Supplemental Materials

In the following supplemental materials we provide a more elaborative explanation of the compound-decision signal-detection model (SDT-CD, Duncan, 2006; Palmer, Brewer, & Weber, 2010) and how we used this model to fit high-similarity and low-similarity lineup data from Colloff, Wade, and Strange (2016). The results demonstrate that low-similarity lineups produced superior memory performance. We then run several supplementary simulations, including a replication of Colloff et al.'s simulation, and demonstrate that all models converge on the finding that memory performance is superior with low similarity fillers than with high-similarity fillers.

Signal Detection Theory Compound-Decisions (SDT-CD)

The most common signal-detection task is the simple-detection task. In a simple-detection task, participants are presented with a single stimulus that is either a lure (e.g., innocent suspect) or a target (e.g., culprit). The task of the participant is to determine whether the presented stimulus is from the lure or target distribution. A common eyewitness identification procedure called a showup is a good example of a simple-detection task. In a showup, police present a single suspect – who might or might not be the culprit – to the eyewitness and ask the eyewitness if this is the individual who committed the crime. If the individual is the culprit, the eyewitness should respond affirmatively and if the individual is not the culprit, the eyewitness should respond negatively. The ability of an eyewitness to correctly categorize the suspect as the culprit or innocent suspect is referred to as discriminability and this is often quantified by plotting the empirical Receiver Operating Characteristic (ROC) curve or by calculating its theoretical stand-in, the discriminability index (*d'*, see Macmillan & Creelman, 2005).

Very simply, *d'* represents the distance, in standard units, between the mean of a standard normal distribution representing the innocent suspect and the mean of a standard normal distribution representing the culprit.

Because the ROC analysis reported by Colloff et al. (2016) only considered suspect identifications and ignored filler identifications, it effectively treated the lineup identification procedure as a simple-detection task, i.e., as if only one individual was presented in each identification procedure (who was either the culprit or an innocent suspect). Likewise, because Colloff et al.'s analysis of their simulated data only considered suspect distributions and ignored the filler distributions, it too treated the lineup identification task as a simple-detection task. But, lineups are not simple-detection tasks. Six-person lineups, like those examined by Colloff et al., present participants with a compound-decision task that has two components: 1) a present/absent detection task, and 2) a 6-Alternative-Forced-Choice (*m*-AFC) identification task (Duncan, 2006). Detection in a six-person lineup is measured by comparing the frequency with which participants make affirmative responses (suspect identifications or filler identifications) when the culprit is present to the frequency with which participants make affirmative responses (suspect identifications or filler identifications) when the culprit is absent. To the extent that affirmative responses are more probable when the culprit is present, detection performance is good. The identification component of the task is concerned with the proportion of affirmative responses in the culprit-present lineup that land on the culprit. To the extent that a high proportion of affirmative responses in the culprit-present lineup land on the culprit, identification is good.

SDT-CD assumes that a single discriminability parameter is common to both the detection and identification components of lineup procedures (Duncan, 2006). Because the identification component of a lineup task is relevant only to eyewitnesses who make an identification from culprit-present lineups, estimations of the decision criterion (the tendency for eyewitnesses to respond affirmatively) are based solely on performance in the detection component of the lineup task.

There are two forms of the SDT-CD model based on different decision rules that eyewitnesses might adopt. The *independent observation* rule assumes that eyewitnesses evaluate each lineup member individually and make an identification if one or more lineup members exceed the decision criterion. The *integration* rule assumes that eyewitnesses make a more global assessment of the lineup and make an identification if the sum of the match-to-memory values exceeds the decision criterion. Past research uniformly finds that the integration rule provides better fits to lineup data, therefore, we use the integration rule in the present research (e.g., Duncan, 2006; Palmer et al., 2010).

Fitting the SDT-CD Model to Colloff et al.'s (2016) High- and Low-Similarity Lineups

We fit the SDT-CD model to Colloff et al.'s (2016) high-similarity ("replication") lineup and to their low-similarity ("do-nothing") lineup. Following the simulations presented by Colloff et al., we assumed that there was a set of five decision criteria reflecting different levels of confidence by combining confidence ratings of 0-20 (c_1), 30-40 (c_2), 50-60 (c_3), 70-80 (c_4), and 90-100 (c_5). Consistent with the SDT-CD model, we assumed that a single discriminability parameter (d') guided both detection and identification decisions. Thus, in total, our model had six parameters (d', c_1, c_2, c_3, c_4, c_5).

The culprit-absent lineups had five degrees of freedom because there were five levels of confidence for false-positive responses and the culprit-present lineups had 10 degrees of freedom because there were five levels of confidence for culprit identifications and five levels of confidence for false-positive filler identifications. Accordingly, both lineups had 15 degrees of freedom in total and both lineups had six free parameters in total; thus, the fitting of the model to the lineups was evaluated on nine degrees of freedom (15 – 6 = 9). We determined optimal parameters by using maximum likelihood estimation to minimize the discrepancy function (the negative log likelihood). We assessed goodness-of-fit using the G^2 statistic, which we evaluated on a chi-square distribution.

We present the observed data reported in Colloff et al. (2016) and the model-predicted data in Table S1. It is clear from comparing the predicted and observed values that the SDT-CD model captured the overall trends in the data. The results of the model-fitted data were quite striking. Memory performance in the low-similarity lineup (*d'* = 1.81) was almost twice as good as memory performance in the high-similarity lineup (*d'* = .96). This is also evident from the descriptive analysis we presented in the body of our article: high- and low-similarity culprit-absent lineups resulted in similar false-positive rates (55% vs. 58%), but the low-similarity lineup had superior detection (75% culprit-present affirmative - 58% culprit-absent affirmative = 17%) compared to the high-similarity lineup (65% culprit-present affirmative – 55% culprit-absent affirmative = 10%); and, the low-similarity lineup also produced superior identification performance (75.3% of culprit-present identifications were correct) compared to the high-similarity lineup (47.6% of culprit-present identifications were correct).

Table S1: *Observed and SDT-CD Predicted Identification Percentages in each Confidence bin from the High-Similarity ("Replication") and Low-Similarity ("Do Nothing") Lineups*

Panel A: High-Similarity Fillers

	Confidence					
	90-100	70-80	50-60	30-40	0-20	Reject
TA False Affirmatives						
Observed	8.42	14.74	18.16	8.68	5.00	45.00
Predicted	7.88	13.81	17.55	8.06	5.41	47.29
TP Filler Identifications						
Observed	5.78	8.27	11.29	4.62	4.00	35.20
Predicted	8.64	10.93	11.22	4.49	2.82	32.29
Culprit Identifications						
Observed	7.82	9.42	8.53	3.20	1.87	
Predicted	6.71	8.49	8.71	3.49	2.19	
Best-Fitting Parameters	c_5	c_4	c_3	c_2	c_1	d'
	1.41	0.78	0.27	0.07	-0.07	0.96
Model-Fit Statistics	$G^2(9) = 16.59$, $p = .06$					

Panel B: Low-Similarity Fillers

	Confidence					
	90-100	70-80	50-60	30-40	0-20	Reject
TA False Affirmatives						
Observed	14.16	12.09	17.90	8.55	4.62	42.67
Predicted	10.56	11.27	17.19	6.83	4.24	50.00
TP Filler Identifications						
Observed	2.16	3.96	6.31	3.24	2.88	24.77
Predicted	8.50	5.02	5.37	1.68	0.94	23.02
Culprit Identifications						
Observed	23.96	13.96	14.05	3.15	1.53	
Predicted	21.93	12.96	13.84	4.33	2.42	
Best-Fitting Parameters	c_5	c_4	c_3	c_2	c_1	d'
	1.25	0.78	0.28	0.11	0.00	1.81
Model-Fit Statistics	$G^2(9) = 52.69$, $p < .001$					

It should be noted that neither the equal-variance SDT model that Colloff et al. (2016) fit, nor the SDT-CD model that we fit provided an adequate account of the low-similarity lineup data. The SDT-CD model assumes that fillers and innocent suspects are equally similar to the culprit. But, with low-similarity lineups, the innocent suspect is

more similar to the culprit than are the fillers. As a result, a disproportionate number of false positives in the culprit-absent lineup are of the innocent suspect. Therefore, SDT-CD has difficulty determining a single *d'* parameter that is optimal for both detection and identification. This difficulty arises because the large number of false positives that land on the innocent suspect in the culprit-absent lineup decrease detection performance, but the innocent suspect is not in the culprit-present lineup to reduce identification performance. As a result, with low-similarity lineups, identification performance will exceed detection performance and SDT-CD will have difficulty determining a single *d'* value that is common to both detection and identification. To address this issue, we ran another simulation in which we first fit a detection model to the high-similarity and low-similarity lineups and then, using the best-fitting decision criteria determined by the detection model, fit an identification model to the data. This method improved the goodness-of-fit for both the high-similarity $G^2(9) = 11.16$, $p = .26$, and the low-similarity lineups $G^2(9) = 35.22$, $p < .001$. The model still did not provide an adequate fit to the low-similarity lineup, but the fit was comparable to that achieved by Colloff et al. ($\chi^2(13)=36.10$, $p < .001$). Most importantly, this method converged with the standard SDT-CD model in demonstrating that memory performance was better for the low-similarity lineup (*d'* = 1.56) than for the high-similarity lineup (*d'* = .85).

Low-Similarity Fillers Produced Better Memory Performance than High-Similarity Fillers in Colloff et al.'s Simulations

The SDT-CD model demonstrates that memory performance was superior for low-similarity fillers. Here, we show that Colloff et al.'s simulations actually converge with the SDT-CD model in demonstrating that memory performance for low-similarity

lineups was superior to memory performance for high-similarity lineups. Colloff et al.'s simulation correctly *classified* filler identifications, but the authors relied on a simple detection analysis to analyze those simulated data and the result is that the analysis functionally *misclassified* filler identifications as rejections in exactly the same manner that the analysis of the empirical data did. A simple detection analysis is no more valid for analyzing memory performance from simulated data than it is for analyzing memory performance from empirical data. When these data are analyzed with a more appropriate SDT-CD model, they too show that memory performance was superior for low-similarity lineups than for high-similarity lineups.

We used the same model-fitting strategy from above to fit the SDT-CD model to the simulated data presented by Colloff et al. (2016). The results converged with our fitting of the SDT-CD model to their observed data: the low-similarity fillers (d' = 1.70) produced better, not worse, memory performance than the high-similarity fillers (d' = 0.97). It is clear that high-similarity fillers did not improve memory performance in either the observed or simulated data. Exactly as signal detection theory predicts, the use of high-similarity fillers decreased memory performance.

Table S2: *Colloff et al.'s (2016) Predicted Identification Percentages and Best-Fitting SDT-CD Predicted Identification Percentages in Each Confidence bin from the High-Similarity ("Replication") and Low-Similarity ("Do Nothing") Lineups*

Panel A: High-Similarity Fillers

	Confidence					
	90-100	70-80	50-60	30-40	0-20	Reject
TA False Affirmatives						
"Observed"	7.59	13.69	17.81	8.27	5.63	47.01
Predicted	7.96	13.52	17.32	8.00	5.44	47.76
TP Filler Identifications						
"Observed"	6.05	10.10	12.00	5.18	3.39	33.43
Predicted	8.71	10.68	11.05	4.45	2.83	32.56
Culprit Identifications						
"Observed"	8.39	8.58	7.98	3.02	1.87	
Predicted	6.87	8.42	8.71	3.51	2.23	
Best-Fitting Parameters	c_5	c_4	c_3	c_2	c_1	d'
	1.41	0.79	0.28	0.08	-0.06	0.97
Model-Fit Statistics	$G^2(9) = 8.04$, $p = .53$					

Panel B: Low-Similarity Fillers

	Confidence					
	90-100	70-80	50-60	30-40	0-20	Reject
TA False Affirmatives						
"Observed"	13.34	13.29	19.29	8.21	5.35	40.52
Predicted	11.60	12.01	18.09	7.98	5.30	45.03
TP Filler Identifications						
"Observed"	2.58	3.77	6.15	2.72	1.78	25.71
Predicted	9.57	5.64	6.07	2.10	1.24	20.60
Culprit Identifications						
"Observed"	23.81	13.12	13.35	4.43	2.58	
Predicted	21.29	12.56	13.51	4.66	2.76	
Best-Fitting Parameters	c_5	c_4	c_3	c_2	c_1	d'
	1.20	0.72	0.21	0.01	-0.13	1.70
Model-Fit Statistics	$G^2(9) = 38.06$, $p < .001$					

In fact, one does not even need the SDT-CD model to demonstrate that Colloff et al.'s low-similarity lineups produced better memory performance than did their high-similarity lineups. When arguing that memory performance was superior for high-similarity lineups than low-similarity lineups, Colloff et al. focused on the fact that the

difference between culprit and innocent suspect distributions was greater for the high-similarity lineup (*d'* = .86) than for the low-similarity lineup (*d'* = .54). But as we have already explained in the body of our manuscript, *d'* on suspect identifications can vary because of either a change in rejections or because of a change in filler identifications. Because of this, one cannot simply look at the distance between culprit and innocent distributions and claim to be measuring memory performance. Rather, one needs to look at the average distance between the culprit distribution and all innocent distributions including both the innocent suspect and fillers. In the high-similarity lineup, all fillers were assumed to be equally similar to the culprit; the result is that the culprit was equally distant (*d'* = .86) from all six innocent distributions and the value that summarized memory performance for the high-similarity lineup was *d'* = .86. In the low-similarity lineup, the distance between the culprit and innocent suspect distributions was *d'* = .54; but, on average, the difference between the innocent suspect distributions and each of the five filler distributions was *d'* = 1.25. The result is that the distance between the culprit and filler distributions was, on average, *d'* = 1.79. To estimate memory performance for the low-similarity lineup, one must find the average distance between the mean of the culprit distribution and the means of each of the six innocent-person distributions ([1.79 × 5 + .54]/6), *d'* = 1.58. The results converge with the SDT-CD model in demonstrating that memory performance was better in the low-similarity lineup than in the high-similarity lineup.

Conclusion

If Colloff et al. were correct and the distance between the culprit and innocent suspect distributions was larger for high-similarity lineups than for low-similarity lineups

because high-similarity lineups increased memory performance, then this would be evidenced by an increase in culprit identifications, an increase in correct rejections, or both. As we demonstrated in the body of our manuscript, high-similarity lineups resulted in fewer culprit identifications and no more correct rejections than did the low-similarity lineups. So, we know that high-similarity lineups could not have improved memory performance; that is not why the distance between culprit and innocent distributions (i.e., suspect-identification value) was larger for high-similarity lineups than for low-similarity lineups. Rather, suspect-identification value was superior for high-similarity lineups because high-similarity fillers decreased innocent suspect identifications to a greater extent than they decreased culprit identifications. As we demonstrated in these supplemental materials, when the data are analyzed with a model that accounts for all eyewitness behaviors, the longstanding finding that memory performance is inversely related to lure-target similarity replicates.

References

Colloff, M., Wade, K. A., & Strange, D. (2016). Unfair lineups make eyewitnesses more likely to confuse innocent and guilty suspects. *Psychological Science,* Advance Online Publication. doi: 10.1177/0956797616655789

Duncan, M. J. (2006). *A signal detection model of compound decision tasks* (Tech. Rep. Np. No. TR2006-256). Toronto, ON: Defence Research and Development Canada.

Macmillan, N. A. & Creelman, C. D. (2005). *Detection Theory: A User's Guide.* (2nd Ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

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