Supplemental Materials for Smith, Ying, Goldstein, & Fitzgerald (2024)

The purpose of this supplemental materials document is twofold. First, we present a series of simulations examining the impact of correlated memory signals on the predictions of the absolute (MAX) and relative (BEST-REST) judgment models. Second, we fit the relative model to the data from both Experiments 1 and 2 and assessed whether its conclusions converged with those of the absolute model.

Assessing the Impact of Correlated Memory Signals on the Predictions of the Absolute and Relative Models

There is an emerging consensus in the eyewitness identification literature that the memory signals emanating from members of the same lineup should be correlated (e.g., Akan et al., 2021; Smith et al., 2022; Wixted et al., 2018). In other words, when one lineup member provides a strong match to the witness' memory for the culprit, other members of that same lineup should also tend to provide a strong match. Likewise, when one lineup member provides a weak match, the other lineup members should also tend to provide a weak match. For simplicity, the model predictions that we generated in the main body of our paper assumed that memory signals were independent or uncorrelated ($\rho = 0$). We made that simplifying assumption because the qualitative predictions of both the absolute model and the relative model were consistent across large variations in the degree of correlation among memory signals. As we will demonstrate below, the absolute model consistently predicts more rejections of low-similarity culprit-absent lineups than high-similarity culprit-absent lineups and the relative model consistently predicts no difference in rejection rates. For fair versus biased culprit-absent lineups, the absolute model consistently predicts more rejections from biased than fair lineups and the relative model consistently predicts more rejections from fair lineups than biased lineups. For the

culprit-present condition, the absolute model predicts either a slight increase in rejections from biased lineups or no change in rejection rates and the relative model consistently predicts more rejections of fair lineups.

For the analyses that we present below, we used R (R Core Team, 2021), RStudio (RStudio Team, 2022), Tidyverse (Wickham, 2019), faux (DeBruine, 2023), and grid (Murrell, 2005).

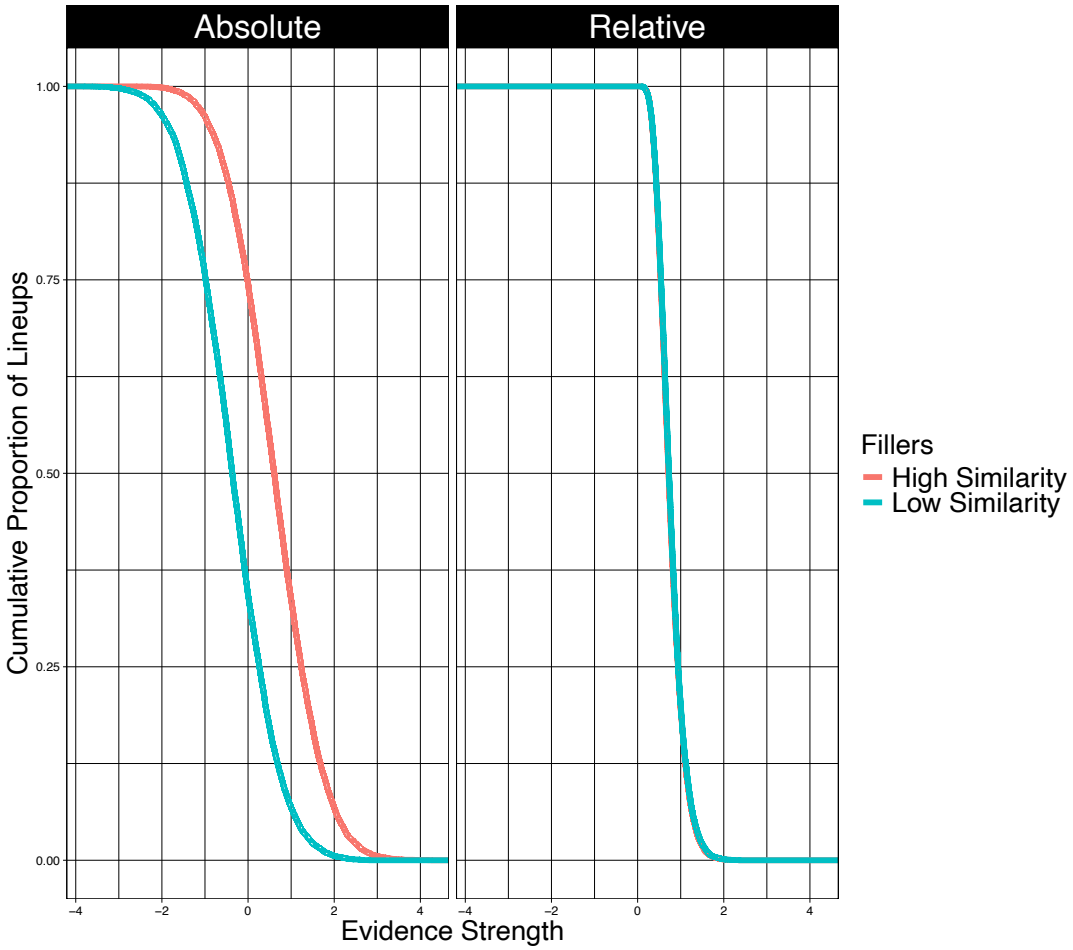
Rejection Rates for Low-Similarity Versus High-Similarity Culprit-Absent Lineups Across Various Levels of Signal Correlation (Experiment 1 Predictions). We examined what impact assuming correlated memory signals among lineup members had on the predictions of both the absolute (MAX) and relative (BEST-REST) models. We started with predictions for rejection rates from low-similarity versus high-similarity culprit-absent lineups (Experiment 1). As with the predictions that we present in the main body of this article, predictions were derived from simulations (N = 10,000) of six-person low-similarity and high-similarity culprit-absent lineups under the assumptions of absolute (MAX) and relative (BEST-REST) decision rules. Low-similarity lineups were comprised of six random draws from the low-similarity filler distribution [$X \sim N(\mu = -1, \sigma = 1)$] and high-similarity lineups were comprised of six random draws from the high-similarity filler distribution [$X \sim N(\mu = 0, \sigma = 1)$]. Critically, we designed our experiments so that relative similarity within culprit-absent lineups was constant and so that only absolute signal strength varied. We did this by creating pairs of targets (A-A', B-B',…, H-H'), selecting relatively high-similarity fillers for each target (e.g., A fillers for target A and A' fillers for target A') and then manipulating whether witness-participants viewed lineups with the high-similarity fillers (e.g., A fillers for target A) or the low-similarity fillers (e.g., A' fillers for target A). Hence, low-similarity fillers were as similar to other low-similarity fillers as high-

similarity fillers were to other high-similarity fillers and the filler signals were equally correlated for both high- and low-similarity culprit-absent lineups. Although filler signals were equally correlated in low-similarity and high-similarity lineups, the high-similarity fillers should have been more strongly correlated with the culprit than the low-similarity fillers (e.g., Shen et al., 2023; Smith et al., 2022). We examined whether this had any impact on model predictions.

Our simulations included two correlation parameters: ρ_1 and ρ_2. The correlation among culprit-absent lineup fillers was governed by ρ_1. For high-similarity fillers, ρ_1 also governed the correlation between the fillers and the culprit. For low-similarity fillers, ρ_2 governed the degree of correlation between the fillers and the culprit and we constrained ρ_2 so that it was equal to or less than ρ_1. We systematically varied both correlation parameters over a wide range of values from .00 to .75. Across all simulations, the absolute model consistently predicted more rejections from low-similarity culprit-absent lineups than from high-similarity culprit-absent lineups and the relative model consistently predicted no change in rejection rates. Figure S1 displays the predictions of the absolute (MAX) and relative (BEST-REST) models for the following parameter settings: low-similarity filler distribution [$X \sim N(\mu = -1, \sigma = 1, \rho_1 = .75, \rho_2 = .00)$] and high-similarity filler distribution [$X \sim N(\mu = 0, \sigma = 1, \rho_1 = .75)$].

Although the above simulations did not generate predictions directly from the model likelihood functions, we verified that the likelihood functions generated the same predictions. Finally, we also considered what the models would have predicted if the signals of high-similarity fillers were more strongly correlated with one another than were the signals of low-similarity fillers. In that case, the absolute model continued to predict more rejections of low-similarity lineups than high-similarity lineups, but the relative model predicted more rejections of high-similarity lineups than low-similarity lineups.

Figure S1: ***Evidence Strength Distributions Predicted by the Absolute and Relative Models on High-Similarity and Low-Similarity Culprit-Absent Lineups***



Rejection Rates for Fair Versus Biased Lineups Across Various Levels of Signal Correlation (Experiment 2 Predictions). For fair and biased culprit-present and culprit-absent lineups (Experiment 2) we also used two correlation parameters. The first parameter specified the correlations between (1) fair fillers and the culprit, (2) fair fillers and the biased innocent-suspect, and (3) fair fillers with other fair fillers. The second parameter specified the correlations between (1) the biased fillers and the culprit, (2) the biased fillers and the innocent suspect, and (3) the biased fillers with other biased fillers. For brevity, we refer to these as memory signal correlations for fair and biased lineups, respectively. We varied both correlation parameters from

.00 to .75 in increments of .25, but with the added constraint that the correlation for biased fillers could not exceed the correlation for fair fillers, which makes sense because as fillers become more similar to the culprit, the signals should become increasingly correlated (Shen et al., 2023; Smith et al., 2022; Wixted et al., 2018).

As with the predictions that we present in the main body of this article, predictions were derived from simulations (N = 10,000) of six-person fair and biased culprit-absent and culprit-present lineups under the assumptions of absolute (MAX) and relative (BEST-REST) decision rules. Fair culprit-absent lineups were comprised of six random draws from the fair filler distribution [$X \sim N(\mu = 0, \sigma = 1)$] and biased culprit-absent lineups were comprised of one draw from the fair filler distribution and five draws from the biased filler distribution [$X \sim N(\mu = -1, \sigma = 1)$]. For the culprit-present lineups depicted in this supplementary materials document, we assumed that the culprit was drawn from a normal distribution with a mean of 1.5 and variance of 1: [$X \sim N(\mu = 1.5, \sigma = 1)$]. We also considered several other parameter settings and consistently found the same qualitative patterns of results that we report here. Fair culprit-present lineups were comprised of one random draw from the culprit distribution and five random draws from the fair filler distribution and biased culprit-present lineups were comprised of one random draw from the culprit distribution and five random draws from the biased filler distribution.

Figure S2 shows the predictions of the absolute model for culprit-absent lineups. The absolute model consistently predicts more rejections from biased culprit-absent lineups than from fair culprit-absent lineups. This is evidenced by the fact that the biased evidence distribution is shifted to the left of the fair evidence distribution, meaning that the absolute model

predicts that witnesses would see less evidence for making an affirmative identification on a biased lineup compared to a fair lineup.

There is one exception to this prediction. In the extreme case where the memory signal correlations in fair lineups are extremely high ($\rho_1 = .75$) and the memory signals in biased lineups are uncorrelated ($\rho_1 = .00$), whether the absolute model predicts more rejections from fair or biased lineups depends on the placement of the witness' decision criterion (see upper righthand corner of Figure S2). This is evident from the fact that the evidence distributions cross-over. Typically, the absolute model predicts more rejections from biased lineups than from fair lineups, because the strength of the MAX signal should tend to be weaker on a biased lineup than on a fair lineup. But in the extreme case where there is almost no variability among members of the same fair culprit-absent lineup—which is what $\rho_1 = .75$ implies—there will be instances where all the fair lineup members provide an extremely weak match to memory. Conversely, because the biased lineup members are uncorrelated, typically there will be at least one that does not provide an *extremely* weak match to memory and whom cannot be rejected with the same confidence as the fair lineup members. Hence, in this extreme case of almost no variability among fair lineup members and lots of variability among biased lineup members, the absolute model predicts that whether fair or biased lineups lead to more rejections depends on criterion placement. But we want to reiterate that this is a rather extreme example and that, in all other situations the absolute model predicts more rejections from biased culprit-absent lineups compared to fair culprit-absent lineups.

Figure S2: ***Evidence Strength Distributions Predicted by the Absolute Model on Fair and Biased Culprit-Absent Lineups***

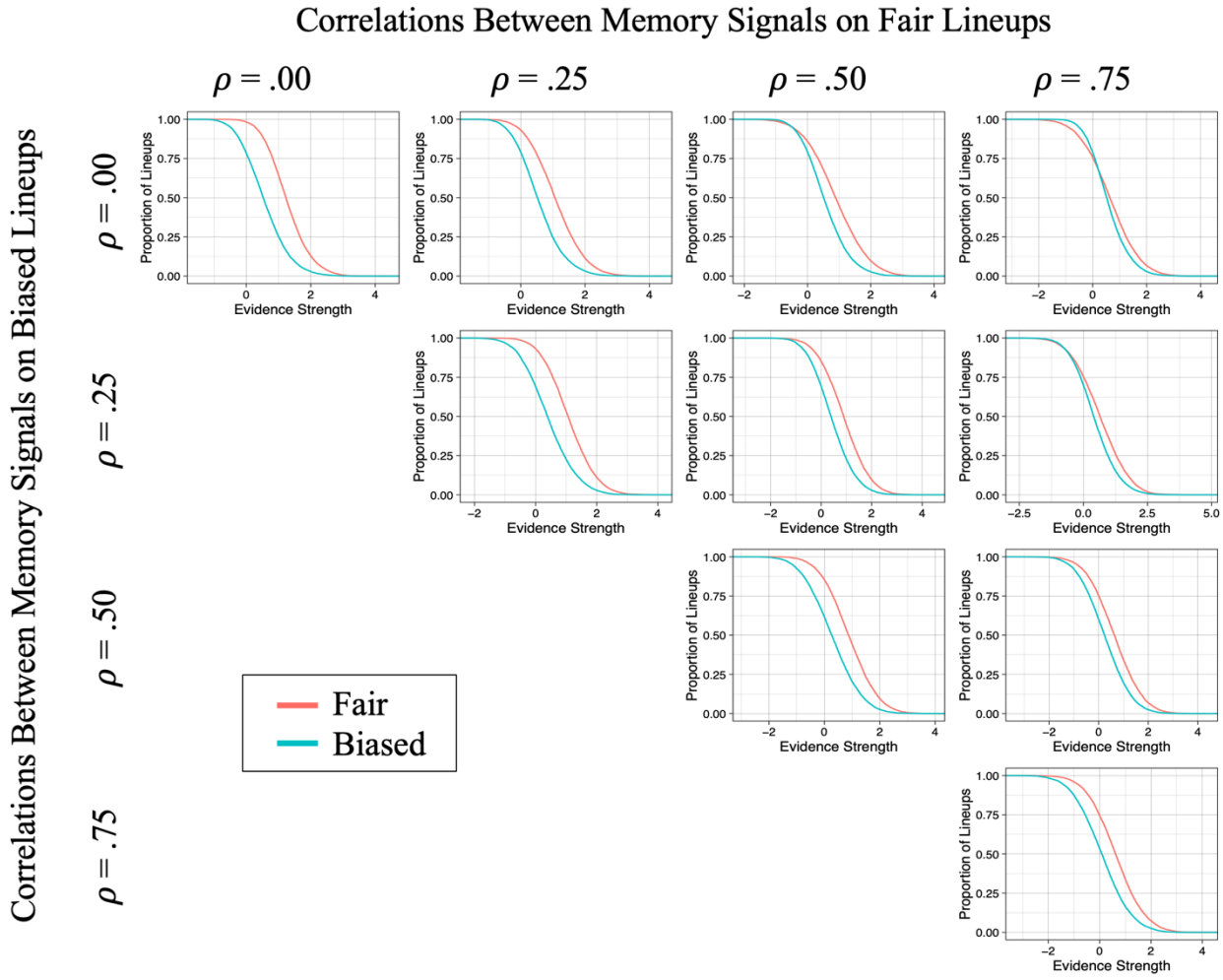


Figure S3 shows the predictions of the absolute model for culprit-present lineups. The absolute model consistently predicts either a slight increase in rejections of biased culprit-absent lineups compared to biased culprit-present lineups or no change in rejection rates.

Figure S3: ***Evidence Strength Distributions Predicted by the Absolute Rule on Fair and Biased Culprit-Present Lineups***

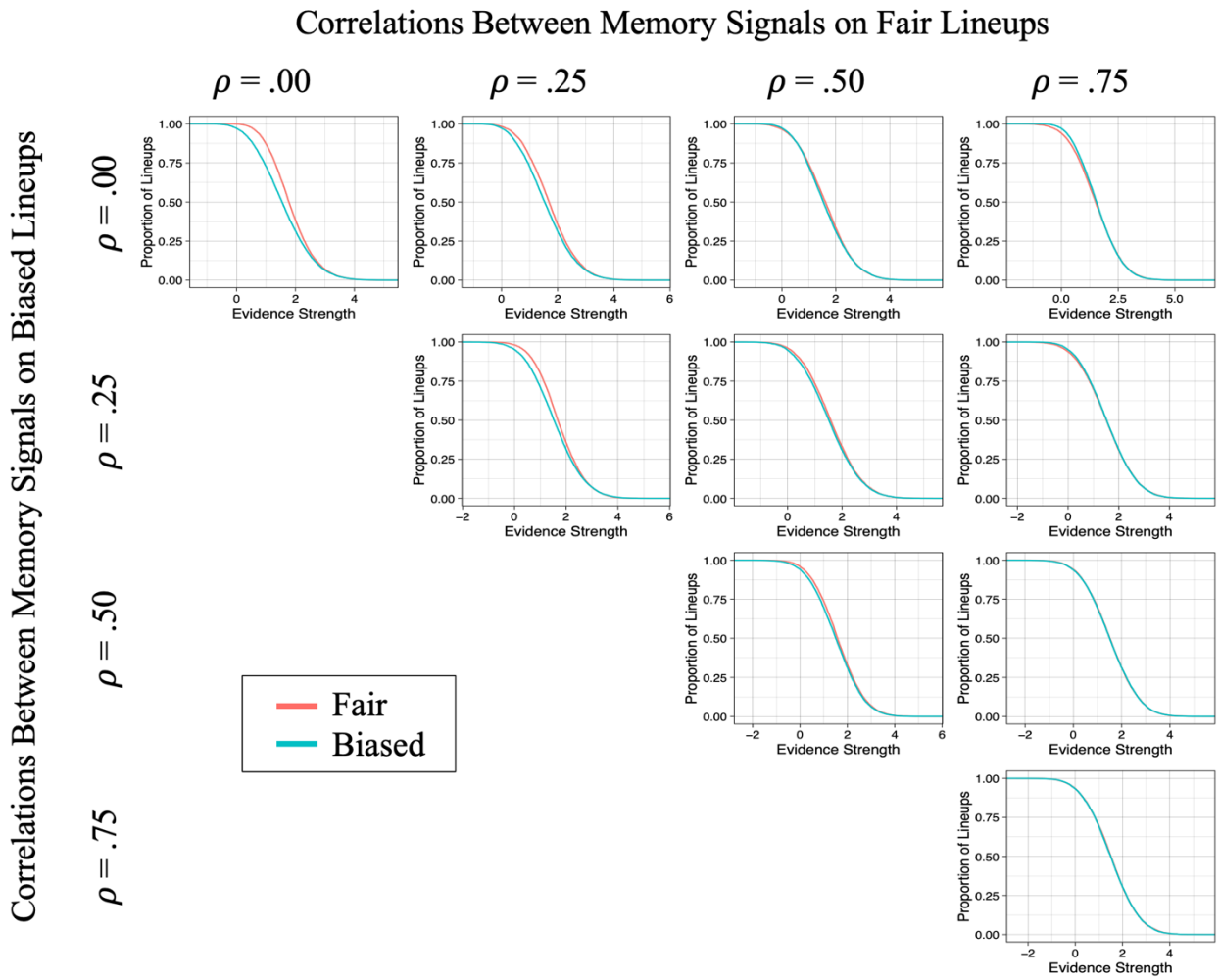


Figure S4 and Figure S5 show the predictions of the relative model for culprit-absent and culprit-present lineups, respectively. For both culprit-absent and culprit-present lineups, the relative model consistently predicts more rejections of fair lineups than biased lineups. This is evident from the fact that the fair distribution is shifted to the left of the biased distribution, meaning that the relative model predicts that there would be less evidence for a witness to make an affirmative identification from a fair lineup compared to a biased lineup. It is also noteworthy that if you compare the magnitude of the predicted differences in Figures S4 and S5, the relative model predicts a larger difference in rejection rates for culprit-present conditions than for culprit-

absent conditions, which is the complete opposite of what the absolute model predicts and the complete opposite of what the empirical data show.

Figure S4: ***Evidence Strength Distributions Predicted by the Relative Model on Fair and Biased Culprit-Absent Lineups***

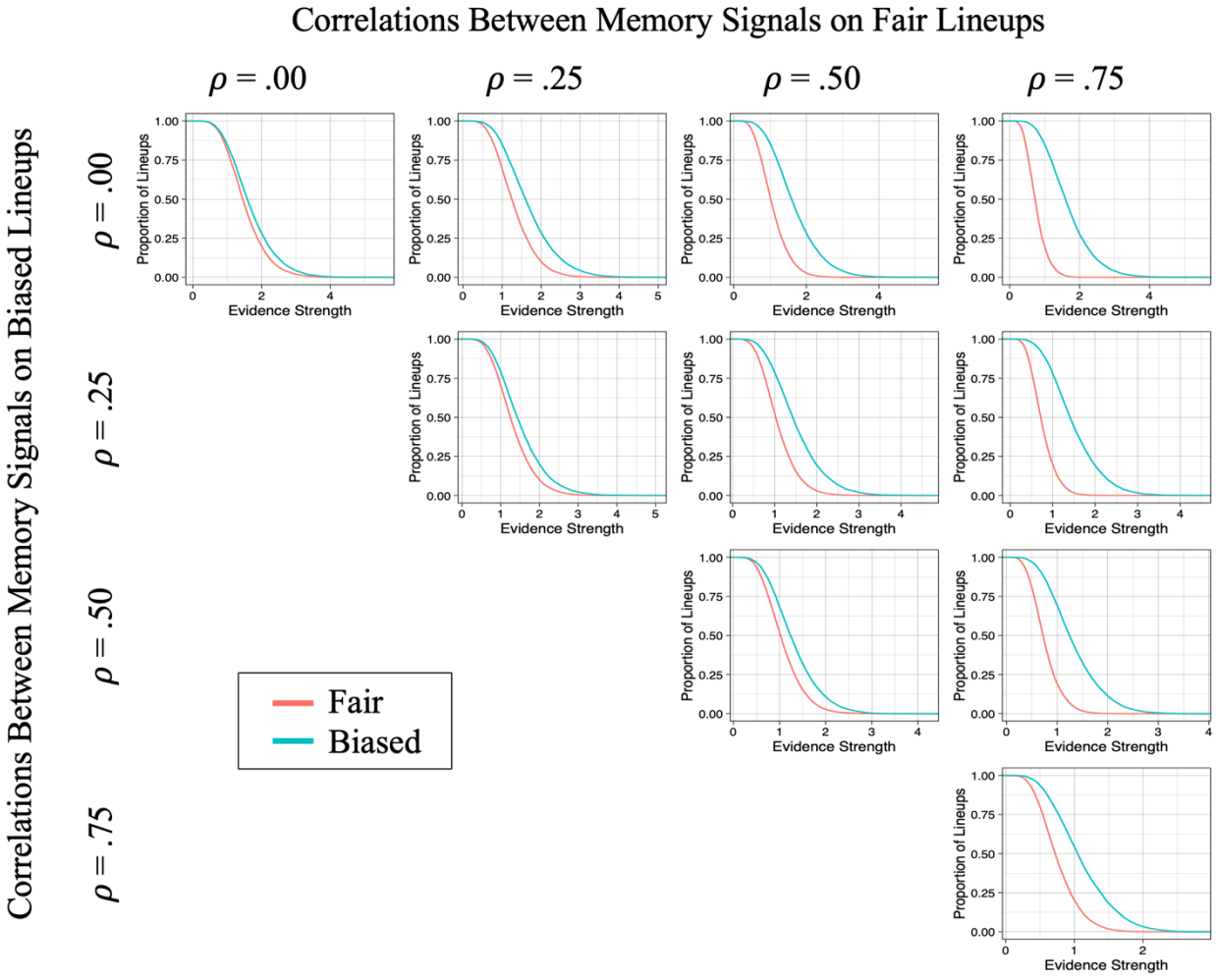
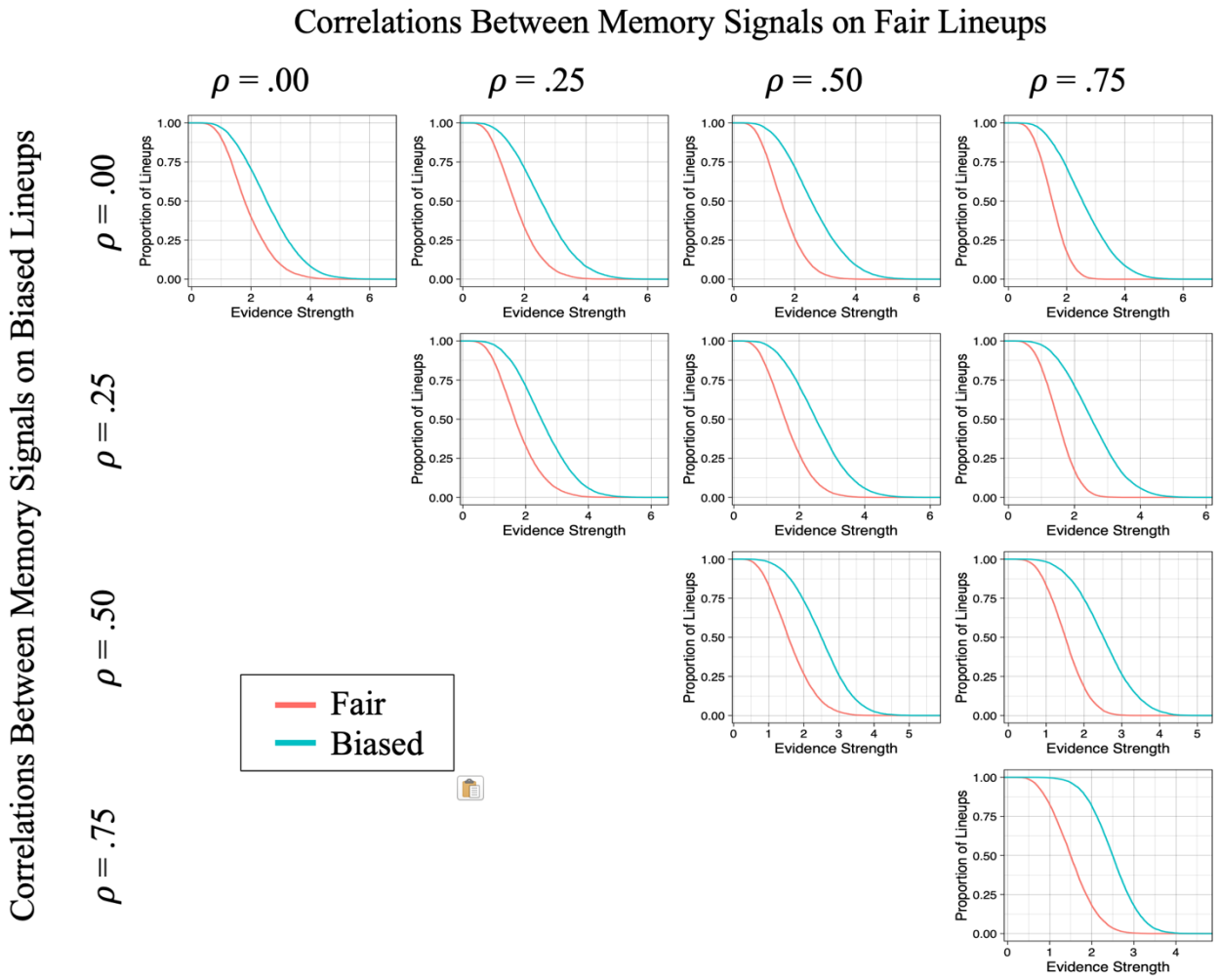


Figure S5: ***Evidence Strength Distributions Predicted by the Relative Model on Fair and Biased Culprit-Present Lineups***



Once again, although the above simulations did not generate predictions directly from the model likelihood functions, we verified that the likelihood functions generated the same predictions.

Fitting the Relative-Judgment Model to the Experimental Data

The purpose of the present section is to demonstrate that the relative model leads to the same general conclusions as the absolute model. Namely, low-similarity lineups better

discriminate between guilty-suspect identifications and innocent-suspect identifications than do high-similarity lineups and fair lineups better discriminate between guilty-suspect identifications and innocent-suspect identifications than do biased lineups. For reasons that we explain in the main body of this article, our intention was not to compare the absolute fit of non-nested models. There are several reasons why that is not an appropriate criterion for assessing construct validity. Nevertheless, it will become apparent below that the relative model provided a suboptimal absolute fit to both the data from Experiment 1 (comparing low-similarity and high-similarity lineups) and to the data from Experiment 2 (comparing fair and biased lineups).

In the main body of this paper, we refer to the theoretical models as the absolute and relative models and to their decision rules as the MAX and BEST-REST rules, respectively. In this supplemental materials document, we fit the relative model to the data from both Experiments 1 and 2. More specifically, we fit the ensemble model to the experimental data, which is the mathematical equivalent of the BEST-REST model (Wixted et al., 2018). For transparency, in this supplemental materials document we refer to the relative model as the ensemble model. We used R (R Core Team, 2021) and RStudio (RStudio Team, 2022) to facilitate the model-fitting routine. We wish to thank Akan et al. (2021) for making their Matlab code publicly available. As a starting point, we converted their Matlab code into R code.

Using the Ensemble Model to Assess the Impact of Absolute Filler Similarity on Suspect-Identification Discriminability (Experiment 1). As in the main body of this paper, we binned affirmative identifications into four confidence bins (90% - 100%, 70% - 80%, 50% - 60%, and 0% - 40%) and included a fifth bin comprised of lineup rejections collapsed over all levels of confidence. High and low similarity lineups each had 12 degrees of freedom: culprit identifications at each confidence bin [4], culprit-present filler identifications at each confidence

bin [4], and culprit-absent mistaken identifications at each confidence bin [4]. There were 24 degrees of freedom in total. The ensemble model included 10 free parameters. We permitted the location of the culprit distribution to vary freely in both low-similarity and high-similarity lineup conditions and estimated the locations of four decision criteria for each lineup condition. Hence, the unconstrained model had 10 free parameters and 14 degrees of freedom.

The best-fitting parameter estimates for the unconstrained model are summarized in Table S1 and Table S2 contrasts observed and predicted proportions. The unconstrained model provided a suboptimal fit to the data $\chi^2(14) = 31.36,\ p = .01$. As expected, the low-similarity lineup better discriminated between guilty-suspect identifications and innocent-suspect identifications than did the high-similarity lineup. To test whether this difference was significant we fit a simpler model in which we constrained the distance between the culprit and filler distributions to be equivalent across low-similarity and high-similarity lineups. The constrained model provided a poor absolute fit to the data, $\chi^2(15) = 93.02, p < .001$, and a significantly worse fit than the unconstrained model, $\chi^2(1) = 61.67, p < .001$. Hence, consistent with the absolute model, the ensemble model also leads to the conclusion that low-similarity lineups better discriminate between guilty-suspect identifications and innocent-suspect identifications than do high-similarity lineups.

Table S1: ***Best-Fitting Parameter Estimates of the Ensemble Model to Low- and High-Similarity Lineups***

Parameter	Low-Similarity Lineup	High-Similarity Lineup
μ_{culprit}	2.53	1.82
λ_{90-100}	2.60	2.24
λ_{70-80}	2.07	1.75
λ_{50-60}	1.75	1.45
λ_{0-40}	1.51	1.20

Table S2: ***Observed Values and Ensemble-Predicted Values for Low-Similarity and High-Similarity Culprit-Present and Culprit-Absent Lineups***

	Culprit Present			Culprit Absent		
Lineup	Hit Culprit	FA Filler	False Rejection	FA Innocent Suspect	FA Filler	Correct Rejection
Low Similarity						
90-100	.28 (.30)	.01 (.00)		.00 (.00)	.01 (.01)	
70-100	.49 (.51)	.03 (.01)		.01 (.01)	.07 (.06)	
50-100	.64 (.65)	.06 (.03)		.02 (.03)	.12 (.14)	
0-100	.71 (.73)	.09 (.05)	.20 (.22)	.04 (.05)	.22 (.24)	.73 (.72)
High Similarity						
90-100	.20 (.21)	.02 (.01)		.01 (.01)	.05 (.03)	
70-100	.38 (.40)	.09 (.05)		.03 (.03)	.14 (.13)	
50-100	.50 (.52)	.15 (.10)		.05 (.05)	.26 (.27)	
0-100	.59 (.61)	.21 (.16)	.19 (.24)	.08 (.09)	.40 (.42)	.53 (.49)

Note. Number in parentheses are predicted. FA = False Alarm.

Using the Ensemble Model to Assess the Impact of Lineup Bias on Suspect-Identification Discriminability (Experiment 2). As in the main body of this paper, we binned affirmative identifications into four confidence bins (90% - 100%, 70% - 80%, 50% - 60%, and 0% - 40%) and included a fifth bin comprised of lineup rejections collapsed over all levels of confidence. The fair lineup had 12 degrees of freedom in total: culprit identifications at each confidence bin [4], culprit-present filler identifications at each confidence bin [4], and culprit-absent mistaken identifications at each confidence bin [4]. For biased lineups the innocent suspect was drawn from a stronger strength distribution than were the lineup fillers and so we distinguished between culprit-absent mistaken identifications of innocent suspects and culprit-absent mistaken identifications of fillers. As a result, the biased lineup had 16 degrees of freedom: culprit identifications at each confidence bin [4], culprit-present filler identifications at each confidence bin [4], innocent suspect identifications at each confidence bin [4], and culprit-

absent filler identifications at each confidence bin [4]. Hence, there were 28 degrees of freedom total.

The ensemble model included 11 free parameters: three location parameters and eight decision criteria [four criteria for the biased lineups and four criteria for the fair lineups]. We permitted the location of the culprit distributions to vary freely on fair and biased lineups and we also permitted the location of the innocent-suspect distribution to vary freely on biased lineups. We also estimated the location of four decision criteria on both fair and biased lineups. Hence, the unconstrained model included 11 free parameters and 17 degrees of freedom.

The best-fitting parameter estimates for the unconstrained model are summarized in Table S3, and Table S4 contrasts observed and predicted proportions. The unconstrained model provided a suboptimal fit to the data $\chi^2(17) = 42.41,\ p < .001$. As expected, the fair lineup better discriminated between guilty-suspect identifications and innocent-suspect identifications than did the biased lineup. This is evidenced by the fact that the distance between the culprit and innocent-suspect distributions is greater for the fair lineup (1.19 – 0.00 = 1.19) than for the biased lineup (2.11 – 1.16 = 0.95). To test whether this difference was significant we fit a simpler model in which we constrained the distance between the culprit and filler distributions to be equivalent across fair and biased lineups. The constrained model provided a poor absolute fit to the data, $\chi^2(18) = 62.11, p < .001$, and a significantly worse fit than the unconstrained model, $\chi^2(1) = 19.70, p < .001$. Hence, consistent with the absolute model, the ensemble model also leads to the conclusion that fair lineups better discriminate between guilty-suspect identifications and innocent-suspect identifications than do biased lineups.

Table S3: ***Best-Fitting Parameter Estimates of the Ensemble Model to Fair and Biased Lineups***

Parameter	Fair Lineup	Biased Lineup
μ_{culprit}	1.19	2.11
$\mu_{\text{Innocent Suspect}}$	-	1.16
λ_{90-100}	2.14	2.54
λ_{70-80}	1.67	2.02
λ_{50-60}	1.38	1.71
λ_{0-40}	1.08	1.48

Table S4: ***Observed Values and Ensemble-Predicted Values for Fair and Biased Culprit-Present and Culprit-Absent Lineups***

	Culprit Present			Culprit Absent		
Lineup	Hit Culprit	FA Filler	False Rejection	FA Innocent Suspect	FA Filler	Correct Rejection
Fair						
90-100	.10 (.10)	.03 (.03)		.01 (.01)	.05 (.05)	
70-100	.22 (.23)	.11 (.09)		.03 (.03)	.17 (.16)	
50-100	.31 (.32)	.21 (.18)		.06 (.06)	.31 (.31)	
0-100	.41 (.41)	.33 (.30)	.26 (.29)	.10 (.10)	.48 (.50)	.42 (.39)
Biased						
90-100	.19 (.20)	.01 (.00)		.04 (.04)	.02 (.01)	
70-100	.37 (.38)	.04 (.02)		.12 (.12)	.06 (.04)	
50-100	.51 (.51)	.07 (.05)		.20 (.21)	.10 (.09)	
0-100	.61 (.61)	.10 (.08)	.29 (.31)	.27 (.28)	.14 (.15)	.59 (.57)

Note. Number in parentheses are predicted. FA = False Alarm.

References

Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2021). The effect of lineup size on eyewitness identification. *Journal of experimental psychology. Applied*, *27*, 369–392. https://doi.org/10.1037/xap0000340.

DeBruine, L. (2023). *Faux: Simulation for Factorial Designs.* doi:10.5281/zenodo2669586, R package version 1.2.1 https://debruine.github.io/faux/.

Murrell, P. (2005). *R Graphics.* Chapman & Hall/CRC Press.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Shen, K. J., Colloff, M F., Vul, E., Wilson, B. M., & Wixted, J. T. (2023). Modeling face similarity in police lineups. *Psychological Review, 130*(2), 432 – 461. https://doi.org/10.1037/rev0000408.

Smith, A. M., Smalarz, L., Wells, G. L., Lampinen, J. M., & Mackovichova, S. (2022). Fair lineups improve outside observers' discriminability, not eyewitness' discriminability: Evidence for differential filler-siphoning using empirical data and the WITNESS computer-simulation architecture. *Journal of Applied Research in Memory and Cognition.* Advance online publication. https://doi.org/10.1037/mac0000021

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, https://doi.org/10.21105/joss01686.

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology, 105,* 81 – 114. https://doi.org/10.1016/j.cogpsych.2018.06.001.