# Supplemental Material:

**Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T.** 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). *Psychological Science.*

**Table of Contents**

**Showup Experiment Testing a Diagnostic-Feature-Detection Mechanism…..…..Page 1**

Method……..……...……..……..…….....Page 2

Procedure……..……..……..……..…….Page 3

Results & Discussion……..……..……...……..…Page 4

**An Empirical Comparison of the Models: Fitting the BEST Model vs. the INTEGRATION Model to the Colloff et al. (2016) Data…….…..…….……..Page 12**

## Showup Experiment Testing a Diagnostic-Feature-Detection Mechanism

Colloff, Wade, and Strange (2016) measured people's ability to discriminate between innocent and guilty suspects in fair versus unfair lineups. Subjects first watched a video in which a culprit with a distinctive feature, such as a black eye, committed a simulated crime. In the unfair lineup (our do-nothing condition), the suspect, whether innocent or guilty, had that distinctive feature, but none of the foils did. In the fair lineups, either everyone in the lineup had the distinctive feature (replication condition) or no one did (block condition and pixelation condition; see Colloff et al., 2016, Figure 1). Subjects were asked to identify the culprit that they had seen in the video, and they were much less likely to choose the suspect (innocent or guilty) when the lineup was fair. This is a well-known phenomenon that reflects increased filler-siphoning in fair lineups. Crucially, the use of fair lineups also increased people's ability to discriminate between innocent and guilty suspects—an outcome that is uniquely predicted by the diagnostic-feature-detection theory (Wixted & Mickes, 2014; see also Colloff, Wade, Wixted, & Maylor, 2017).

To further test the prediction made by the diagnostic-feature-detection theory, we reran two conditions tested by Colloff et al. (2016), but we removed the foils from the identification procedure. That is, we tested our original block (fair) versus do-nothing (unfair) conditions, but we used showups instead of lineups. In a showup, subjects are tested with a

single suspect who is either innocent or guilty. In the "fair" (block) showup condition, neither the innocent nor the guilty suspect had a visible distinctive feature, as in the corresponding block condition of Colloff et al. (2016). In the "unfair" (do-nothing) showup condition, both the innocent suspect and the guilty suspect had the distinctive feature, as in the corresponding do-nothing condition of Colloff et al. (2016). The key premise of the diagnostic-feature-detection account is that procedures that prevent reliance on non-diagnostic features (those features shared by innocent and guilty suspects) improve people's ability to discriminate between innocent and guilty suspects (i.e., increase $d_{Innocent\text{-}Guilty}$). Thus, a diagnostic-feature-detection mechanism predicts that $d_{Innocent\text{-}Guilty}$ will be larger in fair showups than unfair showups because in fair showups subjects cannot rely on the non-diagnostic distinctive feature so they will be less likely to confuse innocent and guilty suspects. Critically, the diagnostic-feature-detection theory makes this prediction for the same reason it made that prediction in the corresponding lineup condition in the Colloff et al. (2016) study. In both cases, the prediction has nothing to do with identifications of foils. If $d_{Innocent\text{-}Guilty}$ is indeed larger in fair showups than unfair showups, then this difference could not have arisen from a differential filler-siphoning mechanism. However, if our original result were due to filler-siphoning or to any filler-dependent phenomenon, then the predicted effect should no longer be observed.

## Method

### *Design*

We used a 2 (showup type: block, do-nothing) × 2 (target: present, absent) mixed design, with target manipulated within subjects. We collected two data points per subject because each subject watched two mock crime videos and completed two identification tasks (one target-present, one target-absent). ROC analyses in lineup research requires large samples, but the techniques for conducting power analysis are not well defined. ROC lineup studies usually recruit between 300 and 500 subjects per condition. Therefore, our data collection stopping rule was to recruit at least 2,000 subjects with useable data, so that we had at least 500 subjects in each condition. Using the mean difference and standard deviations observed in our original lineup study as a guide (Colloff et al., 2016), a power analysis indicated that, with 500 subjects per condition, power for this showup experiment would exceed 80%. We pre-registered our study before we started data collection and our data are available online (see https://osf.io/nr24b/).

### *Subjects*

The subjects were 2,368 adults who completed the task online. In total, we excluded 290 subjects (12%; between 64 and 91 in each of the eight cells) because they experienced technical difficulties while watching the video ($n$ = 25, <1%), stated that they had viewed the videos before or completed the study more than once ($n$ = 210, 9%), or incorrectly answered an attention-check question on the content of the video ($n$ = 55, 2%). These exclusions resulted in a final sample size of 2,078; with between 518 and 522 subjects in each of the eight cells (884 male, 1,133 female, 61 other or prefer not to say; age range = 17–78 years, $M$ = 33.59, $SD$ = 12.40). The majority of the sample self-identified as Caucasian (52.60%), the remainder identified as Asian (24.92%), Latin, Hispanic, or Mexican-American (8.47%), Black, African, African-American or Caribbean (5.92%), Filipino (0.87%), Native-American (0.67%), or Other (4.91%), while 1.64% chose not to disclose their race or ethnicity. Of the final sample, 1,667 subjects were recruited via Amazon Mechanical Turk and received $0.35, 380 were recruited via snowball sampling from social-networking sites and email advertisements and were entered into a prize drawing for a £25 Amazon voucher, and 31 students were recruited from John Jay College of Criminal Justice and received extra credit in a course. We combined all data for the analyses.

### *Materials*

The materials were from Colloff et al. (2016). We used two 30 s mock crime videos. In the mugging video, the male culprit had a tribal tattoo on his cheek. In the graffiti video, the male culprit had a black-eye. Target-present showups were an image of the guilty suspect (i.e., the culprit from the video), whereas target-absent showups were an image of an innocent suspect (i.e., not the culprit from the video). Colloff et al. compiled a pool of 40 matched-to-description faces for each culprit. For each subject, we randomly selected an individual from these pools to serve as the innocent suspect. In fair (block) showups, the area of the culprit's distinctive feature was concealed by a solid black rectangle. In unfair (do-nothing) showups, the suspect had a visible distinctive feature.

## Procedure

Subjects were told that the study was about perception and memory and were randomly assigned into conditions. First, subjects watched a mock crime video (mugging or graffiti) labelled as "Video A." We told subjects that they should pay close attention, because they

would be asked questions about the content of the video. Following the video, we asked subjects if they experienced any technical difficulties while playing the video, and then gave subjects 4 min to complete some spatial reasoning questions.

After 4 min, the study automatically advanced and we told subjects: "On the next page you will be presented with a photograph which may or may not be the male perpetrator you saw in Video A." On the next page, a single image of the suspect was displayed. The showup technique (block or do-nothing) and format (target-present or target-absent) depended on the condition to which the subject had been randomly assigned. We asked subjects whether or not the person in the photograph was the person that they saw in Video A and then asked them to rate their confidence in their identification decision on an 11-point Likert-type scale ranging from 0% (*completely uncertain*) to 100% (*completely certain*). Following these questions, we asked subjects to answer an attention-check question about the content of the video.

Next, we had subjects complete the same sequence of tasks again, this time viewing the alternate mock crime video (mugging or graffiti) and showup format (target-present or target-absent). Here, the video and tasks were labelled as "Video B." Subjects remained in the same showup technique condition to which they had been assigned: Subjects who were presented with a fair (block) showup after "Video A," for instance, were also presented with a fair (block) showup after "Video B." The order of the videos and target conditions was counterbalanced. At the end of the study, we checked if subjects had seen either of the videos before and we asked several demographic questions (e.g., age, gender, ethnicity).

**Results & Discussion**

To date, ROC lineup studies have focused on the performance of subjects who made an identification (i.e., subjects who responded: "Yes, that is the culprit") and the diagnostic-feature-detection theory was developed to account for these findings. As such, we specified in our pre-registration that our showup analysis would be performed on Yes responses. That is, in our pre-registration, we specifically stated that we would construct partial ROC curves and measure partial Area Under the Curve (*p*AUC) by plotting the hit rate (HR; subjects who *made an identification* of a guilty suspect ÷ number of target-present showups) against the false alarm rate (FAR; subjects who *made an identification* of an innocent suspect ÷ number of target-absent showups) over decreasing levels of confidence. We also stated that we would fit a signal-detection model using counts of hits from target-present showups and false alarms

from target-absent showups. ROC showup studies, however, also allow for the analysis of subjects who respond: "No, that is not the culprit", because subjects who respond No can rate how confident they are that the suspect is not the culprit. This means that full ROC curves (which extend to HR = 1 and FAR = 1) can be constructed and the full Area Under the Curve (AUC) can be measured. Correspondingly, a signal-detection model can be fit to the full ROC data. Here, we report both our pre-registered analysis on Yes responses (*p*AUC and modelling) as it directly corresponds to how we analysed the lineup data in our original study, followed by additional analyses (AUC and modelling) based on the full ROC data.

***Yes Responses: Partial Area Under the Curve (pAUC) Analysis***

Figure 2A in our main reply shows the partial ROC curves. We used the statistical package *pROC* (Robin et al., 2011) with RStudio (RStudio Team, 2015) and the R software environment (R Development Core Team, 2015) to calculate *p*AUC and *D,* a measure of effect size: $D = (AUC1 - AUC2)/s$, where *s* is the standard error of the difference between the two AUCs and is estimated using bootstrapping. We defined the specificity (1 – FAR) using the smallest false alarm rate (FAR) range in the comparison (i.e., specificity = .77). As reported in our main reply, people were better able to discriminate between innocent and guilty suspects in fair than unfair showups. This pattern of results was observed in both the mugging and graffiti stimulus sets, which indicates that our findings are not driven by one particular set of encoding and test conditions (these analyses are available at https://osf.io/nr24b/). Note that defining the specificity using the smallest FAR range means that the *p*AUC analysis only includes the identification decisions made with the highest confidence levels in the unfair showup condition. Limiting the *p*AUC analysis to a small subset of the unfair curve did not affect our conclusions. We found the same results when we fit a signal-detection model to our data, which includes responses made at every confidence level in fair and unfair showups. We describe this model-fitting next.

***Yes Responses: Signal-detection Model***

A *p*AUC analysis is atheoretical and need not agree with the interpretation provided by fitting a signal-detection model to the ROC data. Therefore, we fit a signal-detection model to further confirm our findings. This signal-detection-based analysis is the one that bears most directly on the key prediction made by the diagnostic-feature-detection account (i.e., that $d_{\text{Innocent-Guilty}}$ should be higher in fair showups), because a model-based analysis allows us

to directly measure $d_{\text{Innocent-Guilty}}$. The signal-detection model for showups assumes two Gaussian distributions representing the memory strength values for innocent suspects and guilty suspects. The distance between the $\mu_{innocent}$ and $\mu_{guilty}$ distributions ($d_{Innocent\text{-}Guilty}$) reflects ability to discriminate between innocent and guilty suspects. The model also assumes that there is a set of response criteria that reflect different levels of confidence. To reduce the number of parameters, we combined confidence ratings of Yes 0-20 ($c_1$), Yes 30-40 ($c_2$), Yes 50-60 ($c_3$), Yes 70-80 ($c_4$), and Yes 90-100 ($c_5$) to create a 5-point confidence scale. The model assumes that a positive (Yes) identification is made when the suspect's face is familiar enough to exceed $c_1$, and the confidence in the identification is determined by the highest criterion that is exceeded.

We fit an unequal-variance model, because this provided a significantly better fit to the data than an equal-variance model, as is typically true of list-memory tasks involving single test items. Usually, in list-memory studies, $\sigma_{target}$ (analogous to $\sigma_{guilty}$ here) is greater than 1, but here the analogous parameter was estimated to be less than 1. This result is to be expected because the target (i.e., the guilty suspect) was fixed across participants, whereas the innocent suspect was randomly selected from a pool of faces and therefore varied across participants. This additional source of item variance would be expected to (and did) increase the variance of the innocent suspect distribution relative to the guilty suspect distribution. Hence, $\sigma_{guilty}$ should be (and was) less than 1.[1] Moreover, it is clear from Figure that responding is more liberal in the unfair (do-nothing) showup than the fair (block) showup. Therefore, to minimize the number of free parameters, we also constrained the confidence criteria across the fair and unfair showup conditions using an additive factor. The model had 9 parameters ($\mu_{guilty(fair)}$, $\mu_{guilty(unfair)}$, $c_1$, $c_2$, $c_3$, $c_4$, $c_5$, $\sigma_{guilty,}$ and the additive factor) and both the fair and unfair showups had 10 degrees of freedom in the data (5 levels of confidence for guilty suspect identifications and 5 levels of confidence for innocent suspect identifications). Thus, the fit of the model to the data involved 20 – 9 = 11 degrees of freedom. We fixed $\mu_{innocent}$ and $\sigma_{innocent}$ to 0 and 1, respectively. Once the frequencies of positive identifications were entered into the model, the frequencies of reject identifications were fixed. We fit the model to the data by minimizing the chi-square goodness-of-fit statistic.

We first fit the unequal-variance model allowing $d_{Innocent\text{-}Guilty}$ to differ across the fair

[1] Note that we set $\sigma_{guilty}$ to be the same in fair and unfair showups, because adding an extra parameter and allowing $\sigma_{guilty}$ to vary across fair and unfair showups (i.e., using $\sigma_{guilty(fair)}$ and $\sigma_{guilty(unfair)}$) did not significantly improve the fit.

and unfair showup conditions (full model)[2]. We used the best-fitting model parameters to draw the lines of best fit on Figure 2A in our main reply. It is clear from the correspondence between the empirical data points and the model-predicted lines of best fit, that the model was able to capture the trends in our data. This is corroborated by the non-significant chi-square goodness-of-fit statistic in Table S1 (Yes response analysis, full model column). Looking at the best-fitting parameters in the full model columns in Table S1, it is clear that $d_{Innocent\text{-}Guilty}$ is higher in fair showups than unfair showups. Next, we fit the same model but we constrained $d_{Innocent\text{-}Guilty}$ to be equal across the fair and unfair showups, while allowing the confidence criteria to differ across conditions by an additive factor (constrained model; note that the additive factor was allowed to vary between the full and constrained models). The constrained model provided a significantly worse fit of the data than the full model, $\chi^2$ (1) = 10.17, $p$ = .001. This indicates that $d_{Innocent\text{-}Guilty}$ is significantly larger in the fair (block) showup condition than the unfair (do-nothing) showup condition.

Taken together, the results of our model-fitting exercise are concordant with our partial ROC analyses: Both suggest that fair showups enhance people's ability to discriminate between innocent and guilty suspects more than unfair showups. This pattern occurs, even though there is no opportunity for filler-siphoning. Thus, our experiment supports the prediction made by the diagnostic-feature-detection theory; an effect that cannot be attributed to filler-siphoning because there are no foils in showups.

[2] $d$ (the distance between the means of the innocent and guilty distributions in units of $\sigma_{innocent}$) provides a useful discriminability measure when unequal variance is assumed and the magnitude of the unequal variance parameter does not vary across conditions (i.e., $\sigma_{guilty}$ is the same across fair and unfair showups). Under such conditions, it is linearly related to the standard discriminability measure used in the unequal-variance case, namely, $d_a$.

Table S1

*Full and Constrained Model Fits for the Fair (Block) vs. Unfair (Do-nothing) Comparisons in the "Yes Response" Analysis and "Yes and No Response" Analysis*

| | Yes Response Analysis | | | | Yes and No Response Analysis | | | |
|---|---|---|---|---|---|---|---|---|
| | Full Model | | Constrained Model | | Full Model | | Constrained Model | |
| Estimate | Fair | Unfair | Fair | Unfair | Fair | Unfair | Fair | Unfair |
| $d_{Innocent\text{-}Guilty}$ | 1.13 | 0.92 | 1.02 | 1.02 | 1.12 | 0.93 | 1.03 | 1.03 |
| $\sigma_{Guilty}$ | 0.65 | 0.65 | 0.67 | 0.67 | 0.88 | 0.73 | 0.82 | 0.82 |
| $c_1$ | 0.75 | 0.14 | 0.66 | 0.19 | –0.50 | –0.75 | –0.49 | –0.77 |
| $c_2$ | 0.78 | 0.17 | 0.69 | 0.22 | 0.00 | –0.26 | 0.00 | –0.27 |
| $c_3$ | 0.88 | 0.27 | 0.80 | 0.33 | 0.66 | 0.09 | 0.63 | 0.09 |
| $c_4$ | 1.12 | 0.51 | 1.03 | 0.57 | 1.11 | 0.48 | 1.05 | 0.51 |
| $c_5$ | 1.58 | 0.97 | 1.51 | 1.04 | 1.70 | 0.98 | 1.60 | 1.06 |
| Overall $\chi^2$ | 19.25 | | 29.42 | | 46.26 | | 56.78 | |
| Overall df | 11 | | 12 | | 6 | | 8 | |
| Overall *p* | .057 | | .003 | | < .001 | | < .001 | |

*Note.* For the Yes response analysis, the full model allows *d* to differ between the fair vs. unfair showup comparison, while the constrained model holds *d* constant across the fair vs. unfair comparison. In the Yes and No response analysis, the full model allows *d* and σ (i.e., $d_a$) to differ between the fair vs. unfair showup comparison, while the constrained model holds *d* and σ (i.e., $d_a$) constant across the fair vs. unfair comparison. The overall $\chi^2$, df and *p* goodness-of-fit statistics show that in both analyses, the fit was worse in the constrained model compared to the full model.

### *Yes and No Responses: Area Under the Curve (AUC) Analysis*

We next analyzed the full ROCs, which is how old/new ROCs are typically analyzed in list-memory studies. In brief, to construct the full ROC curves in Figure S1 we formed a single 21-point confidence scale, ranging from Yes 100 to Yes 0 then No 0 to No 100. We collapsed Yes 0 and No 0 into one category. In Figure S1, every hit rate includes in the denominator the total number of trials which presented the guilty suspect at test. Every false alarm rate includes in the denominator the total number of trials which presented the innocent suspect at test. All that changes across the ROC is the proportion of trials included in the numerator of each measure. The leftmost ROC point includes only those IDs made with the highest level of confidence (Yes 100). The next point includes only those IDs made with highest and the second-highest level of confidence (Yes 100 and Yes 90). This continues all the way down the scale, crashing right through the arbitrary Yes/No point on the confidence scale. As such, the bottom half of each curve in Figure S1 (i.e., on the left side of the graph) corresponds exactly to the curves displayed in Figure 1B in our main reply. In Figure S1, the curves have just been extended (i.e., on the right side of the graph) to also take into account gradations of confidence in No responses.

Figure S1B shows that the ROC curve for fair (block) showups is higher than the ROC curve for unfair (do-nothing) showups at the left side of the graph (i.e., Yes responses), but

the curves begin to overlap around the mid-point of the curves where No responses are included. AUC analysis showed that people were better at discriminating between innocent and guilty suspects in fair showups, AUC = 0.800, 95% CI = [0.781, 0.819], than in unfair showups, AUC = 0.775, 95% CI = [0.755, 0.795]), $D$ = 1.80, but the effect was not statistically significant, $p$ = .07. Note this experiment was conducted with an a priori directional prediction that was specified in the pre-registration (i.e., in our pre-registration we hypothesized: "unfair (do-nothing) showups will impair people's ability to discriminate between innocent and guilty suspects more than fair (block) showups."). Thus, it would not be unreasonable to apply a one-tailed test, in which case the result would be statistically significant at $p$ = .035. Whether a one-tailed or a two-tailed test is used, the results suggest that the fair showup advantage that was clearly evident in subjects who said "Yes, that is the culprit" is reduced when subjects who said "No, that is not the culprit" are included. We next fit a signal-detection model to the full ROC data. As noted earlier, a fit of a theoretical signal-detection model provides the most direct test of the prediction made by the diagnostic-feature-detection account.

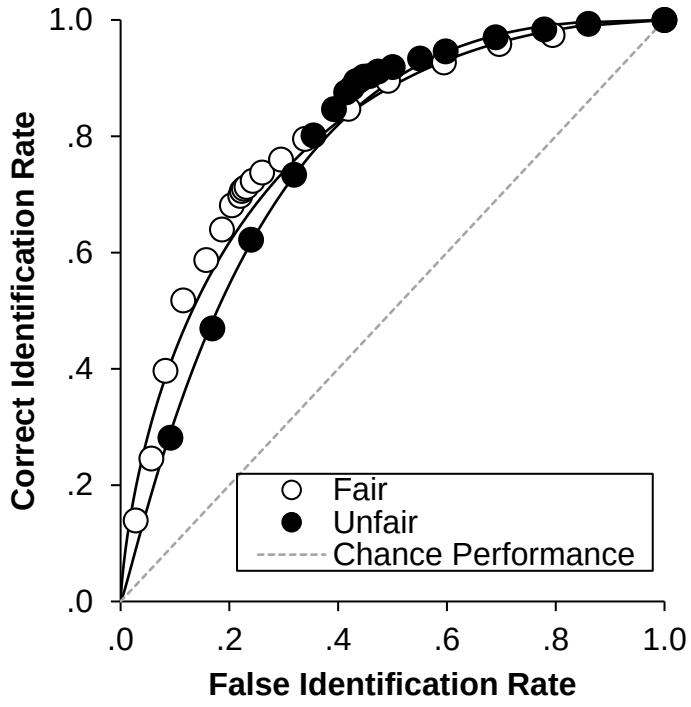


**Figure S1.** Receiver Operating Characteristic (ROC) curves for the fair (block) and unfair (do-nothing) showup conditions. The lines of best fit are drawn using the best-fitting parameters from the corresponding signal-detection model analysis.

***Signal-detection Model: Yes & No Responses***

For this test, we created a single 5-point confidence scale by combining confidence ratings of No 80-70 ($c_1$), No 60-0 ($c_2$), Yes 0-60 ($c_3$), Yes 70-80 ($c_4$), and Yes 90-100 ($c_5$). Note that the final confidence category—No 100-90—is fixed once the other categories are specified. The model assumes that a positive identification is made when the suspect's face is familiar enough to exceed $c_3$, and the confidence in the identification is determined by the highest criterion that is exceeded. As before, we fixed $\mu_{innocent}$ and $\sigma_{innocent}$ to 0 and 1, respectively, and fit the model to the data by minimizing the chi-square goodness-of-fit statistic. The model had 14 parameters ($\mu_{guilty(fair)}$, $\mu_{guilty(unfair)}$, $\sigma_{guilty(fair)}$, $\sigma_{guilty(unfair)}$, *and* $c_1$, $c_2$, $c_3$, $c_4$, $c_5$ for fair and unfair showups) and both fair and unfair showups had 10 degrees of freedom in the data (5 levels of confidence for guilty suspect identifications and 5 levels of confidence for innocent suspect identifications). Thus, the fit of the model to the data involved 20 – 14 = 6 degrees of freedom.

We first fit the unequal-variance model allowing $d_{Innocent\text{-}Guilty}$ and $\sigma_{guilty}$ to differ across the fair and unfair showup conditions (full model). As with the fit to the Yes response data, an unequal-variance model fit significantly better than an equal-variance model. Unlike the fit to the Yes response data, $\sigma_{guilty(fair)}$ differed significantly from $\sigma_{guilty(unfair)}$, so these parameters were not constrained to be equal to each other. Note that it is sensible that $\sigma_{guilty(unfair)}$ was estimated to be smaller than $\sigma_{guilty(fair)}$. The innocent suspects in the unfair showups have more variability than in the fair showups (i.e., $\sigma_{innocent(unfair)} > \sigma_{innocent(fair)}$), because each innocent suspect in the pool for the unfair showups also had a somewhat different distinctive feature (whereas the feature was fixed for the guilty suspect). This variability shows up in the $\sigma_{guilty}$ parameters (i.e., $\sigma_{guilty(unfair)} < \sigma_{guilty(fair)}$) because the $\sigma_{innocent}$ parameter was always fixed to 1 (thus, a change in $\sigma_{innocent}$ will show up as a change in $\sigma_{guilty}$). Also, unlike the fit to the Yes response data, the confidence criteria across the two conditions were not shifted in lockstep (i.e., constraining them to differ by a constant additive factor across conditions significantly worsened the fit), so $c_1$, $c_2$, $c_3$, $c_4$, $c_5$ were also free to vary across the fair and unfair showups.

Because an unequal-variance model fit the data best, the relevant discriminability parameter was not *d'*. Moreover, because $\sigma_{guilty(fair)} \neq \sigma_{guilty(unfair)}$, we also could not use the *d* parameter as a proxy for discriminability, as we did in the analysis of yes responses. We therefore measured discriminability using the standard $d_a$ formula: $(\mu_{guilty} - \mu_{innocent}) / \sqrt{[.5(\sigma^2_{guilty} + \sigma^2_{innocent})]}$. Setting $\mu_{innocent} = 0$ and $\sigma_{innocent} = 1$ by convention, the equation

reduces to $d_a = \mu_{guilty} / \sqrt{[.5(\sigma^2_{guilty} + 1)]}$. For the unconstrained fit, $d_{a\ Innocent\text{-}Guilty}$ was higher in fair showups ($d_{a\ Innocent\text{-}Guilty} = 1.18$) than unfair showups ($d_{a\ Innocent\text{-}Guilty} = 1.05$). Next, to determine if the difference was significant, we fit the same model but we constrained $d_{Innocent\text{-}Guilty}$ and $\sigma_{guilty}$ (i.e., $d_{a\ Innocent\text{-}Guilty}$) to be equal across the fair and unfair showups while still allowing the confidence criteria to differ across conditions. The constrained model provided a significantly worse fit of the data than the full model, $\chi^2\ (2) = 10.52$, $p = .005$. Again, this indicates that $d_{a\ Innocent\text{-}Guilty}$ is significantly better in the fair (block) showup condition than the unfair (do-nothing) showup condition. Thus, whether a signal-detection model is fit to the partial ROC data or to the full ROC data, the prediction made by the diagnostic-feature-detection account was confirmed—people's ability to discriminate between innocent and guilty suspects was better in the fair (block) showup than the unfair (do-nothing) showup. The next logical step would be to develop alternative theoretical interpretations of this interesting finding and empirically pit those theoretical accounts against each other and against the diagnostic-feature-detection account. But, critically, because showups do not contain foils, this finding cannot be attributed to differential filler-siphoning.

Finally, we should note that we also fit more complex versions of the "Yes and No Responses" (full ROC) signal-detection model. These analyses are available at https://osf.io/nr24b/. In short, the fit of the "Yes and No Responses" model was significantly improved when Yes responses and No responses were allowed to differ from each other, suggesting that Yes responses and No responses differed significantly in terms of discriminability (i.e., $d_{Innocent\text{-}Guilty}$ and $\sigma_{guilty}$). This finding corresponds with the pattern of results observed in the ROC analyses (Figure S1). For Yes responses, the data for fair (block) showups fall higher on the ROC plot than the data for unfair (do-nothing) showups. But when No responses are included and full ROC curves are constructed, the data for the fair and unfair showups converge and lie on top of each other. Although we had not anticipated this difference, it is perhaps unsurprising: People who respond Yes and select a person from a lineup (called choosers in previous research) and people who respond No and state that the culprit is "not present" (called non-choosers in previous research) are often analyzed separately and found to differ from each other in important ways in eyewitness ID research (e.g., Sporer, Penrod, Read, & Cutler, 1995). Future research should continue to examine how and why Yes responses and No responses differ on eyewitness identification tasks.

## An Empirical Comparison of the Models: Fitting the BEST Model vs. the INTEGRATION Model to the Colloff et al. (2016) Data

We used the BEST model to analyze our lineup data (Colloff et al., 2016), but Smith et al. (2017) endorsed the INTEGRATION model instead. These two models differ only in the assumed decision rule. The BEST model assumes that after the most familiar face is detected, the identification decision is based on the familiarity of that face alone (e.g., Clark, Erickson, & Breneman, 2011; Macmillan & Creelman, 2005). The INTEGRATION model assumes that after the most familiar face in the lineup is detected, the identification decision is based on the sum of the familiarity values of all of the faces in the lineup. If the summed familiarity variable is high enough, the most familiar face is identified (e.g., Duncan, 2006; Palmer, Brewer, & Weber, 2010). The only way to test which model is more appropriate for the analyses is to empirically compare them. Here, we fit the INTEGRATION model to our (Colloff et al., 2016) fair (replication) and unfair (do-nothing) lineup data to determine: (a) which model—the INTEGRATION model or the BEST model—provided a better fit to the data; and (b) whether the INTEGRATION model led to the same or different conclusions about $d'_{Innocent-Guilty}$ as the BEST model.

We used the same model-fitting procedure that we used when we fit the BEST model in our original paper (see Colloff et al., 2016, Supplemental Materials, p.3). In brief, the model for the fair lineup had 6 parameters ($\mu_{guilty}$, $c_1$, $c_2$, $c_3$, $c_4$, $c_5$) and there were 15 degrees of freedom in the data (5 levels of confidence for: guilty suspect identifications, foil identifications in target-present lineups, and foil/innocent suspect identifications in target-absent lineups). Thus, the fit of the model to the fair lineup data involved 15 – 6 = 9 degrees of freedom. The model for the unfair lineup had 7 parameters ($\mu_{guilty}$, $\mu_{foil}$, $c_1$, $c_2$, $c_3$, $c_4$, $c_5$) and there were 20 degrees of freedom in the data (5 levels of confidence for: guilty suspect identifications, foil identifications in target-present lineups, innocent suspect identifications, and foil identifications in target-absent lineups). Thus, the fit of the model to the unfair data involved 20 – 7 = 13 degrees of freedom. For both the fair and unfair models, we fixed $\mu_{innocent}$ to 0 and set the standard deviations for each distribution to 1, for simplicity. Once the frequencies of positive identifications were entered into the model, the frequencies of reject identifications were fixed. We fit the model by minimizing the chi-square goodness-of-fit statistic.

Table S2 shows the observed identification responses and the identification responses

predicted by the BEST model from Colloff et al. (2016), along with the identification responses predicted by the INTEGRATION model. It is clear that both models are able to capture the basic trends in the Colloff et al. data, but which model provides the best fit? To answer that question, we examine the $\chi^2$, df and *p* goodness-of-fit statistics in Figure 2B in our main reply. They show that, for both fair (replication) and unfair (do-nothing) lineups, the INTEGRATION model offers a worse fit than the BEST model. Although all useful models are simplified approximations to the truth, the basic idea of model comparison is that, all else being equal, the better fitting model is the one that should be used to interpret the data, at least until an even better-fitting model is developed. It is therefore incorrect to assert that the INTEGRATION model is the proper model to use to interpret these data.

Table S2

*Observed Identification Responses and Identification Responses Predicted by the BEST and the INTEGRATION Models in Each Confidence bin in the Fair (Replication) and Unfair (Do-nothing) Conditions of Colloff et al. (2016)*

| | Fair | | | | | Unfair | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Target present | | | Target absent | | Target present | | | Target absent | | |
| Confidence | Guilty Suspect | Foil | Incorrect Rejection | Foil | Correct Rejection | Guilty Suspect | Foil | Incorrect Rejection | Innocent Suspect | Foil | Correct Rejection |
| 0-20 | | | | | | | | | | | |
| Observed | 21.00 | 45.00 | - | 57.00 | - | 17.00 | 32.00 | - | 18.00 | 29.00 | - |
| Best Predicted | 20.99 | 38.17 | - | 64.20 | - | 28.65 | 19.74 | - | 26.35 | 28.04 | - |
| Integration Predicted | 26.46 | 33.21 | - | 64.43 | - | 43.14 | 13.88 | - | 32.43 | 25.46 | - |
| 30-40 | | | | | | | | | | | |
| Observed | 36.00 | 52.00 | - | 99.00 | - | 35.00 | 36.00 | - | 37.00 | 50.00 | - |
| Best Predicted | 33.96 | 58.33 | - | 94.28 | - | 49.22 | 30.15 | - | 42.38 | 41.12 | - |
| Integration Predicted | 40.76 | 51.12 | - | 92.43 | - | 64.48 | 20.71 | - | 46.17 | 36.20 | - |
| 50-60 | | | | | | | | | | | |
| Observed | 96.00 | 127.00 | - | 207.00 | - | 156.00 | 70.00 | - | 113.00 | 69.00 | - |
| Best Predicted | 89.77 | 135.01 | - | 203.00 | - | 148.19 | 68.25 | - | 110.26 | 85.91 | - |
| Integration Predicted | 100.57 | 125.78 | - | 197.99 | - | 160.18 | 51.19 | - | 104.52 | 81.62 | - |
| 70-80 | | | | | | | | | | | |
| Observed | 106.00 | 93.00 | - | 168.00 | - | 155.00 | 44.00 | - | 74.00 | 49.00 | - |
| Best Predicted | 96.58 | 113.68 | - | 156.10 | - | 145.58 | 41.84 | - | 87.71 | 47.48 | - |
| Integration Predicted | 96.08 | 119.52 | - | 152.55 | - | 132.87 | 42.14 | - | 76.74 | 59.59 | - |
| 90-100 | | | | | | | | | | | |
| Observed | 88.00 | 65.00 | - | 96.00 | - | 266.00 | 24.00 | - | 122.00 | 22.00 | - |
| Best Predicted | 94.42 | 68.03 | - | 86.49 | - | 264.29 | 28.69 | - | 106.32 | 29.32 | - |
| Integration Predicted | 74.93 | 93.40 | - | 87.49 | - | 210.48 | 66.83 | - | 99.13 | 77.21 | - |
| Total | | | | | | | | | | | |
| Observed | - | - | 396.00 | - | 513.00 | - | - | 275.00 | - | - | 434.00 |
| Best Predicted | - | - | 376.05 | - | 535.93 | - | - | 285.40 | - | - | 412.11 |
| Integration Predicted | - | - | 363.18 | - | 545.11 | - | - | 304.10 | - | - | 377.94 |

*Note*. The total row displays all reject identification decisions because the models do not account for the confidence level with which lineup rejections are made.

Despite the comparatively poor performance of the INTEGRATION model, we nevertheless considered Smith et al.'s (2017) use of it to interpret the data. We do so to highlight that Smith et al. mistakenly fit a fair lineup model with two-distributions to the unfair do-nothing lineup data which has three-distributions (see section two *'An empirical comparison of Smith et al.'s model versus our model'* in our reply), and to show that, when fit correctly, the INTEGRATION model finds the same result as the BEST model. Again, we fit the INTEGRATION model using the same model-fitting procedure that we used in our original paper when we fit the BEST model (Colloff et al., 2016, Supplemental Materials, p.3). We first fit the model allowing $d'_{Innocent\text{-}Guilty}$ to differ across the fair and unfair lineup conditions (full model), then we fit the same model but we constrained $d'_{Innocent\text{-}Guilty}$ to be equal across the fair and unfair lineups, while allowing the confidence criteria ($c_1$, $c_2$, $c_3$, $c_4$, $c_5$) to differ across conditions (constrained model). We would conclude that the difference in $d'_{Innocent\text{-}Guilty}$ between the fair and unfair lineups is statistically significant, if the constrained model provides a significantly worse fit to the data than the full model.

Table S3 shows the full and constrained model fits for the BEST model from Colloff et al. (2016) and for the INTEGRATION model. First, looking at the best-fitting parameters in the full model columns, it is clear that both models estimate that $d'_{Innocent\text{-}Guilty}$ is higher in fair lineups than in unfair lineups. Fitting the BEST model in Colloff et al. showed that the constrained model provided a significantly worse fit of the fair (replication) and unfair (do-nothing) data than the full model, $\chi^2$ (1) = 24.32, $p < .001$. We found the exact same pattern of results using the INTEGRATION model: The constrained model provided a significantly worse fit of the data than the full model, $\chi^2$ (1) = 17.41, $p < .001$. In short, both the BEST model and the INTEGRATION model tell the same story: fair lineups make it significantly easier for people to discriminate between innocent and guilty suspects than unfair lineups. These results are consistent with the diagnostic-feature-detection interpretation (Wixted & Mickes, 2014), and, notably, the findings from our new showup experiment.

Finally, note that, in one sense, memory performance in unfair lineups is better in that, by design, it is easier to tell the difference between the foils (with no distinctive feature) and the guilty suspect (who has the distinctive feature) compared to the fair lineup condition. In other words, $d'_{Foil\text{-}Guilty}$, should be larger in unfair lineups compared to fair lineups. Indeed, Figure 2B in our reply shows that, according to both the BEST model and the INTEGRATION model, $d'_{Foil\text{-}Guilty}$ in the unfair lineups is much larger than it is in the fair lineups. This is logical. Making the foils less similar to the suspect as we move from fair to

unfair lineups—that is, the experimental manipulation itself—increases discriminability between foils and suspects. But measuring how the foils differ from the suspects only serves as a manipulation check; it does not tell us anything about the theoretically interesting measure (i.e., $d'_{Innocent\text{-}Guilty}$) which we tested in our original paper. This is because, in unfair lineups, it is also easier to tell the difference between the foils (with no distinctive feature) and the *innocent* suspect (who has the distinctive feature) compared to the fair lineup condition. Therefore, as shown in Figure 2D of our reply, but contrary to how Smith et al. (2017) modeled the data, when the lineup is unfair, the innocent suspect distribution (which is needed to measure $d'_{Innocent\text{-}Guilty}$) should be estimated separately from the foil distribution (which has moved further away as it is the experimental manipulation). Smith et al. modelled the data by combining innocent suspect and foil IDs into one distribution to calculate an "omnibus" measure of memory performance in unfair lineups. This confounds the measure required for testing the diagnostic-feature-detection theory ($d'_{Innocent\text{-}Guilty}$), with the experimental manipulation ($d'_{Foil\text{-}Guilty}$).

Table S3

*Full and Constrained ($d'_{Innocent\text{-}Guilty}$) Model Fits Using the BEST Model and the INTEGRATION Model for the Fair (Replication) vs. Unfair (Do-nothing) Comparison from Colloff et al. (2016)*

| | BEST | | | | INTEGRATION | | | |
|---|---|---|---|---|---|---|---|---|
| | Full model | | Constrained model | | Full model | | Constrained model | |
| Estimate | Fair | Unfair | Fair | Unfair | Fair | Unfair | Fair | Unfair |
| $d'_{Innocent\text{-}Guilty}$ | 0.86 | 0.54 | 0.73 | 0.73 | 0.99 | 0.66 | 0.86 | 0.86 |
| $d'_{Foil\text{-}Guilty}$ | 0.86 | 1.79 | 0.73 | 1.88 | 0.99 | 2.00 | 0.86 | 2.08 |
| $c_1$ | 1.18 | 0.22 | 1.16 | 0.32 | –0.17 | –0.83 | –0.24 | –0.74 |
| $c_2$ | 1.27 | 0.31 | 1.25 | 0.41 | 0.18 | –0.47 | 0.10 | –0.37 |
| $c_3$ | 1.41 | 0.46 | 1.39 | 0.56 | 0.69 | 0.03 | 0.62 | 0.13 |
| $c_4$ | 1.76 | 0.85 | 1.74 | 0.95 | 1.97 | 1.22 | 1.88 | 1.33 |
| $c_5$ | 2.22 | 1.25 | 2.20 | 1.35 | 3.54 | 2.31 | 3.46 | 2.43 |
| Overall $\chi^2$ | 49.41 | | 73.73 | | 224.52 | | 241.93 | |
| Overall df | 22 | | 23 | | 22 | | 23 | |
| Overall *p* | <.001 | | <.001 | | <.001 | | <.001 | |

*Note.* The full model allows $d'_{Innocent\text{-}Guilty}$ to differ between the fair vs. unfair lineup comparison. The constrained model holds $d'_{Innocent\text{-}Guilty}$ constant across the fair vs. unfair comparison. A lower chi-square indicates a better fit. For both the BEST model and the INTEGRATION model, the overall $\chi^2$, df and *p* goodness-of-fit statistics show that the fit was worse in the constrained model compared to the full model.

## References

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

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal- detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging, 32*, 243–258. doi:10.1037/pag0000168

R Development Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.0) [Computer software]. Retrieved from https://www.r-project.org/index.html

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77–84. doi:10.1186/1471-2105-12-77

RStudio Team. (2015). RStudio: Integrated development for R (Version 0.98.1103) [Computer software]. Retrieved from http://www.rstudio.com/

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.*

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

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