unfair (do-nothing) conditions of Colloff et al. (2016); and signal-detection interpretations of (C) fair lineups, where $d'_{Innocent\text{-}Guilty} = d'_{Foil\text{-}Guilty}$, and (D) unfair lineups, where $d'_{Innocent\text{-}Guilty} \neq d'_{Foil\text{-}Guilty}$.

References

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35,* 364–380. doi:10.1007/s10979-010-9245-1

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27,* 1227–1239. doi:10.1177/0956797616655789

Duncan, M. (2006). *A signal detection model of compound decision tasks.* (Tech Note DRDC TR 2006-25). Toronto, ON: Defence Research and Development Canada.

Flowe, H. D., Klatt, T., & Colloff, M. F. (2014). Selecting fillers on emotional appearance improves lineup identification accuracy. *Law and Human Behavior, 38,* 509–519. http://doi.org/10.1037/lhb0000101

Green, D. M., & Swets J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Macmillan N. A. & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361–376. doi:10.1037/a0030609

Mickes, L., Seale-Carlisle, T.M., Wetmore, S.A., Gronlund, S.D., Clark, S.E., Carlson, C.A., Goodsell, C.A., Weatherford, D., & Wixted, J.T. (2017). ROCs in Eyewitness Identification: Instructions vs. Confidence Ratings. *Applied Cognitive Psychology, 31,* 467–477. doi:10.1002/acp.3344

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied, 16,* 387–398. doi: 10.1037/a0021034

Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications,* 1:10. doi: 10.1186/s41235-016-0006-7.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review, 22,* 944-954.

Smith, A. M., Wells, G. L., Smalarz, L., & Lampinen, J. M. (2017). Increasing the Similarity of Lineup Fillers to the Suspect Improves the Applied Value of Lineups Without Improving Memory Performance: Commentary on Colloff, Wade, & Strange (2016). *Psychological Science.*

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121,* 262–276. doi:10.1037/a0035940

Wixted, J. T., & Mickes, L. (2018). Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications,* 3:9. doi:10.1186/s41235-018-0093-8

Author Contributions

M. F. Colloff and J. T. Wixted developed the showup study concept. All authors contributed to the study design. M. F. Colloff., K. A. Wade., and D. Strange collected data. J. T. Wixted provided the MATLAB model-fitting routines and M. F. Colloff performed the data analysis and interpretation under the supervision of J. T. Wixted. M. F. Colloff and K. A. Wade drafted the manuscript and J. T. Wixted and D. Strange provided critical revisions. All authors approved the final version of the manuscript for submission.