Reply

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 (2018)

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

[1]Centre for Applied Psychology, School of Psychology, University of Birmingham; [2]Department of Psychology, University of Warwick; [3]Department of Psychology, John Jay College of Criminal Justice, City University of New York; and [4]Department of Psychology, University of California, San Diego

In our original article (Colloff, Wade, & Strange, 2016), we tested 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 that are nondiagnostic 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., $d'_{innocent-guilty}$) should be better in fair lineups than in unfair lineups (i.e., lineups in which the suspect does not or does stand out, respectively). Indeed, our data confirmed that prediction.

Smith, Wells, Smalarz, and Lampinen (2018) argue instead that (a) fair lineups do not improve but instead worsen people's memory performance and (b) 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 fitted the wrong signal detection model to the data. With regard to the second point, Smith et al. proposed differential-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 hypothesized to be differential, with similar-looking foils attracting more identifications when the suspect in the lineup is innocent than when he or she is guilty. Thus, differential 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 modeled our data. In what follows, we first 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 $d'_{innocent-guilty}$ that we observed. We then present new data from an experiment that tested the same prediction that was tested in our original article, but this time with no foils involved (eliminating the possibility of filler siphoning). Finally, we illustrate that the signal detection model we fitted to the data was appropriate and that the model preferred by Smith et al., when fit to the data as it should be, confirms that $d'_{innocent-guilty}$ was higher in the fair-lineup condition, as predicted by diagnostic-feature-detection theory.

Filler Siphoning Does Not Make a Prediction About $d'_{innocent-guilty}$

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

Corresponding Author:
Melissa F. Colloff, University of Birmingham, School of Psychology, Birmingham, United Kingdom, B15 2TT
E-mail: M.Colloff@bham.ac.uk

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 the false alarm rate to innocent suspects decrease differentially: 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 the data reported in our original article are fully consistent with differential-filler-siphoning theory in this respect. Indeed, we said so (Colloff et al., 2016, Supplemental Material, p. DS7).

Critically, however, a manipulation that decreases identifications of innocent suspects more than it decreases identifications of guilty suspects (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 $d'_{\text{innocent-guilty}}$. Therefore, filler-siphoning theory does not make a prediction about the specific change in $d'_{\text{innocent-guilty}}$ that we observed in the data reported in Colloff et al. (2016). Diagnostic-feature-detection theory, however, specifically predicts the change in $d'_{\text{innocent-guilty}}$ 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 that $d'_{\text{innocent-guilty}}$ 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 (area shaded gray), and only these foils compete for identifications 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 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 the far-left plot of Figure 1b, $d'_{\text{innocent-guilty}}$ remains unchanged as lineups become fairer (contrary to diagnostic-feature-detection theory); in the middle plot, $d'_{\text{innocent-guilty}}$ decreases as lineups become fairer (again, contrary to diagnostic-feature-detection theory); and in the far-right plot, $d'_{\text{innocent-guilty}}$ 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, 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 ID rates for innocent and guilty suspects both decrease. In all three scenarios, the foil distribution overtakes a greater proportion of the innocent-suspect distribution than the guilty-suspect distribution. This means that the innocent-suspect ID rate will decrease more than the guilty-suspect ID rate. Hence, differential filler siphoning is predicted to occur 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—filler-siphoning theory. 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.

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 suspect in the target-absent showup and the guilty suspect in the target-present showup 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 because
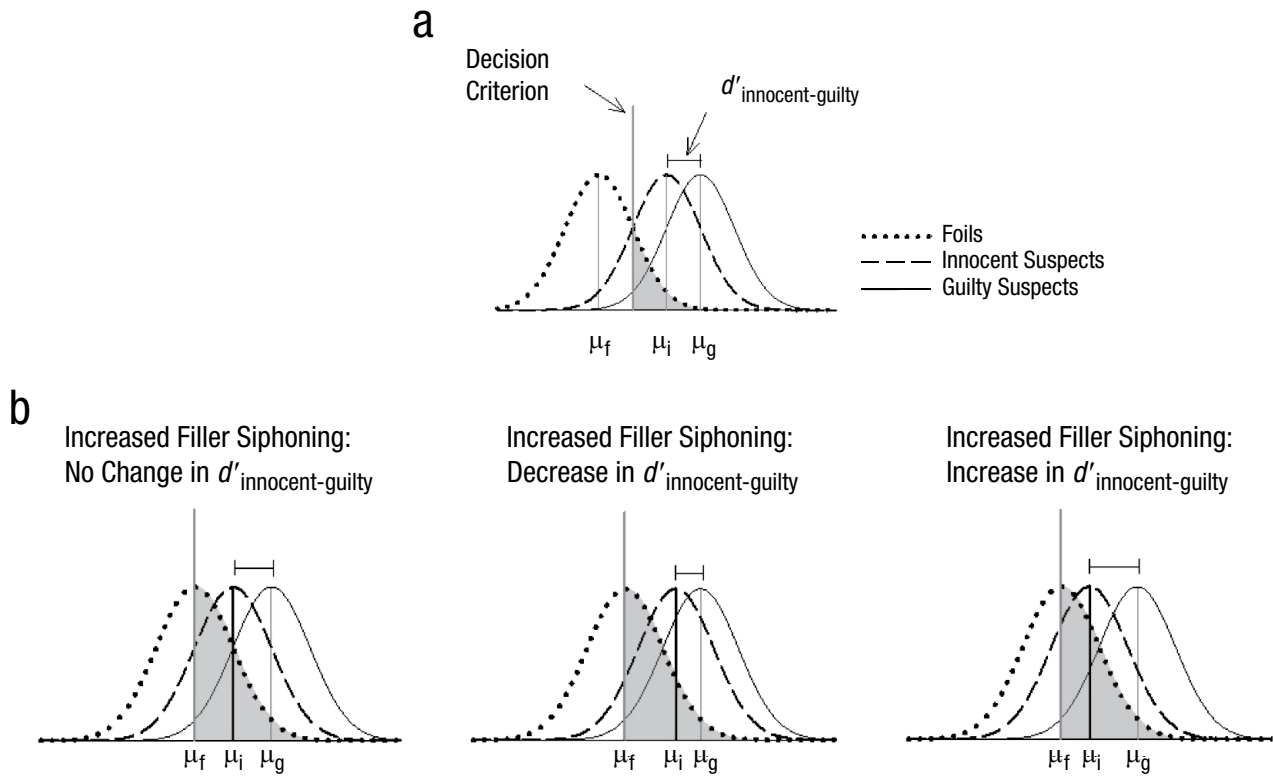
a

b

Fig. 1. Signal detection interpretation of (a) an unfair lineup and (b) three different ways in which $d'_{\text{innocent-guilty}}$ (the ability to discriminate between innocent and guilty suspects) can change, independently of filler siphoning, when a fairer lineup is used. The dotted, dashed, and solid distributions represent the memory-strength values of foils, innocent suspects, and guilty suspects, with mean memory strengths of μ_f, μ_i, and μ_g, respectively. The vertical line inside each distribution depicts the mean memory strength for that distribution. When fairer lineups are used, filler siphoning increases because a larger proportion of the filler distribution exceeds the decision criterion (gray-shaded area) and overlaps with the innocent-suspect and guilty-suspect distributions. Filler siphoning is differential (i.e., reduces the ID rate for innocent suspects more than the ID rate for guilty suspects) because the filler distribution overlaps a greater proportion of the innocent-suspect distribution than the guilty-suspect distribution. Panel (b) illustrates that differential filler siphoning occurs no matter what the effect of changing to fairer lineups might be on $d'_{\text{innocent-guilty}}$. Diagnostic-feature-detection theory predicts the outcome illustrated in the right-hand plot in (b), namely that $d'_{\text{innocent-guilty}}$ should increase when a fairer lineup is used, for reasons briefly described in this Reply and in more detail in Colloff, Wade, and Strange (2016). Note that in (b), we depict lineups that are less unfair (not perfectly fair) compared with (a) to clearly show the predictions made by the filler-siphoning and diagnostic-feature-detection accounts as lineups become increasingly fair. However, the point illustrated in (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 (a), including to perfectly fair lineups (as depicted in Fig. 2b).

the area of the feature was covered with a black rectangle. 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 the fair-showup condition). Theoretically, a fair showup prevents witnesses from relying on a nondiagnostic feature by removing it altogether from the decision, enhancing the ability of witnesses to discriminate between innocent and guilty suspects. Analogously, 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 analyzed the showup data in the same way that we analyzed the lineup data in our original article and found the same result. Receiver-operating-characteristic (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—partial area under the curve (pAUC) = .102, 95% confidence interval (CI) = [.091, .112]—than in unfair showups that allowed people to rely on the nondiagnostic distinctive feature—pAUC = .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 for fair showups ($d_{\text{innocent-guilty}} = 1.13$) than unfair showups ($d_{\text{innocent-guilty}} = 0.92$). Note that these analyses are based on participants who identified
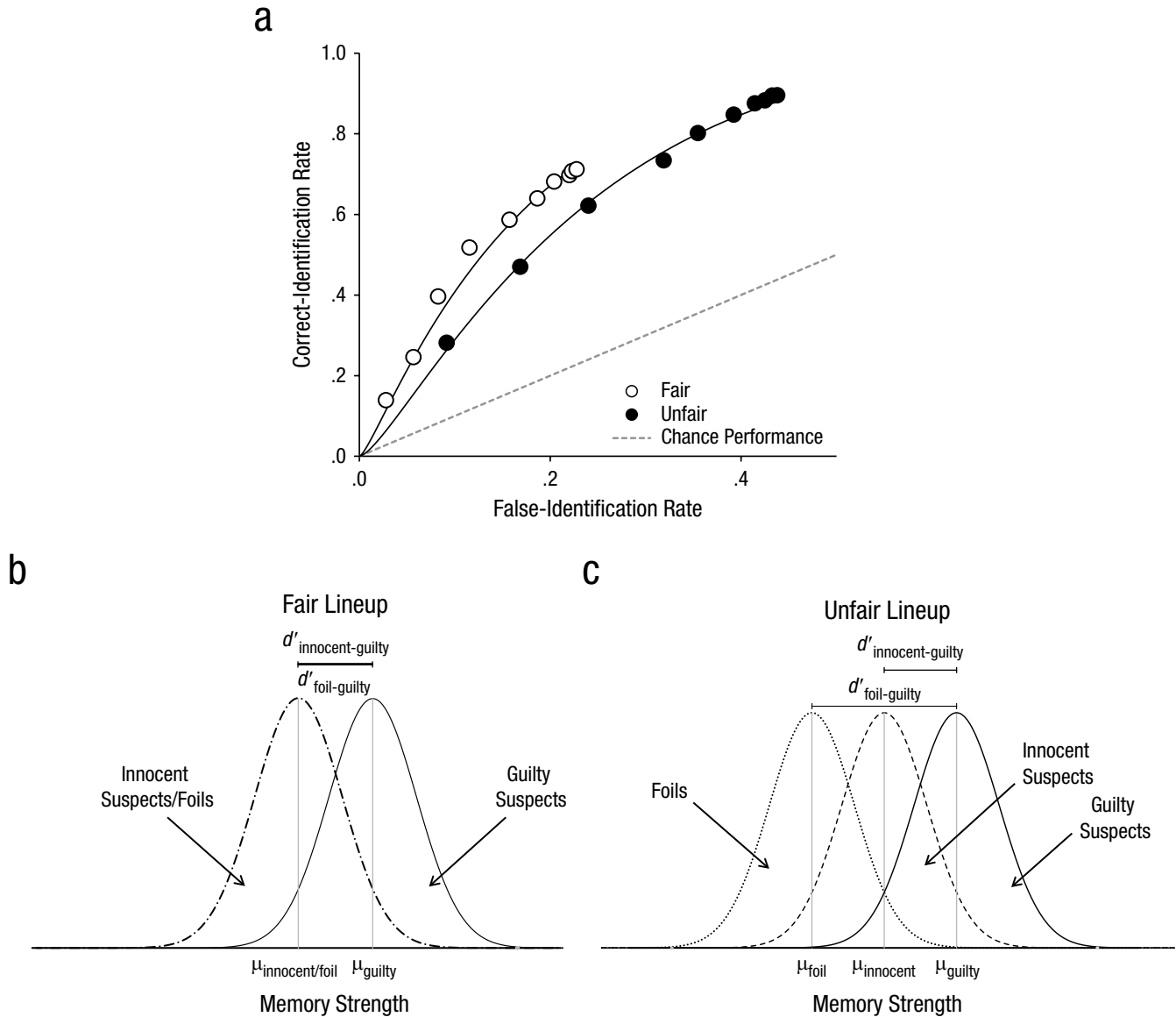
Fig. 2. Showup-study results and signal detection models for fair and unfair lineups. The graph in (a) shows partial receiver-operating-characteristic curves for the fair (block) and unfair (do-nothing) showup conditions ($p < .001$), with lines of best fit drawn using the best-fitting parameters from a signal detection model. The graphs in (b) and (c) show signal detection interpretations of (b) fair lineups, where $d'_{\text{innocent-guilty}} = d'_{\text{foil-guilty}}$, and (c) unfair lineups, where $d'_{\text{innocent-guilty}} \neq d'_{\text{foil-guilty}}$.

innocent or guilty suspects in accordance with our preregistered plans; when full ROC curves are plotted and modeled, the conclusions remain the same (see the Supplemental Material available online). Critically, these findings cannot be explained by filler siphoning.

An Empirical Comparison of Smith et al.'s Model With Our Model

Smith et al. also argued that the model we used to estimate $d'_{\text{innocent-guilty}}$ was inappropriate because it (a) misclassified "filler identifications as rejections"

(p. 1550) and (b) was a "simple-detection model" that did not "have both detection and identification components" (p. 1549). Smith et al. fitted a different signal detection model to the data that they argued was more appropriate. On the basis of 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 Colloff et al., 2016, Table S2 in the Supplemental Material). Also, our model is a compound

Table 1. Discriminability Estimates and Goodness-of-Fit Statistics for the Best-Fitting Versions of the Best-Above-Criterion and Integration Models to Data From the Fair (Replication) and Unfair (Do-Nothing) Conditions of Colloff, Wade, and Strange (2016)

Estimate	Best-above-criterion model		Integration model	
	Fair	Unfair	Fair	Unfair
$d'_{\text{innocent-guilty}}$	0.86	0.54	0.99	0.66
$d'_{\text{foil-guilty}}$	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 value indicates a better fit.

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 (Clark, Erickson, & Breneman, 2011; Macmillan & Creelman, 2005), whereas Smith et al.'s uses the *integration* rule (Palmer, Brewer, & Weber, 2010). When we empirically compared the two models, we found that the best-above-criterion model offered a noticeably better fit (see Table 1). Nonetheless, even Smith et al.'s integration model supports our original conclusion: $d'_{\text{innocent-guilty}}$ is higher in fair lineups than in unfair lineups according to both models (see Table 1 and the Supplemental Material).

If the best-above-criterion model is not 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 (Fig. 2b) but not when the lineup is unfair (Fig. 2c). 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 $d'_{\text{foil-guilty}}$ (the ability to discriminate between foils and guilty suspects) because the innocent suspect and the foils are equally similar to the culprit (Fig. 2b)—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 $d'_{\text{foil-guilty}}$ (Fig. 2c). Even when analyzing the unfair-lineup data, Smith et al. combined innocent-suspect and foil IDs from target-absent lineups, as if they were drawn from the same memory-strength distribution (reducing a three-distribution model to a two-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 the Supplemental Material).

To summarize, we agree that filler siphoning occurs to a greater extent in fair than unfair lineups, reducing identifications of innocent suspects more than identifications of guilty suspects. But 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 in Colloff et al. (2016) do not prove that 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 that makes specific, testable predictions about $d'_{\text{innocent-guilty}}$, whereas the filler-siphoning account does not.

Action Editor

D. Stephen Lindsay served as action editor for this article.

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 the data. J. T. Wixted provided the MATLAB (The MathWorks, Natick, MA) model-fitting routines, and M. F. Colloff analyzed and interpreted the data 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.

ORCID iDs

Melissa F. Colloff (iD) https://orcid.org/0000-0001-6401-4872
Kimberley A. Wade (iD) https://orcid.org/0000-0002-7134-9600

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618786459

Open Practices

All data have been made publicly available via the Open Science Framework (OSF) and can be accessed at osf.io/nr24b. Stimulus materials can be obtained from the corresponding author. The design and analysis plans for the experiment were preregistered at the OSF and can be accessed at osf.io/7upqk. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618786459. This article has received badges for Open Data and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

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* (Technical Report No. TR 2006-256). Toronto, Ontario, Canada: 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. doi:10.1037/lhb0000101

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: 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., . . . Wixted, J. T. (2017). ROCs in eyewitness identification: Instructions versus 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*, Article 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. (2018). Increasing the similarity of lineup fillers to the suspect improves the applied value of lineups without improving memory performance: Commentary on Colloff, Wade, and Strange (2016). *Psychological Science*, *29*, 1548–1551.

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*, Article 9. doi:10.1186/s41235-018-0093-8