Absolute-Judgment Models Better Predict Eyewitness Decision-Making than do Relative-Judgment Models

Andrew M. Smith Rebecca C. Ying Alexandria R. Goldstein

Department of Psychology, Iowa State University

Ryan J. Fitzgerald

Department of Psychology, Simon Fraser University

Author Note

Andrew M. Smith https://orcid.org/0000-0002-4184-9364

Rebecca C. Ying https://orcid.org/0000-0002-5666-7790

Alexandria R. Goldstein https://orcid.org/0000-0003-1118-8558

Ryan J. Fitzgerald https://orcid.org/0000-0002-7249-7269

Correspondence concerning this article should be addressed to Andrew M. Smith, Department of Psychology, Lagomarcino Hall, 1347 Stange Rd., Ames, IA 50011. Email: amsmith@iastate.edu.
The preregistration, data analysis plan, anonymized data, and analysis code are available here: https://osf.io/t34pr/?view_only=75b9a275488b4b388f42101fece18efd

Abstract

When presented with a lineup, the witness is tasked with identifying the culprit or indicating that the culprit is not present. The witness then qualifies the decision with a confidence judgment. But how do witnesses go about making these decisions and judgments? According to absolute-judgment models, witnesses determine which lineup member provides the strongest match to memory and base their identification decision and confidence judgment on the absolute strength of this MAX lineup member. Conversely, relative-judgment models propose that witnesses determine which lineup member provides the strongest match to memory and then base their identification decision and confidence judgment on the relative strength of the MAX lineup member compared to the remaining lineup members. We took a *critical test approach* to test the *predictions* of both models. As predicted by the absolute-judgment model, but contrary to the predictions of the relative-judgment model, witnesses were more likely to correctly reject low-similarity lineups than high-similarity lineups (Experiment 1), and more likely to reject biased lineups than fair lineups (Experiment 2). Likewise, witnesses rejected low-similarity lineups with greater confidence than high-similarity lineups (Experiment 1) and rejected biased lineups with greater confidence than fair lineups (Experiment 2). Only a single pattern was consistent with the relative model and inconsistent with the absolute model: suspect identifications from biased lineups were made with greater confidence than suspect identifications from fair lineups (Experiment 2). The results suggest that absolute-judgment models better *predict* witness decision-making than do relative-judgment models and that pure relative-judgment models are unviable.

Keywords: eyewitness memory; eyewitness lineup; signal detection theory; relative judgment; absolute judgment; memory

Absolute-Judgment Models Better Predict Eyewitness Decision-Making than do Relative-Judgment Models

In the process of solving crimes, police investigators often present witnesses with lineups. A lineup is a procedure in which the photograph of a suspect is surrounded by the photographs of known-innocent persons called fillers and presented to a witness for an identification attempt. The task of the witness is to indicate which lineup member, if any, is the culprit. But how *do* witnesses go about making that judgment? How *should* witnesses go about making that judgment? Psychological scientists have long been interested in both descriptive and normative models of eyewitness decision-making. In the present work, we examined the cognitive processes that give rise to lineup decisions and associated expressions of confidence. More generally this work sheds light on the cognitive processes that underlie tasks of *unforced choice*—tasks in which respondents have the option of choosing from a limited array of options or rejecting the entire array (Cervantes & Benjamin, 2024). Specifically, we examined whether eyewitness identification decisions were more consistent with the predictions of *absolute-judgment models* or *relative-judgment models.*

The distinction between absolute-judgment strategies and relative-judgment strategies has long influenced theories of eyewitness decision-making (Wells, 1984, 1993). However, the historical distinction between absolute- and relative-judgment strategies in eyewitness identification procedures is distinct from how the absolute-relative distinction is defined more broadly in cognitive psychology. Wells (1984) introduced the *absolute-relative* distinction to explain why innocent persons were often mistakenly identified despite bearing only a modest resemblance to the culprit. According to Wells (1984), a witness who does not appreciate that the culprit might not be present in the lineup is at risk of adopting a *relative-judgment strategy,* and

identifying whichever lineup member provides the strongest match to memory for the culprit even if the absolute match is not particularly strong. Because the culprit might not be present in the lineup, what a witness should do instead is adopt an *absolute-judgment strategy* and only identify a lineup member if the *absolute* match between that individual and her memory for the culprit is strong. Hence, historically, the absolute-relative distinction was intended to explain why witness criterion setting was often too lax (Wells, 1984).

With the introduction of formal modeling to the identification literature, the relative-absolute framework evolved into a distinction in how witnesses assess the strength of memory evidence (Clark, 2003; Clark et al., 2011). It is in this vein that we examine absolute and relative judgment processes. As a starting point, we describe a lineup procedure from the perspective of signal-detection theory. Signal detection theory describes how judgments are formed under situations of uncertainty and has been applied to a wide range of contexts, including eyewitness identification (Green & Swets, 1966; Wickens, 2002). To that end, we start with an overview of the simplest of lineup models: the equal-variance signal-detection model with an absolute-judgment strategy. We refer to this as the absolute model, but it has also been referred to as the MAX model, the best-above model, the independent-observations model, and the dependent-observations model (Akan et al., 2021; Clark et al., 2011; Duncan, 2006; Macmillan & Creelman, 2005; Smith et al., 2022; Starns et al., 2023). After we introduce the absolute model, we then review an alternative model that assumes witness decision-making results from a relative-judgment process: the BEST-REST model (Clark, 2011; Sauer et al., 2008). The BEST-REST model is also commonly referred to as the ensemble model (Akan et al., 2021; Meyer-Grant & Klauer, 2022; Wixted et al., 2018).[1] We will refer to this model generally as the relative

[1] Although mathematically equivalent, there is one minor difference between the BEST-REST and ensemble model. As the name suggests, the BEST-REST rule assumes that the witness subtracts the average memory strength of the

model, but when greater precision is needed, we will refer to this as the BEST-REST model because that moniker more clearly articulates the presumed decision-making process than does the term ensemble model.[2]

Eyewitness Lineups from the Perspective of Signal Detection Theory

Figure 1 depicts the equal-variance signal detection model for eyewitness identifications (Green & Swets, 1966; Wickens, 2002). The horizontal axis is a random variable that reflects the match-to-memory strength of a given lineup member (absolute strength) and increases as one pans from left to right. The two Gaussian distributions reflect the range of latent match-to-memory values for innocent (novel) persons and culprits (previously seen persons), respectively. Because the witness has seen the culprit before, on average, a culprit should provide a stronger match-to-memory than should a given innocent person. Hence, the distribution of potential match-to-memory values for culprits is shifted to the right of the distribution of potential match-to-memory values for innocent persons. The standardized distance between the means of these two distributions is referred to as the discriminability index $\left(d' = \frac{\mu_{\text{Culprit}} - \mu_{\text{Innocent}}}{\sigma}\right)$ and reflects the ability of a witness to discriminate between the culprit and a given innocent person. It follows that a six-person culprit-present lineup can be represented by one random draw from the culprit distribution and five random draws from the innocent-person distribution. A six-person culprit-absent lineup can be represented by six random draws from the innocent-person distribution.

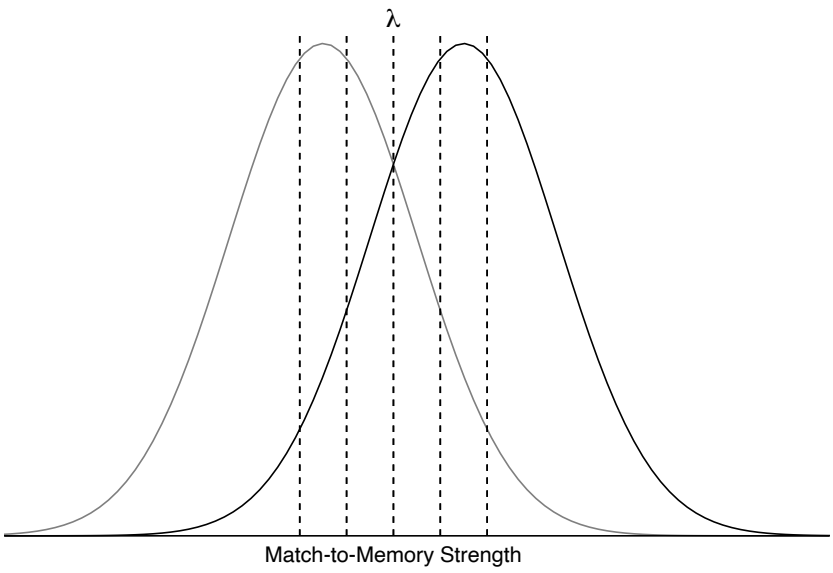
---

remaining lineup members from the strength of the best-matching lineup member. Conversely, the ensemble model assumes that the witness subtracts the average of all memory signals (including the best) from the best-matching lineup member. Despite this minor computational difference, the models are mathematically equivalent (Wixted et al., 2018).

[2] There are two other models that we do not address in the present manuscript: the integration model (e.g., Duncan, 2006; Palmer & Brewer, 2012) and the BEST-NEXT model (Clark, 2003; Clark et al., 2011). Both models have fallen out of favor in the literature. The decision rule proposed by the integration model is implausible, even at face, and the model has struggled to explain lineup data (e.g., Wixted et al., 2018). Although the BEST-NEXT model proposes a reasonable decision rule, it was supplanted long ago by the BEST-REST rule (Clark et al., 2011; Sauer et al., 2008; Wixted et al., 2018) and data patterns are more consistent with the BEST-REST rule than with the BEST-NEXT rule (Charman et al., 2011; Horry & Brewer, 2016).

The Absolute-Judgment Model. But how do witnesses go about making identification decisions? According to the absolute model, the witness compares each lineup member to her memory for the culprit and determines which one provides the strongest match-to-memory (i.e., the MAX signal). If the degree of match between the MAX lineup member and the witness' memory for the culprit exceeds the witness' decision criterion (λ), the witness identifies that person and otherwise the witness rejects the lineup. The absolute model can be extended further to account for expressions of confidence. Confidence reflects the extent to which the signal of the MAX lineup member exceeds or falls short of the witness' decision criterion. This is easily captured by assuming that the witness holds a series of decision criteria rather than a single criterion. Confidence is equal to the rightmost criterion that the MAX memory signal exceeds. Hence, if the match-to-memory strength of the MAX lineup member exceeded the rightmost criterion, the witness would identify that person with high confidence, and if the match-to-memory strength of the MAX lineup member fell below the leftmost criterion, the witness would reject the entire lineup with high confidence.

Figure 1: ***The Equal Variance Signal Detection Model***



Note. The rightmost distribution (black) reflects the range of latent match-to-memory strength values for the culprit and the leftmost distribution (grey) reflects the range of latent match-to-memory strength values for innocent persons. The dashed vertical lines represent decision and confidence criteria.

The Relative-Judgment Model. In contrast, the relative-judgment model assumes that witnesses base their decisions not on the absolute strength of the MAX lineup member, but on the difference in strength between the MAX lineup member and the average strength of the remaining lineup members. In other words, witnesses transform the raw strength variable depicted on the horizontal axis of Figure 1 into a difference score between the MAX lineup member (AKA BEST) and the average strength of the remaining lineup members (REST). The BEST-REST rule assumes that a witness identifies the MAX lineup member if the difference in match-to-memory strength between that lineup member and the average match-to-memory strength of the remaining lineup members exceeds the witness' decision criterion (Clark et al., 2011; Sauer et al., 2008; Wixted et al., 2018). Like the absolute-judgment model, the relative-judgment model can easily be extended to accommodate expressions of confidence. Confidence reflects the extent to which the BEST-REST score exceeds or falls short of the witness' criterion. As with the absolute model, this is easily captured by assuming that witnesses do not hold a single decision criterion but rather a series of decision criteria. Confidence reflects the rightmost criterion that the BEST-REST score exceeds.

Notice that these two models hold many assumptions in common. Both models assume that witnesses start by examining the lineup and determining which lineup member provides the strongest match to memory. Both models also assume that a witness would never identify anyone but the MAX lineup member. And, both models assume that the witness' decision to identify the MAX lineup member or to reject the lineup and the expression of confidence associated with that decision is based on comparing the evidence strength to their internal decision criteria. The defining difference between these two variables lies in how they operationalize the decision rules

(evidence strength). Whereas the absolute model operationalizes the decision rule as the absolute strength of the MAX lineup member, the relative model operationalizes the decision rule as the strength of the MAX lineup member minus the average strength of the remaining lineup members (BEST – REST). Hence, distinguishing between these two models boils down to determining which decision rule more accurately predicts outcomes from eyewitness lineups.

The Use of Absolute- and Relative-Judgment Strategies on Tasks of Recognition Memory

Fundamental research on human recognition memory focuses primarily on two tasks: the old/new task and the 2-alternative forced-choice (2-AFC) task. On an old/new recognition task, respondents study a list of to-be-remember items (e.g., 40 words) and after a short delay, they are presented with a test list. The test list is comprised of both the studied (old) items and non-studied (new) items. Studied and non-studied items are presented to respondents one at a time in a random order. On each trial, the respondent is tasked with indicating whether the item is old (studied) or new (non-studied) and providing an expression of confidence in that decision. Respondents base their recognition decisions and confidence judgments on the absolute match-to-memory strength of the test probe (Macmillan & Creelman, 2005; Wickens, 2002).

On a 2-AFC task, respondents study a list of to-be-remember items (e.g., 40 words) and after a short delay, they are presented with a test list. Like the old/new test, the test list is comprised of both the studied (old) items and non-studied (new) items. Unlike an old/new test each trial is comprised of one old item and one new item, and the task of the respondent is to indicate which of the two items is old and to provide an expression of confidence in that decision. It is commonly assumed that decisions and confidence judgments on a 2-AFC task are based on computing the difference in match-to-memory strengths between the two items (a relative-judgment process) (Jang et al., 2009; Macmillan & Creelman, 2005). However, the

results of several experiments suggest that in at least some situations, respondents might ignore the strength of the non-MAX signal and rely on only the absolute strength of the MAX signal (Hanczakowski et al., 2021; Hockley, 1984; Starns et al., 2017; Zawadzka et al., 2017).

An eyewitness lineup is a hybrid of the old/new and alternative-forced choice tasks that have dominated the fundamental recognition literature. On a lineup task, the witness is exposed to a crime, and then after a delay is presented with either a culprit-present or culprit-absent lineup. Like old/new tasks, some lineups do not include the culprit and the correct response is to indicate that the culprit is not present. But like AFC tasks, when the culprit is present, it is insufficient to merely indicate that the culprit is present; the witness must also identify which person is the culprit. Given that lineups include elements of both old/new and alternative-forced choice tasks, both absolute judgments and relative judgments are candidate strategies for making identification decisions.

There is some evidence that witnesses' confidence judgments result from relative-judgment strategies. Adding four implausible individuals (duds) to a lineup comprised of two plausible individuals increased witness confidence that the plausible lineup members were the culprit (Charman et al., 2011; see also Windschitl & Chambers, 2004). Likewise, replacing high-similarity lineup fillers with low-similarity lineup fillers increased witness confidence in correct identifications of the culprit (Horry & Brewer, 2016). Both patterns are predicted by the relative model: adding duds to a lineup or replacing plausible lineup members with duds decreases the average signal strength of the non-MAX lineup members (REST) and increases the magnitude of the BEST-REST difference score. In contrast, the absolute model predicts that witness confidence would be invariant to changes in the strength of the non-MAX lineup members.

That the match-to-memory strength of non-identified lineup members impacts eyewitness expressions of confidence has been taken by some to imply that lineup data are more consistent with a relative-judgment process than they are with an absolute-judgment process (Wixted et al., 2018). Yet, the *decision-making* patterns in these same studies paint a different picture. Adding duds to a lineup does not increase the identification of plausible lineup members (Charman et al., 2011). Likewise, decreasing the similarity of lineup fillers to the suspect, if anything, increased correct rejections, which is the opposite of what a relative-judgment model predicts (Horry & Brewer, 2016). Both patterns are consistent with the absolute model which predicts that witness decision-making is influenced only by the absolute strength of the MAX lineup member and not by the strength of non-MAX lineup members. Although confidence-judgment data are consistent with the predictions of the relative model, the identification decisions themselves are consistent with the predictions of the absolute model. Hence, these data suggest that witness decision-making might be driven by absolute strength, but witness confidence judgments by relative strength.

More recent work investigating eyewitness decision strategies has taken to fitting models to empirical data. This work has concluded that both the relative model and the absolute model can account for eyewitness decision-making on lineups (Akan et al., 2021) but that the relative model provides a closer fit to the data (Shen et al., 2023; Wixted et al., 2018). Despite many virtues, one problem with a retrospective-fitting approach is that models can accommodate data patterns that they would not have predicted *a priori*. Rather than retrospectively fitting models to data to determine which model provides a closer fit, we took a *critical test approach* (e.g., Allais, 1953; Birnbaum, 2011; Cha & Dobbins, 2021; Dobbins, 2023; Kellen & Klauer, 2015; Kellen et al., 2021; Ma, Starns, & Kellen, 2021). That is, we determined instances in which the absolute-

and relative-judgment models made divergent predictions and then carried out experiments to assess the viability of both pure absolute models and pure relative models and to determine whether eyewitness decision-making was more consistent with the predictions of the absolute model or the relative model.

Rather than comparing closeness-of-fit for absolute and relative models, our aim was to test competing predictions derived from the two models. Closeness-of-fit will play no role in our consideration about whether the absolute or relative judgment model has superior construct validity (viz. better captures how witnesses make identification decisions). Historically, fit statistics have figured prominently in the comparison of non-nested decision models, but more recent work has emphasized the importance of the critical-testing approach (e.g., Dobbins, 2023; Kellen et al., 2021; Ma et al., 2021). By and large, the shift away from retrospective model fitting and towards the critical testing approach is attributable to the fact that retrospective model fitting does not typically inform on which of two models has superior construct validity (Dobbins, 2023). Indeed, even models with psychologically meaningless parameters (e.g., a polynomial regression model) can retrospectively provide a good fit to empirical decision data. Conversely, due to measurement error, a psychologically valid model will sometimes provide a poor fit to empirical decision data. If invalid models can provide good fits and valid models can provide poor fits, then retrospective model fit is not an appropriate arbiter on which of two models has greater construct validity (see Dobbins, 2023). Finally, it has already been established through model-recovery simulations that both the absolute and relative models have the potential to retrospectively fit data that was generated by its counterpart (Akan et al., 2021; Cervantes & Benjamin, 2024). Hence, there are numerous good reasons to use a critical testing approach rather than focusing on retrospective closeness of fit. Following that logic, we opted not to use

retrospective model fitting, and instead generated predictions from competing decision models, determined situations in which the models made opposing predictions, and used experimentation to determine a winner.

The Present Study: Using Critical Tests to Distinguish Between Absolute- and Relative-Judgment Strategies

Determining critical tests is complicated by the fact that, despite assuming fundamentally different decision-making strategies, the absolute and relative models often predict similar patterns of results. This is especially true for *suspect-identification rates,* which has long been the primary focus of the identification literature. Consider for example a manipulation of lineup composition bias. A typical biased lineup is one in which the suspect matches the witness' description of the culprit, but the fillers do not (Fitzgerald et al., 2013; Lindsay & Wells, 1980). When this happens, even an innocent suspect would tend to provide a stronger match to memory than the lineup fillers. In essence, the fillers have been drawn from a weaker match-to-memory strength distribution than the innocent suspect. Contrast this with a fair lineup where the fillers match the witness' description of the culprit and are ostensibly drawn from the same match-to-memory strength distribution as the innocent suspect (as assumed in Figure 1). Both absolute and relative models assume that the probability of a suspect identification is equal to the joint probability that the suspect provides the strongest match-to-memory AND that the memory signal for the suspect exceeds the witness' decision criterion. The absolute strength of the suspect is constant across biased and fair lineups, but the relative strength of the suspect is higher in the biased lineup. Hence, the expected BEST-REST value is larger for biased lineups than for fair lineups and the relative-judgment model predicts more suspect identifications from biased lineups. But despite assuming an absolute decision rule, the absolute model also predicts more

suspect identifications from biased lineups because the suspect is more likely to provide the MAX signal on a biased lineup compared to a fair lineup.

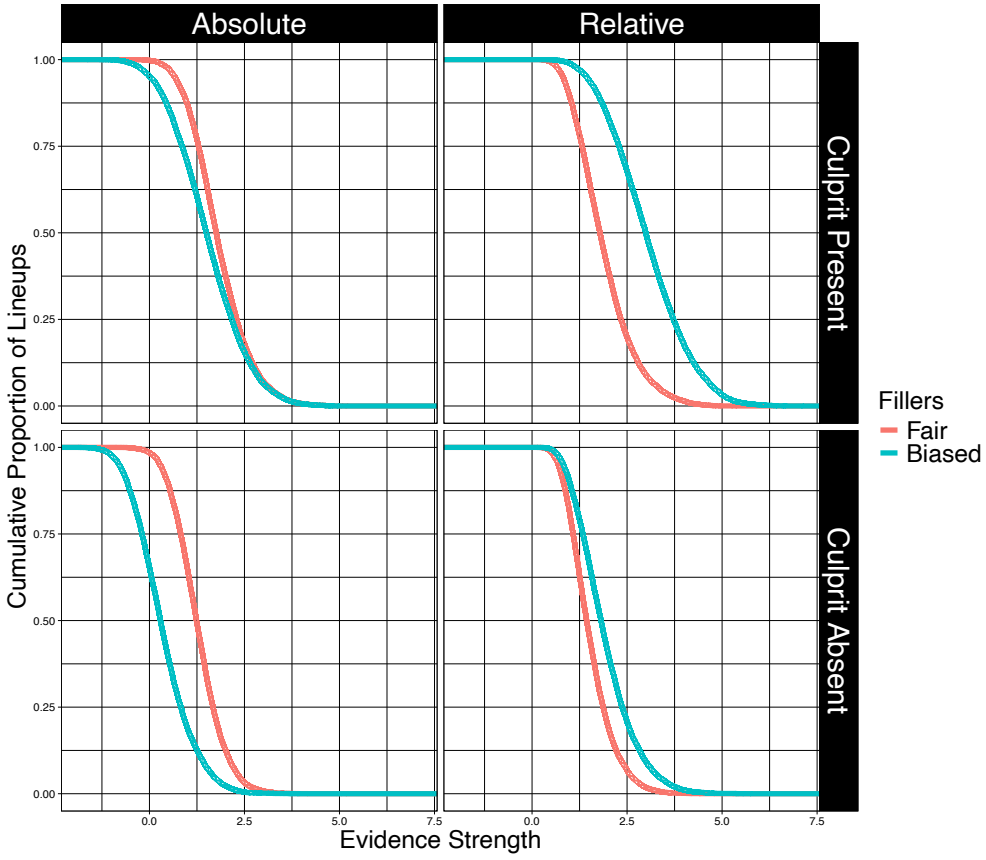
Rejection Rates from Fair and Biased Lineups. However, when it comes to rejection rates, the two decision rules often make divergent predictions. For example, the absolute model predicts more rejections from biased lineups than from fair lineups, and the relative model predicts the opposite pattern. The absolute model predicts more rejections from biased lineups because the expected MAX value is less for a biased lineup than for a fair lineup. In other words, compared to fair lineups, biased lineups decrease the expected absolute strength of the MAX signal. Indeed, if you take six random draws from a standard normal distribution [$X \sim N(\mu = 0, \sigma = 1)$], the expected value of the MAX draw is 1.27 (Smith et al., 2023; Yang & Burke, 2022). But if you draw only one memory signal from that distribution (the innocent suspect) and the remaining memory signals from a distribution with weaker strength (the fillers), the expected MAX value is less than 1.27.[3] Conversely, the relative model predicts more rejections from fair lineups because the expected difference between the MAX lineup member and the remaining lineup members is smaller when all lineup members are drawn from the same underlying distribution.

Figure 2 shows the complements of the cumulative distribution functions predicted by the MAX and BEST-REST decision rules for fair and biased culprit-present and culprit-absent lineups. In other words, what the graphs show are the cumulative proportions of lineups (on the

[3] An alternative way to manipulate lineup bias is by holding the similarity between the fillers and the culprit constant and manipulating whether the innocent suspect is as similar to the culprit as are the fillers (fair lineup) or more similar to the culprit than are the fillers (biased lineup). In this context, the expected MAX signal would be greater for the biased lineup than for the fair lineup and therefore the absolute model would predict fewer correct rejections from the biased lineup than from the fair lineup. Although this could easily be manipulated in laboratory experiments, in criminal investigations it is typically the case that police investigators generate a suspect and then have control over who they select as fillers. Hence, the similarity between the innocent-suspect and the culprit is typically fixed and the similarity between the fillers and the culprit is free to vary. For simplicity, we focus on the typical situation.

vertical axis) that exceed any fixed level of evidence strength (on the horizontal axis). What is important to note is that for the absolute (or MAX) model, the distribution for the biased lineup is shifted to the left of the distribution for the fair lineup. This means that the absolute model predicts more rejections from biased lineups than from fair lineups. Conversely, for the relative (or BEST-REST) model, the distribution for the biased lineup is shifted to the right of the distribution for the fair lineup. This means that the relative model predicts more rejections from fair lineups than from biased lineups. Hence, one of our two critical tests involves manipulating lineup composition bias and comparing rejection rates. Details on the simulation used to generate these predictions are included in the Figure note.

Figure 2: ***Evidence Strength Distributions Predicted by the Absolute and Relative Models on Fair and Biased Lineups***



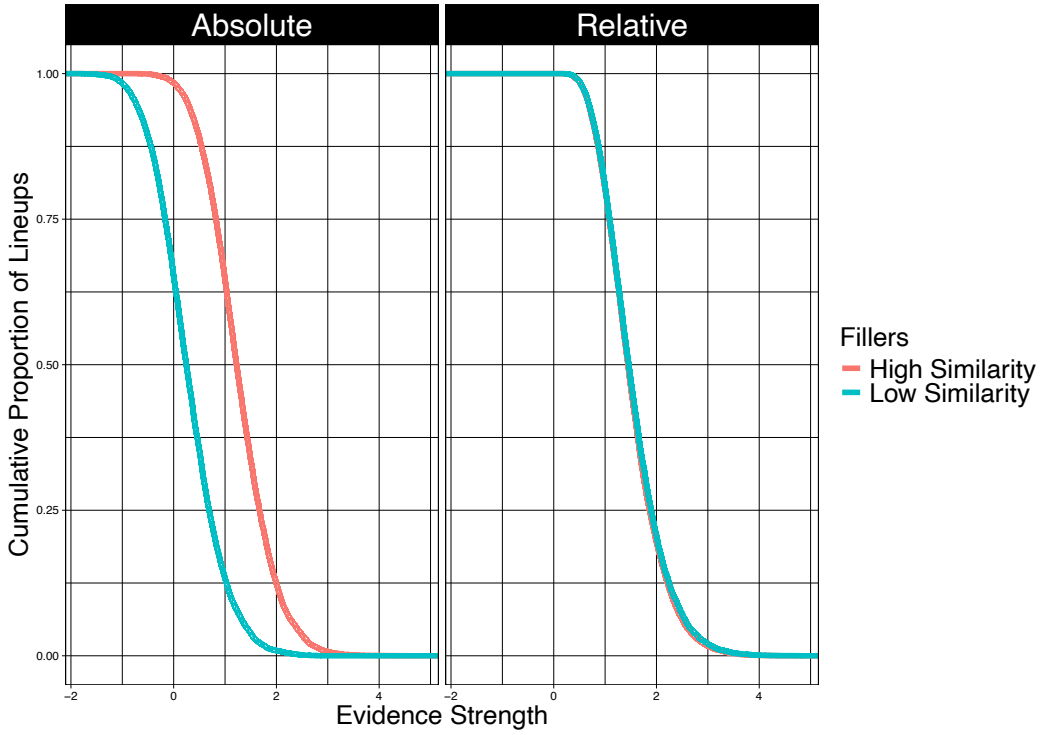
Note. Predictions were derived from simulations (N = 10,000) of six-person fair and biased culprit-present and culprit-absent lineups under the assumptions of MAX and BEST-REST decision rules. Fair culprit-present lineups were comprised of one random draw from the culprit distribution [$X \sim N(\mu = 1.5, \sigma = 1)$] and five random draws from the fair filler distribution [$X \sim N(\mu = 0, \sigma = 1)$]. Biased culprit-present lineups were comprised of one random draw from the culprit distribution and five random draws from the biased filler distribution [$X \sim N(\mu = -1.5, \sigma = 1)$]. Fair culprit-absent lineups were comprised of six random draws from the fair culprit-absent distribution and biased culprit-absent lineups were comprised of one draw from the fair culprit-absent distribution and five draws from the biased culprit-absent distribution. We also examined predictions for various other parameter settings and the MAX rule never predicted fewer rejections from fair lineups than from biased lineups and the BEST-REST rule never predicted fewer rejections from biased lineups than from fair lineups. See footnote 3 for discussion of a different type of lineup bias where the absolute model predicts more rejections of fair lineups than of biased lineups.

Rejection Rates from Low- and High-Similarity Culprit-Absent Lineups. The absolute and relative models also make divergent predictions about what impact changes in absolute match-to-memory strength have on rejection rates. In theory, these predictions apply to both culprit-present and culprit-absent lineups, but in practice it would be very difficult to manipulate absolute strength independently of relative strength in culprit-present lineups. That would require strengthening culprit and filler strengths by equal amounts, probably with two separate manipulations, and it would always be debatable as to whether culprits and fillers had been strengthened to the same extent. Accordingly, for absolute strength, we focus only on culprit-absent lineups.

The purest way to manipulate match-to-memory strength on culprit-absent lineups is by holding encoding conditions constant and manipulating the similarity of innocent lineup members to the culprit. Imagine two culprit-absent lineups, one where all the lineup members are high in similarity to the culprit and a second where all the lineup members are low in similarity to the culprit. The absolute model predicts more rejections from the low-similarity lineup because the expected MAX signal strength is weaker for the low-similarity lineup compared to the high-similarity lineup. Conversely, the relative model predicts no difference in correct rejection rates because the expected BEST-REST value is invariant to changes in absolute strength.

These predictions are illustrated in Figure 3, which was based on a simulation in which we manipulated the absolute similarity of culprit-absent lineup members to the culprit. Notably, for the absolute (MAX) model, the cumulative evidence strength for the low-similarity lineup is shifted to the left of the cumulative evidence strength for the high-similarity lineup, which means the absolute model predicts more rejections from low-similarity lineups than from high-similarity lineups. For the relative (BEST-REST) model, the cumulative evidence strength for the low-similarity lineup falls right on top of the cumulative evidence strength for the high-similarity lineup, meaning that the relative model predicts no change in rejection rates. Hence, another critical test of absolute and relative models involves comparing correct rejection rates for low-similarity and high-similarity lineups. Details on the simulation used to generate these predictions are included in the Figure note.

Figure 3: ***Evidence Strength Distributions Predicted by the Absolute and Relative Models on High-Similarity and Low-Similarity Culprit-Absent Lineups***



Note. Predictions were derived from simulations (N = 10,000) of six-person low-similarity and high-similarity culprit-absent lineups under the assumptions of MAX and 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 [$X \sim N(\mu = 0, \sigma = 1)$]. We also examined predictions for various other parameter settings and the MAX rule always predicted more correct rejections from low-similarity lineups and the BEST-REST rule never predicted a difference in correct rejections.

The Assumption of Correlated Memory Signals Does Not Moderate the Critical Test Predictions. Many recent attempts to model decision-making on eyewitness lineups assume that the probability of being identified is more similar for members *within* lineups than it is for members who are in different (*between*) lineups. There are several good reasons to assume this type of interdependency. For instance, different witnesses adopt different decision criteria. When a witness with a lenient criterion encounters a lineup that increases the probability of identification for all members of that lineup compared to if that witness had a more stringent criterion (Smith et al., 2017). Variations in levels of attention at encoding also leads to the prediction that the probability of identification ought to be more similar for members *within* lineups than it is for members *between* lineups (Wetmore et al., 2017).

Another reason for predicting that the probability of identification ought to be more similar for members within lineups than for members between lineups is because members within the same lineup tend to be matched to the culprit on a common set of features, either because they were matched to the appearance of the suspect or to the witness' description of the culprit. Either way, their signal strengths should be correlated (Akan et al., 2021; Shen et al., 2023; Smith et al., 2022; Wixted et al., 2018). The implication is that if one innocent lineup member tends to provide a relatively strong match to the witness' memory for the culprit then the other innocent lineup members should also tend to provide a relatively strong match to the witness' memory for the culprit and if one innocent lineup member tends to provide a weak

match to the witness' memory for the culprit, all lineup members should tend to provide a weak match to the witness' memory for the culprit.

There is little doubt that criterial variance, variations in levels of attention at encoding, and correlated signals are all operating in both actual police lineups and in laboratory experiments. But all three parameters tend to have similar impacts on modeling outcomes and so it would make little sense to instantiate all three of these parameters at once (Smith et al., 2022). Because the assumption of correlated memory signals is the most used approach to account for lineup dependencies, that is the approach that we consider here.

To keep the presentation of model predictions depicted in Figures 2 and 3 as straightforward as possible, we assumed statistical independence of the lineup signals. However, in our supplemental materials file, we generated predictions across a broad range of variations in correlated memory signals and the model predictions were consistent with what we found when assuming independent signals (see Figure 2 and Figure 3). Across all variations in the correlation parameter for both culprit-present and culprit-absent conditions, the relative model consistently predicted more rejections of fair lineups than of biased lineups. The absolute model consistently predicted more rejections of biased culprit-absent lineups than fair culprit-absent lineups. For culprit-present lineups, the absolute model predicted slightly more rejections from biased than fair lineups under the assumption of modestly correlated signals and no difference in rejection rates for higher levels of correlation. For low-similarity versus high-similarity culprit-absent lineups, the absolute model consistently predicted more rejections of low-similarity lineups than high-similarity lineups and the relative model consistently predicted no change in rejection rates across low-similarity and high-similarity lineups. We provide greater detail in the supplemental materials document.

Summary of Experiments

In summary, we determined two situations in which the absolute and relative judgment models make contradictory predictions about lineup outcomes. As it turns out, absolute and relative judgment models make contradictory predictions for some of the most fundamental problems in the identification literature. It is hard to imagine something more fundamental to the identification literature than understanding how variations in similarity between culprits and innocent-persons affects correct rejection rates. Any model that does not make accurate predictions about how variations in absolute signal strength impact correct rejection rates is unviable. Likewise, we cannot entertain a decision rule that makes erroneous predictions about what impact lineup bias might have on rejection rates. Accordingly, in our first experiment, we manipulated the absolute strength of culprit-absent lineups and examined the impact on correct rejection rates. In our second experiment, we manipulated lineup composition bias and examined whether the pattern of rejection rates was consistent with the absolute model or the relative model.

Experiment 1: Comparing Correct Rejection Rates from Low-Similarity Culprit-Absent Lineups and High-Similarity Culprit-Absent Lineups

Methods

The preregistration, data analysis plan, anonymized data, and analysis code are available here: https://osf.io/t34pr/?view_only=75b9a275488b4b388f42101fece18efd (Smith et al., 2023). These experiments were declared exempt by the [redacted for review] Institutional Review Board. The primary purpose of our initial experiment was to compare rejection rates for high-similarity and low-similarity culprit-absent lineups. We conducted an *a priori* power analysis using ANOVA_Power Shiny Application (Lakens & Caldwell, 2021). We used a 2-factor

ANOVA design as a proxy for estimating power. This power analysis revealed that a sample size of approximately N = 300 would give us 93% power to detect a small effect of filler similarity (*f* = .03). At the time of planning this project, there were no available tools for determining power for ROC designs that we were aware of and so we used the ANOVA design as a rule of thumb as others have done in the past (e.g., Akan et al., 2021; Ayala & Smith, 2024).

Participants

Participants were 335 MTurk workers that were paid in exchange for their participation. We used CloudResearch for additional data management tools (Litman et al., 2017). We excluded data from two participants who experienced technological issues and 21 participants that demonstrated no ability to discriminate between guilty persons and innocent persons. This left us with a final sample of 312 participants. From the 312 participants that we retained in our final sample we excluded 24 trials (out of 2496) where participants reported technological issues. Out of the final sample, 47% identified as female, 51% identified as male, less than 1% identified as non-binary, and less than 1% opted not to report. When asked about their race, 77% identified as White, 10% identified as Black/African American, 6% identified as Asian, 4% identified as Hispanic or Latino/a, 2% identified as another race, and less than 1% opted not to state their race. On average, participants were 42.75 years of age (*SD* = 12.61).

Design

We used a 2 (Culprit: Present vs. Absent) x 2 (Fillers: High-similarity vs. Low-similarity) within-participants design. Each participant watched eight videos and completed eight lineup tasks (two in each cell of the design). Before conducting the current study, we created eight stimulus groups, each containing eight videos and eight lineups (two for each cell of the design). Participants were randomly assigned to one of these eight preset stimulus groups. Within each

stimulus group, video and lineup conditions were fixed (e.g., participants assigned to the same stimulus group saw the same set of target videos and completed the same type of lineup for each target). Furthermore, the proportion of culprit-present and culprit-absent lineups within a given block varied from one block to the next. We randomized the proportion of culprit-present lineups on each block so that participants could not use their experience on one lineup as basis for what decision they should make on a subsequent lineup.

Culprit-present lineups were comprised of the culprit and five lineup fillers whom the witnesses had not seen before. Culprit-absent lineups were comprised of six lineup fillers whom the witness had not seen before. We describe the filler selection process in the materials subsection, but high-similarity fillers were relatively high in similarity to culprit and low-similarity fillers were relatively low in similarity to culprit. Culprit-absent lineups did not include designated innocent suspects. Instead, we estimated innocent-suspect identification rates by dividing the culprit-absent false alarm rate by the nominal lineup size (six).

Materials

Videos. Videos were sourced from the eyewitness-identification stimulus database (Fitzgerald et al., 2023). We used a total of 16 simulated crime videos. Each video depicted a single individual entering a room, stealing either a laptop or an iPad out of a bag, and then leaving. Video duration ranged from 23 seconds to 37 seconds. For each video there is a clear view of the culprit's face that lasts for approximately 10 seconds.

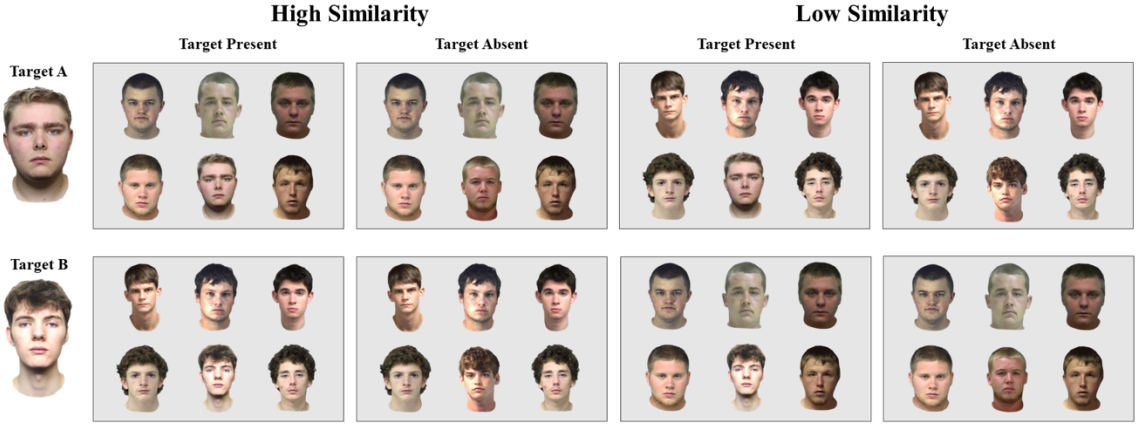
Lineups. All lineups were presented simultaneously and consisted of six faces presented to participants in a 2 rows x 3 columns photo array. Culprit-present lineups were comprised of the culprit from the simulated crime video and five lineup fillers. Culprit-absent lineups were comprised of six lineup fillers. Lineup positions were randomized anew for each participant and

for culprit-present lineups each of the six lineup fillers had a 5/6 chance as being included in the culprit-present lineup.

We used the culprit photos from the Eyewitness Lineup Identity database (Fitzgerald et al., 2023), which includes mugshots created with a recording booth on loan from the Video Identification Parade Electronic Recording (VIPER) Bureau, West Yorkshire Police, England. Images in the database were not quality assured by the VIPER Bureau, and the authors accept full responsibility for their quality. The people depicted in the database are actors, not actual culprits or lineup members in real criminal cases. Filler photos were from the Florida Mugshot Database.

Because our objective was to manipulate absolute similarity independently of relative similarity, we took a somewhat unique approach to constructing lineups. We looked through the culprit pool to find pairs of culprits who fit the same vague description (e.g., white male, early 20s, with brown hair). We created eight of these culprit pairs (i.e., A – A', …, H – H'). We then selected high-similarity fillers for each of the 16 culprits (A, …, H'). To that end, we instructed a team of research assistants to find six fillers that both matched the description of the culprit and who generally resembled the culprit. To create low-similarity lineups, we swapped the fillers for Culprit A with the fillers for Culprit A', the fillers for Culprit B with the fillers for Culprit B', and so on and so forth. This way we ensured that the similarity among individuals within a lineup remained constant across the high-similarity and low-similarity conditions. Sample stimuli for two targets are presented in Figure 4.

Figure 4: ***Examples of High-Similarity and Low-Similarity Target-Present and Target-Absent Lineups***



Procedure

We used the Qualtrics survey platform to facilitate the current study. Participants were only eligible if they: (1) were over 18 years of age, (2) lived in the United States, (3) were fluent in English, (4) had at least an 80% approval rate on at least 100 previous MTurk tasks, and (5) were using a computer or laptop to complete the task. Upon completing informed consent, participants were asked if they agreed to pay attention and follow instructions throughout the study. If they declined, they were instructed that they were not eligible to participate. Subsequently, the participants completed two simple bot-check questions to confirm their humanity, in which they were asked to select a specific letter from a list of multiple-choice options. Failure resulted in dismissal from the experiment.

To limit the amount of time that it took to complete the survey, we grouped the project into four blocks each containing two videos, a single filler task, and two lineups. The order in which the blocks were displayed to participants was randomized across participants, but the order of videos and lineups within each block was fixed. We constrained blocks to always contain two culprits who did not fit the same description so that witnesses would know which

trace was being probed by which lineup. For example, if the first culprit in a block was a White Male the second culprit might have been a Black Female.

On each block, participants were provided with basic task instructions. They were then asked to pay careful attention to each of the two videos as they would be presented with lineups related to these videos at a later point. After watching the two simulated-crime videos, participants completed a one-minute anagram-solving task. After the anagram task participants were instructed that they would complete lineups for each of the two persons that they saw before the anagrams task. Further, participants were admonished that the person from the video may or may not be present in the lineup and were asked to identify that person if present and otherwise to select "Not Present". Following each identification decision, participants were asked to express their level of confidence on a scale from 0% (not at all) to 100% (completely) in 10-point increments. Once they were done with both lineups, they were asked to report any technical difficulties they experienced with the videos or the lineups.

Results

Although retrospective model fit is not an appropriate arbiter for determining which of two non-nested decision models has superior construct validity, it is appropriate for comparing the relative fit of nested models (Dobbins, 2023). Accordingly, we fit two versions of the absolute model to the data to determine whether we had produced the expected finding that low-similarity lineups better discriminate guilty-suspect identifications from innocent-suspect identifications than do high-similarity lineups (e.g., Carlson et al., 2023). In the first model, we permitted discriminability to vary across low-similarity and high-similarity lineups and in the second model, we constrained discriminability to be equal across our manipulation of similarity. We then compared the relative fit of these two nested models. After establishing that our

manipulation had its intended impact, we then present the results of our critical test assessing whether low-similarity lineups led to more correct rejections than did high-similarity lineups. Although both the absolute and relative models predict better discriminability for low-similarity lineups, only the absolute model predicts more correct rejections for low-similarity lineups. The relative model predicts equivalent correct rejection rates for low-similarity and high-similarity lineups. In the process of analyzing the data from both Experiment 1 and Experiment 2, we used R (R Core Team, 2021), RStudio (RStudio Team, 2022), Tidyverse (Wickham et al., 2019), and lme4 (Bates et al., 2015).

Assessing the Impact of Filler Similarity on Suspect-Identification Discriminability

Following standard practice, 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 (e.g., Smith et al., 2018; Wixted et al., 2018). 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]. Hence, there were 24 degrees of freedom in total. The absolute model included 10 free parameters. We fixed the location of the low-similarity filler distribution ($\mu = 0$) and permitted the high-similarity filler distribution (μ_{Filler}) and the culprit distribution (μ_{culprit}) to vary, but we assumed that the location of the culprit distribution was constant across low-similarity and high-similarity lineups. We also estimated locations for the eight confidence criteria, four for the low-similarity lineups and four for the high-similarity lineups. Hence, the unconstrained model included 10 free parameters and 14 degrees of freedom. For simplicity, we assumed that the memory signals were uncorrelated. We made this assumption because by design low-similarity fillers were as correlated with other

low-similarity fillers as high-similarity fillers were with other high-similarity fillers and therefore there was little to be gained from adding an additional parameter to the model.

The best-fitting parameter estimates for the unconstrained model are summarized in Table 1 and Table 2 contrasts observed and predicted proportions. The unconstrained model provided a good fit to the data, $\chi^2(14) = 18.50,\ p = .18$. The good fit between the model and the observed data is further evidenced by the Receiver Operating Characteristic (ROC) curves depicted in Figure 5. The operating points in Figure 5 represent the empirical data, and the curves represent the predictions of the MAX model. 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, $\chi^2(15) = 79.89, p < .001$, and a significantly worse fit than the unconstrained model, $\chi^2(1) = 61.39, p < .001$.[4] We also fit two versions of the relative model to determine whether it also supported the conclusion that low-similarity lineups have better discriminability than high-similarity lineups. We refer the interested reader to supplemental materials.

[4] Consistent with previous modeling of lineup data (see also Akan et al., 2021; Shen et al., 2023; Smith et al., 2022), we combined the data from all participants and fit the model at the aggregate because there were not enough data points to fit the model at the level of individual participants.

Table 1: ***Best-Fitting Parameter Estimates of the Absolute model to Low- and High-Similarity Lineups***

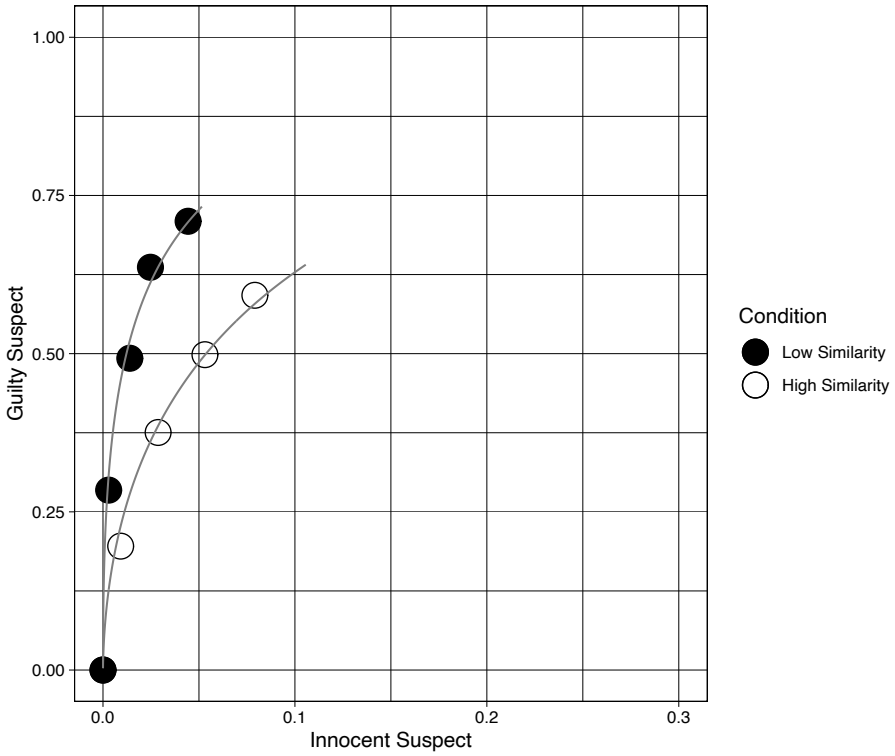
Parameter	Low-Similarity Lineup	High-Similarity Lineup
μ_{culprit}	2.25	2.25
μ_{Filler}	0_{Fixed}	0.64
λ_{90-100}	2.82	3.10
λ_{70-80}	2.25	2.55
λ_{50-60}	1.90	2.20
λ_{0-40}	1.63	1.89

Table 2: ***Observed and 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 (.28)	.01 (.01)		.00 (.00)	.01 (.01)	
70-100	.49 (.49)	.03 (.04)		.01 (.01)	.07 (.06)	
50-100	.64 (.62)	.06 (.07)		.02 (.03)	.12 (.13)	
0-100	.71 (.70)	.09 (.10)	.20 (.20)	.04 (.05)	.22 (.23)	.73 (.73)
High Similarity						
90-100	.20 (.20)	.02 (.03)		.01 (.01)	.05 (.03)	
70-100	.38 (.37)	.09 (.10)		.03 (.03)	.14 (.13)	
50-100	.50 (.48)	.15 (.17)		.05 (.05)	.26 (.26)	
0-100	.59 (.56)	.21 (.24)	.19 (.20)	.08 (.08)	.40 (.41)	.53 (.51)

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

Figure 5: ***Receiver Operating Characteristic (ROC) Curves Depicting the Fit of the Absolute Model to High-Similarity and Low-Similarity Lineups***



Note. The operating points depict the empirical data, and the curves depict the predictions of the absolute model.

Critical Test #1: Do Low-Similarity Lineups Lead to More Correct Rejections than High-Similarity Lineups?

To test the hypothesis that variations in absolute signal strength would affect correct rejection rates, we fit a random intercepts probit regression model to the culprit-absent data. As predicted by the absolute model, the low-similarity lineup led to more correct rejections (73%) than did the high-similarity lineup (53%), $B = 0.80$, $SE = .09$, $z = 8.63$, $p < .001$. Next, we fit a random intercepts linear regression model to determine whether low-similarity rejections were made with higher confidence, on average, than were high-similarity rejections. As predicted by the absolute model, low-similarity rejections were made with greater confidence ($M = 73.79$) than were high-similarity rejections ($M = 64.72$), $B = 9.07$, $SE = 1.42$, $t(556.63) = 6.41$, $p < .001$.

Experiment 2: Comparing Rejection Rates for Fair and Biased Lineups

The results from Experiment 1 clearly contradict the predictions of the relative model. Increasing the similarity of lineup fillers to the culprit increased match-to-memory strength on culprit-absent lineups, leading to a decrease in correct rejections. That increasing the similarity between innocent persons and the culprit would lead to a decrease in correct rejection rates is hardly surprising. What is much more surprising is that the relative model rose to prominence in the identification literature despite predicting that correct rejection rates would be invariant to changes in absolute signal strength. These data refute any model that assumes witness decision-making is invariant to changes in absolute signal strength.

Conversely, the data from Experiment 1 were consistent with the predictions of the absolute model. Decreasing the similarity between the culprit and innocent persons increased correct rejections and the level of confidence that witnesses expressed in their rejection decisions. But this does not necessarily mean that the absolute model is superior to the relative model. A pure relative model is clearly unviable, but it is possible that a pure absolute model will also prove unviable. Perhaps the absolute model was unfairly advantaged on an initial test in which we manipulated absolute signal strength.

For Experiment 2, we contrasted the absolute and relative models with a manipulation of lineup composition bias. Lineup composition bias affects both relative and absolute strength. Because biased lineup fillers are less similar to the suspect than are fair lineup fillers, the expected BEST-REST score on a biased lineup exceeds the expected BEST-REST score on a fair lineup and the relative model predicts more rejections of fair lineups than of biased lineups. But because biased fillers will typically be weaker in absolute strength than fair fillers, the expected

strength of the MAX signal will typically be weaker on a biased lineup compared to a fair lineup and so the absolute model predicts more rejections of biased lineups than of fair lineups.

Methods

The preregistration, data analysis plan, anonymized data, and analysis code are available here: https://osf.io/t34pr/?view_only=75b9a275488b4b388f42101fece18efd (Smith et al., 2023). This experiment was declared exempt by the [redacted for review] Institutional Review Board. The primary purpose of Experiment 2 was to compare rejection rates between fair and biased lineups. However, we decided to collect enough data to have 80% power to detect a discriminability difference between fair and biased lineups. We conducted an a *priori* power analysis using ANOVA_Power Shiny Application (Lakens & Caldwell, 2021). We used a 2 x 2 ANOVA design as a proxy for estimating power. The *a priori* power analysis revealed that a sample size of approximately N = 400 was required to detect a small interaction effect between target-presence and lineup fairness (f = .02).

Participants

Participants were 436 MTurk workers (via CloudResearch; Litman et al., 2017) that were paid in exchange for their participation. One participant was excluded from the sample due to technical issues and 54 participants were excluded because they demonstrated no ability to discriminate between guilty persons and innocent persons. The final sample included a total of 381 participants. From those 381 participants, 58 trials (out of 6096) were dropped due to technological issues. From the final sample, 44% self-reported as female, 54% as male, 1% as non-binary, and 1% opted not to report. Participants were mostly White (75%), with 12% identifying as Black/African American, 5% identifying as Hispanic or Latino/a, 6% identifying

as Asian, 1% identifying as a different group, and 1% opting to not state their race. On average, participants were 40.81 years of age (*SD* = 11.74).

Design

We used a 2 (Target: Present vs. Absent) x 2 (Lineup: Fair vs. Biased) within-participants design. Participants were randomly assigned to one of eight stimulus groups. Each stimulus group viewed a fixed set of culprit-present/culprit-absent and biased/fair lineups (four from each cell in the 2 x 2 experimental design). Within each stimulus group, lineups were split into four blocks of four videos, one filler task, and four lineups. Conditions were randomized for each stimulus set across the entire survey rather than within each block, so that participants would not be able to use their judgments on earlier lineups to inform their decision making for subsequent decisions.

Culprit-present lineups were comprised of the culprit and five lineup fillers whom the witness had not seen before. Culprit-absent lineups were comprised of six lineup fillers whom the witness had not seen before. The fillers in fair lineups were relatively high in similarity to the culprit and the fillers in biased lineups were relatively low in similarity to the culprit. The innocent suspects on biased lineups were randomly sampled from the high-similarity filler pool. This sampling process was done anew for each participant. Fair culprit-absent lineups did not include designated innocent suspects. Instead, we estimated innocent-suspect identification rates by dividing the culprit-absent false alarm rate by the nominal lineup size (six). We describe the lineups in greater detail in the materials section.

Materials

Videos. Videos were sourced from the Eyewitness Lineup Identity database (Fitzgerald et al., 2023). We used a total of 32 simulated crime videos. Each video depicted a single individual

entering a room, stealing either a laptop or an iPad out of a bag, and then leaving. Video duration ranged from 23 seconds to 37 seconds. In each video there was a clear view of the culprit's face that lasted for approximately 10 seconds. Each participant viewed 16 lineup videos and completed 16 lineups. The 32 sets of stimuli that we used in Experiment 2 included the 16 sets from Experiment 1 plus 16 additional sets.

Lineups. All lineups were presented simultaneously and consisted of six faces presented to participants in a 2 rows x 3 columns photo array. The positions of these photos were randomized anew for each participant. We used culprit photos from the Eyewitness Lineup Identity database (Fitzgerald et al., 2023). All filler photos were from the Florida Mugshot Database. All lineup photos were edited to remove backgrounds and clothing (e.g., shirt collars) and culprit photographs were degraded so that they had similar resolution to that of filler photographs.

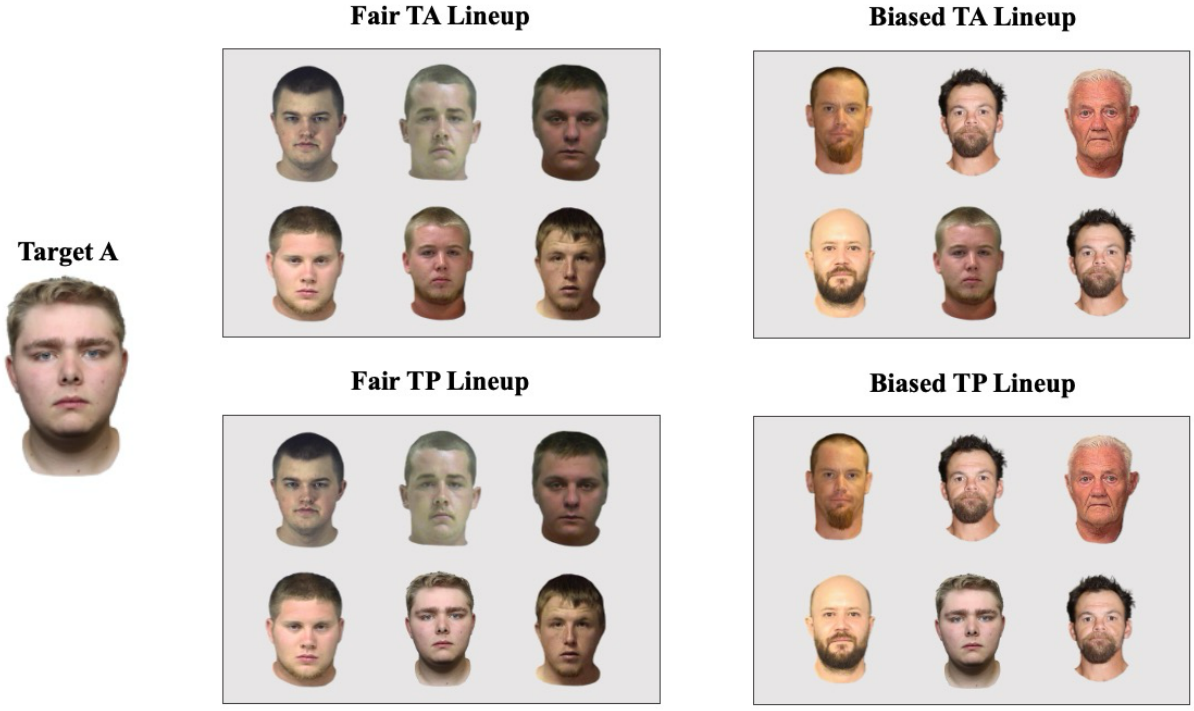
We used a total of 128 lineups, four versions for each of the 32 simulated-crime videos: fair culprit present, fair culprit absent, biased culprit present, and biased culprit absent. After selecting the 32 culprits, we then selected six fair fillers for each culprit. To that end, we instructed a team of research assistants to find six fillers that both matched the description of the culprit and who generally resembled the culprit. For the biased fillers, we told research assistants the race and sex of the culprit and had them find filler photographs based only on those criteria. The research team then reviewed their selections and removed any fillers that coincidentally resembled the culprit.

Fair culprit-absent lineups were comprised of six fair fillers. For fair culprit-present lineups Qualtrics randomly removed one of the six fillers and replaced that person with the culprit. Biased culprit-absent lineups were comprised of five biased fillers and one fair filler that

Qualtrics randomly selected for inclusion. Finally, biased culprit-present lineups were comprised of the culprit and the five biased fillers. All lineup photos were edited to remove backgrounds and clothing (e.g., shirt collars) and culprit photographs were degraded so that their resolution was similar to that of the filler photographs. Where applicable, the random selection of fillers for inclusion or exclusion on lineups was completed anew for each participant and each filler was included / excluded about equally as often as the next. Figure 6 provides examples of fair and biased culprit-present and culprit-absent lineups.

Figure 6

Examples of Fair and Biased Target-Present (TP) and Target-Absent (TA) Lineups



Procedure

We used the Qualtrics survey platform to facilitate the current study. Participants were only eligible if they: (1) were over 18 years of age, (2) lived in the United States, (3) were fluent in English, (4) had at least an 80% approval rate on at least 100 previous MTurk tasks, and (5) were using a computer or laptop to complete the task. Upon completing informed consent,

participants were asked if they agreed to pay attention and follow instructions throughout the study. If they declined, they were instructed that they were not eligible to participate. Subsequently, the participants completed two simple bot check questions to confirm their humanity, in which they were asked to select a specific letter from a list of multiple-choice options. Failure resulted in dismissal from the experiment.

The procedure was split into four blocks and the order in which blocks appeared was randomized across participants. Each block was comprised of four simulated crime videos, a filler task, and four lineup procedures. On each block, participants were provided with basic task instructions. They were then asked to pay careful attention to each of the four videos as they would be presented with lineups related to these videos at a later point. After watching the four simulated-crime videos, participants completed a one-minute anagram-solving task. After the anagram task participants were instructed that they would complete lineups for each of the four persons that they saw before the anagrams task. Prior to each lineup, participants were admonished that the person from the video may or may not be present in the lineup and were asked to identify that person if present and otherwise to select “Not Present”. Following each identification decision, participants were asked to express their level of confidence on a scale from 0% (not at all) to 100% (completely) in 10-point increments. Once they were done with both lineups, they were asked to report any technical difficulties they experienced with the videos or the lineups. At the end of each block, participants were asked to indicate whether they experienced any technological difficulties and if so, to explain as clearly as possible. After completing all sixteen lineups, participants completed some basic demographic questions. They were then thanked for their participation and debriefed.

To limit the likelihood that participants would be confused about what trace was being probed by which lineup, each block was comprised of four target persons that varied on physical characteristics to the extent that they were unlikely to be confused. For example, a given block of culprits might have included a White Male, a White Female, a Black Male, and a Hispanic Female. In addition, the order in which lineups were displayed was fixed to the same order as the simulated crime videos.

Results

As in Experiment 1, we started by fitting two versions of the absolute model to the data to determine whether we had produced the expected finding that fair lineups better discriminate guilty-suspect identifications from innocent-suspect identifications than do biased lineups (e.g., Clark, 2012; Lindsay & Wells, 1980; Smith et al., 2017; Smith et al., 2018) In the first model, we permitted discriminability to vary across fair and biased lineups and in the second model, we constrained discriminability to be equal across our manipulation of lineup composition bias. We then compared the relative fit of these two nested models. After establishing that our manipulation had its intended impact, we then present the results of our critical test assessing whether fair lineups led to more rejections than biased lineups.

Assessing the Impact of Lineup Composition Bias on Suspect-Identification Discriminability

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 (Akan et al., 2021; Smith et al., 2022). 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 fillers were drawn from a weaker strength distribution than the innocent suspect 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 absolute model included 11 free parameters: two location parameters, a correlation parameter, and eight decision criteria [four criteria for the biased lineups and four criteria for the fair lineups]. We assumed that all innocent persons on the fair lineups were drawn from a standard normal distribution ($\mu = 0, \sigma = 1$). Because the innocent suspects on biased lineups were drawn from the same population of faces as the fair fillers, we assumed that they were also drawn from the standard normal distribution ($\mu = 0, \sigma = 1$). We estimated the location of the culprit distribution (μ_{culprit}), but because we used the same culprits for fair and biased lineups, we did not permit the culprit location to vary across the two lineup types. Given that we selected as biased fillers, persons who we expected to provide weaker matches to memory than the fair fillers, we also permitted the location of the biased filler distribution to vary (μ_{Filler}). Finally, because all of the innocent persons included in the fair lineups were selected because they provided a good match to the description of the culprit, we expected that their match-to-memory values would be correlated, and we estimated the strength of that correlation (ρ) (Akan et al., 2021; Shen et al., 2022; Smith et al., 2022; Wixted et al., 2018). Hence, the unconstrained model included 11 free parameters and 17 degrees of freedom.

In Experiment 1, we assumed that the memory signals emanating from lineup members were uncorrelated because, by design, the low-similarity filler signals were as correlated with other low-similarity filler signals as the high-similarity filler signals were with other high-similarity filler signals. But in Experiment 2, we would expect higher signal correlations among fair fillers than we would among biased fillers, and we would also expect higher culprit-to-filler signal correlations in fair lineups and than in biased lineups. This is because, whereas the fair fillers were all selected because they matched the general appearance of the culprit, the biased fillers were selected more haphazardly. Anecdotally, it is evident from a cursory look at the sample stimuli in Figure 6 that there is much more variability in the appearances of biased fillers compared to the appearances of fair fillers (i.e., the signals are less strongly correlated). This is important because as the strength of correlations among signals increases, so too does the potential for lineups to discriminate between guilty-suspects and innocent-suspects. Accordingly, we permitted the memory signals for fair fillers to take on non-zero correlations and fixed the correlation parameters for biased lineup fillers to zero (Smith et al., 2022; see also Akan et al., 2021). The correlation parameter specified both the degree of signal correlations among fair fillers and the degree of correlation between any given fair filler and the culprit.

The best-fitting parameter estimates for the unconstrained model are summarized in Table 3 and Table 4 contrasts observed and predicted proportions. The fit between the unconstrained model and the data was less than optimal, $\chi^2(17) = 33.09, p = .01$. However, it is clear from Figure 7 that the predictions of the absolute model (depicted by the ROC curves) are capturing the major trends in the empirical data (depicted by the operating points). Likewise, Table 4 shows that the observed and predicted proportions are very similar. Critically, we replicated the typical finding that fair lineups better discriminate between guilty-suspect identifications and

innocent-suspect identifications than do biased lineups. This is evidenced by the fact that the fair lineup ROC curve dominated the biased lineup ROC curve over their common region in the ROC space. Given that we fixed the distance between the culprit and innocent-suspect distributions to be equivalent for fair and biased lineups, the observed difference might come across as surprising. The difference between the ROC curves depicted in Figure 7 is attributable to the fact that the memory signals on fair lineups are more strongly correlated with one another than are the memory signals on biased lineups. The result is that fair lineups are better able to discriminate between guilty-suspect identifications and innocent-suspect identifications than are biased lineups. When we constrained the signal correlations to be equivalent for fair and biased lineups, the model provided a poor absolute fit to the data, $\chi^2(18) = 50.83, p < .001$, and a significantly worse fit than the unconstrained model, $\chi^2(1) = 17.74, p < .001$. As in Experiment 1, we also fit two versions of the relative model to determine whether it also supported the conclusion that low-similarity lineups have better discriminability than high-similarity lineups. We refer the interested reader to supplemental materials.

Table 3: ***Best-Fitting Parameter Estimates of the MAX model to Fair and Biased Lineups***

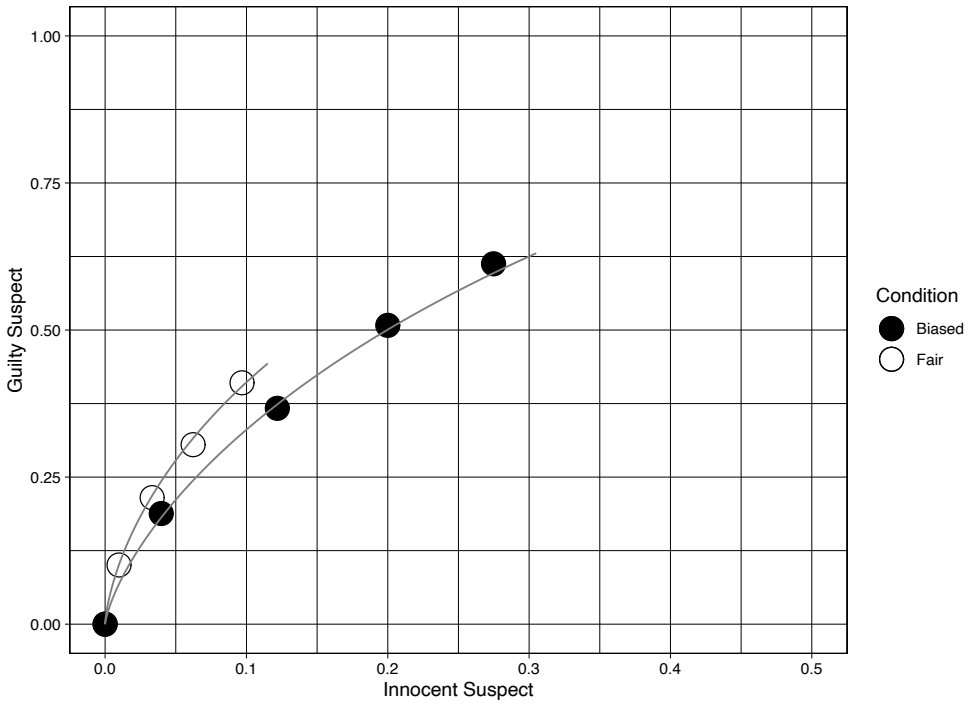
Parameter	Fair Lineup	Biased Lineup
μ_{culprit}	0.84	0.84
μ_{IS}	0_{fixed}	0_{fixed}
μ_{Filler}	0_{fixed}	-1.30
ρ	0.36	0_{fixed}
λ_{90-100}	1.88	1.71
λ_{70-80}	1.46	1.14
λ_{50-60}	1.18	0.78
λ_{0-40}	0.88	0.51

Table 4: ***Observed and 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 (.04)		.01 (.01)	.05 (.04)	
70-100	.22 (.22)	.11 (.14)		.03 (.03)	.17 (.15)	
50-100	.31 (.31)	.21 (.24)		.06 (.06)	.31 (.29)	
0-100	.41 (.40)	.33 (.36)	.26 (.25)	.10 (.10)	.48 (.48)	.42 (.42)
Biased						
90-100	.19 (.19)	.01 (.01)		.04 (.04)	.02 (.01)	
70-100	.37 (.38)	.04 (.03)		.12 (.13)	.06 (.03)	
50-100	.51 (.51)	.07 (.05)		.20 (.21)	.10 (.08)	
0-100	.61 (.60)	.10 (.09)	.29 (.31)	.27 (.29)	.14 (.13)	.59 (.58)

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

Figure 7: ***Receiver Operating Characteristic (ROC) Curves Depicting the Fit of the Absolute Model to Fair and Biased Lineups***



Note. The operating points depict the empirical data, and the curves depict the predictions of the absolute model.

Critical Test #2: Do Biased Lineups Lead to More Correct Rejections than Fair Lineups?

To test whether biased or fair lineups lead to more rejections, we fit a random intercepts probit regression model with a two-way interaction term between culprit-presence and lineup bias. As predicted by the absolute model, there was a two-way interaction between culprit presence and lineup bias, $B = .39$, $SE = .07$, $z = 5.41$, $p < .001$. Simple slope analyses revealed that biased lineups led to more rejections of culprit-absent lineups (59%) than did fair lineups (42%), $B = .50$, $SE = .05$, $z = 10.08$, $p < .001$. Biased lineups also led to more rejections from culprit-present lineups (29%) than did fair lineups (26%), however the effect size was smaller in the present condition compared to the absent condition, $B = .11$, $SE = .05$, $z = 2.13$, $p = .03$. Consistent with the predictions of the absolute model and inconsistent with the predictions of the relative model, biased lineups led to more rejections than fair lineups on both culprit-absent lineups and culprit-present lineups.

Next, we examined whether confidence ratings were consistent with the predictions of the absolute model or the relative model. To that end, we fit random-intercepts linear regression models separately to the confidence ratings associated with lineup rejections and suspect identifications. Starting with lineup rejections, the two-way interaction between culprit presence and lineup bias was marginally significant, $B = 3.34$, $SE = 1.78$, $t(2074.81) = 1.90$, $p = .06$. On average, biased culprit-absent lineups were rejected with greater confidence ($M = 69.56\%$) than were fair culprit-absent lineups ($M = 59.33\%$), $B = 10.22$, $SE = 1.05$, $t(2061.89) = 9.79$, $p < .001$. Likewise, biased culprit-present lineups were rejected with greater confidence ($M = 63.70$) than were fair lineups ($M = 56.82\%$), however the effect size was smaller in the present condition compared to the absent condition, $B = 6.88$, $SE = 1.42$, $t(2083.72) = 4.86$, $p < .001$. Once again,

the confidence ratings associated with lineup rejections were consistent with the predictions of the absolute model and inconsistent with the predictions of the relative model.

Finally, we examined confidence ratings for suspect identifications. The two-way interaction between culprit-presence and lineup bias was significant, $B = 3.72$, $SE = 1.30$, $t(3344.68) = 2.85$, $p = .004$. On average, innocent-suspect identifications were made with greater confidence from biased lineups ($M = 57.67\%$) compared to fair lineups ($M = 53.28\%$), $B = 4.40$, $SE = 1.02$, $t(3348.86) = 4.33$, $p < .001$. Culprit identifications were also made with greater confidence from biased lineups ($M = 67.07$) compared to fair lineups ($M = 58.96$), however the effect size was larger compared to innocent-suspect identifications, $B = 8.12$, $SE = 0.82$, $t(3342.87) = 9.89$, $p < .001$. Expressions of confidence following suspect identifications were consistent with the relative model and inconsistent with the absolute model.

General Discussion

The present work provides overwhelming evidence *against* the relative model and against any model that assumes witness decision-making is unaffected by changes in absolute signal strength. In the first of two experiments, we showed that increasing the absolute match-to-memory strength of culprit-absent lineup members decreased correct-rejection rates and confidence in correct rejections. Both patterns are consistent with the absolute model and inconsistent with the relative model. In the second experiment, witnesses were more likely to reject biased lineups than fair lineups. Witnesses were also more confident in their rejections of biased lineups compared to their rejections of fair lineups. Both patterns are consistent with the predictions of the absolute model and inconsistent with the predictions of the relative model. In fact, across both experiments there was only a single piece of evidence that was consistent with the predictions of the relative model and inconsistent with the predictions of the absolute model:

witnesses expressed greater confidence in suspect identifications from biased lineups compared to fair lineups. At the aggregate, the present work suggests that witnesses not only take absolute signal strength into account, but that witnesses prioritize absolute signal strength.

It would be difficult to overstate the implications of these findings for theoretical models of eyewitness decision-making. Over the past five years, the relative model (AKA the BEST-REST or ensemble model) has risen to prominence in the identification literature (e.g., Shen et al., 2023; Wixted et al., 2018). It has risen to prominence based on its ability to retrospectively fit suspect-identification data. But until now, few had considered what this model *predicts* in foresight. The relative model predicts that so long as relative signal strength is constant, variations in absolute signal strength will have no impact on witness decision-making. In Experiment 1, we manipulated the absolute signal strength of culprit-absent lineup members while holding relative signal strength constant and found that witnesses were more likely to reject a low-similarity culprit-absent lineup than a high-similarity culprit-absent lineup. This simple and intuitive pattern of results is at odds with any model that assumes absolute signal strength has no impact on witness decision-making. To bring the relative model in line with the data from Experiment 1, one would need to assume that when witnesses encountered high-similarity lineups, they lowered their criteria for making affirmative identification decisions. But by making that assumption one is tacitly conceding that witness decision-making is influenced by absolute signal strength which contradicts the most fundamental assumption of the relative model—namely, that witness decision-making is based only on relative strength.

The data from Experiment 2 also contradict the predictions of the relative model. The relative model predicts that fair lineups should lead to more rejections of both culprit-present and culprit-absent lineups when compared to biased lineups. Instead, as predicted by the absolute

model, results revealed that biased lineups led to more rejections on both culprit-absent and culprit-present lineups. Witness confidence was also higher for rejections of biased lineups compared to rejections of fair lineups, which is also the opposite of what the relative model predicts. To bring the relative model in line with the data from Experiment 2, one would need to assume that when witnesses encountered lineups where all members were relatively similar in strength, they respond by lowering their criterion for identification. But at that point the decision rule becomes self-contradictory. The relative model assumes that witnesses use the BEST-REST score as the bases for deciding whether to identify or reject. Witnesses interpret a relatively large BEST-REST score to imply that the BEST-matching lineup member is the culprit and a relatively small BEST-REST score to imply culprit absence. If we were to also assume that witnesses adopt more lenient criteria when the BEST-REST score is small compared to when the BEST-REST score is large, we would be assuming that a small BEST-REST score implies both culprit presence and culprit absence. After all, a witness would not respond to a piece of evidence by lowering their criterion unless they assumed that variable implied culprit presence. The observed data pattern in Experiment 2 is incompatible with a pure relative-judgment model.

There was one pattern that was consistent with the predictions of the relative model and inconsistent with the predictions of the absolute model—suspect-identification confidence was higher on biased lineups than on fair lineups (see also Charman et al., 2011; Horry & Brewer, 2016). This suggests that when rendering confidence ratings for suspect identifications, witnesses considered not only the signal strength of the suspect, but also the signal strengths of the lineup members that they did not identify, which is exactly what the relative model predicts (Shen et al., 2023; Wixted et al., 2018). The absolute model does not predict this pattern but could be made to fit this pattern retrospectively by assuming that witnesses lower their

confidence criteria when the discrepancy between the MAX signal strength and the remaining signal strengths is large. But that retrospective accommodation concedes that witness confidence ratings are influenced by relative signal strength.

Overall, the present work suggests that witness decision-making is driven primarily by absolute signal strength. Yet, the increased confidence in suspect identifications from biased lineups that we observed in Experiment 2 suggests that in at least some situations witnesses make use of relative signal strength (see also Charman et al., 2011; Horry & Brewer, 2016). This suggests that witnesses might make use of both absolute and relative signal strength (Clark, 2003; Clark et al., 2011). One possibility is that witnesses use absolute-judgment rules as part of their primary processing strategy and revert to relative-judgment strategies for more deliberative or effortful tasks like rendering confidence judgments. This is consistent with how some accumulator models distinguish between choices and confidence judgments. Accumulator models assume that as a respondent completes a choice task, evidence accumulates separately for each response option. Once the accumulated evidence for one of the response options passes the respondent's criterion, the respondent selects that option (Brown & Heathcote, 2008; Vickers, 1970). Confidence judgments are then post-computed by comparing the difference in accumulated evidence for chosen and unchosen response options (Smith & Vickers, 1988; Van Zandt, 2000; and see Horry & Brewer, 2016 for a thorough review). This suggests that choices are based on absolute evidence and confidence in those choices is based on relative evidence.

But not all confidence judgments are based on relative evidence. Confidence judgments that follow lineup rejections appear to be driven by absolute signal strength. Indeed, we found that rejections of biased lineups were made with higher confidence than were rejections of fair lineups (see also Horry & Brewer, 2016). Why would witnesses rely on relative signal strength

when expressing confidence judgments in affirmative responses, but rely on absolute strength when rendering confidence judgments in rejection responses? It is possible that witnesses use a different frame of reference when rendering confidence judgments for affirmative identifications versus lineup rejections (Brainerd et al., 2022; Horry & Brewer, 2016; Sakamoto & Miyoshi, 2024). Following an affirmative identification, what might be most salient to the witness are the photographs of the other lineup members that the witness could have identified but did not identify. But following a rejection response, what might be most salient to the witness is her mental representation for the culprit and the discrepancy between that mental representation and the MAX lineup member (Horry & Brewer, 2016; Smith et al., 2023). On this point it is noteworthy that the first formal model of eyewitness decision-making assumed that affirmative identifications were based on a combination of absolute and relative signal strength, but lineup rejections were driven entirely by absolute signal strength (Clark, 2003).

One could also imagine a witness starting the identification-making process with an absolute-judgment strategy and reverting to a relative-judgment strategy if the task proved difficult (e.g., Charman & Wells, 2007; Dunning & Perretta, 2002; Dunning & Stern, 1994). This might explain why some experiments manipulating lineup bias produce patterns that are more in line with the predictions of relative-judgment models (e.g., Colloff et al., 2016; Smith et al., 2022). As discriminability decreases, witnesses might become increasingly inclined to adopt relative-judgment strategies. Alternatively, there are different ways to create biased lineups and it is unclear that each of these variations impacts decision-making in the same way. We used as culprits, individuals who were relatively low in distinctiveness and then manipulated fillers to be as similar to the culprit as the innocent suspect (fair lineup) or less similar to the culprit than the innocent suspect (biased lineup). An alternative approach (for a different applied problem) is to

select culprits with distinctive features (e.g., a black eye), photoshop that same feature onto the innocent suspect and then manipulate whether fillers possess that feature or not (e.g., Colloff et al., 2016; Smith et al., 2022). The literature has not typically distinguished between these and other approaches to manipulating lineup bias, but it is unclear that they would impact decision-making in the same way. After all, if a witness views a lineup where one member has *the exact same black eye* as the culprit, should she not be able to infer that this person must be the culprit? If that person is not the culprit, then why would he have the exact same black eye? Given the tremendous variability in how a lineup can become biased, perhaps we should expect heterogeneity in how these manipulations impact decision strategies. In any case, there is no situation in which the relative model predicts more rejections of biased lineups than fair lineups and therefore the relative model cannot be reconciled with the data from Experiment 2.

Finally, it is worth considering the implications of this work in light of the distinction between psychological process models and mere measurement models. Signal detection theory is unique in that it is used as both a process model and as a measurement model (Macmillan, 1993). From a process-model perspective, the absolute and relative judgment models reflect assumptions about the psychological reality of how witnesses go about completing lineups. Conversely, the measurement-model perspective is agnostic about psychological reality and concerned only with the potential of the model to offer a mapping between observed data and latent psychological constructs. In other words, from a measurement-model perspective, one can fit a model to data without committing to the psychological reality of that model.

As a processing model, the absolute rule implies that the witness' decision to identify or reject is based on a simple comparison between the best-matching lineup member and her memory for the culprit. At least to us, that process seems highly plausible. In contrast, the

relative rule assumes that witnesses engage in the much more cumbersome process of computing a difference score that reflects the relative strength of the BEST-matching lineup member compared to the remaining lineup members (REST). It is unclear to us that witnesses would be both able and motivated to carry out this decision process, especially given the availability of the simpler and more intuitive absolute rule. The absolute rule appears higher in face validity than does the relative rule. Further, the data patterns observed in both Experiments 1 and 2 indicate that witnesses behaved as if they had adopted an absolute-judgment strategy and contrary to how they should have behaved if they had adopted a relative-judgment strategy.

Although the present data weigh against the validity of the relative-judgment model as a psychological processing model, one might wish to argue that the relative-judgment model remains a valid measurement model. But even if we assume the relative model is merely a measurement model and not a psychological processing model, the present results are still problematic. Indeed, the lineup outcomes that we observed in both Experiments 1 and 2 contradict the predictions of the relative model, which implies that the relative model is a flawed measurement device. Psychological realities aside, a measurement model is only as good as the predictions it makes. In both Experiments 1 and 2, the observed data contradicted the predictions of the relative model.

Conclusion

The results of two experiments demonstrated that absolute-judgment models better predict eyewitness decision-making than do relative-judgment models. In fact, the relative-judgment model failed two critical tests. Contrary to the predictions of the relative-judgment model, witnesses were more likely to reject low-similarity culprit-absent lineups than high-similarity culprit-absent lineups (Experiment 1) and more likely to reject biased lineups than fair

lineups (Experiment 2). These same patterns were also evident in witness confidence judgments. Both patterns are consistent with the predictions of the absolute-judgment model. In fact, there was only one aspect of the data that was inconsistent with the predictions of the absolute model: as predicted by the relative model, witnesses were more confident in suspect identifications from biased lineups compared to suspect identifications from fair lineups. It may be that witnesses rely primarily on absolute-judgment strategies but revert to relative-judgment strategies for more deliberative tasks, such as rendering expressions of confidence. One could also envision a witness that is struggling to complete a lineup task reverting to a relative-judgment strategy at the decision-making stage. Future models of eyewitness decision-making might allow for this sort of sequential blending of decision rules.

In addition to raising several questions for future research, the present experiments make two points abundantly clear. First, absolute-judgment models better predict eyewitness decision-making than do relative-judgment models. Second, the *predictions* of pure relative models are directly contradicted by some of the most fundamental data patterns in the identification literature. Contrary to the predictions of the relative-judgment model, witnesses were more likely to reject biased lineups than fair lineups. Likewise, the relative-judgment model's prediction that witnesses would be no more likely to reject low-similarity culprit-absent lineups than high-similarity culprit-absent lineups was also contradicted by the data. Any model that does not predict more correct rejections from a low-similarity culprit-absent lineup than from a high-similarity culprit-absent lineup cannot be considered viable. Future research will be required to determine whether accurately predicting performance on lineups requires models that permit witnesses to make use of both absolute and relative signal strength. In the meantime, the

experimental data appear to be much more in line with the predictions of absolute-judgment models than with the predictions of relative-judgment models.

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.

Allais, P. M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axioms de l'ecole americaine. *Econometrica, 21*(4), 503 – 546. https://doi.org/10.2307/1907921.

Ayala, N. T. & Smith, A. M. (2024). Predicting and postdicting eyewitness identification accuracy on forensic-object lineups. *Journal of Applied Research in Memory and Cognition.* Advance online publication. https://doi.org/10.1037/mac0000171

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

Birnbaum, M. H. (2011). Testing theories of risky decision making via critical tests. *Frontiers in psychology, 2,* 315. https://doi.org/10.3389/fpsyg.2011.00315.

Brainerd, C. J., Bialer, D. M., Chang, M., & Upadhyay, P. (2022). A fundamental asymmetry in human memory: Old ≠ not-new and new ≠ not-old. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(12), 1850 – 1867. https://doi.org/10.1037/xlm0001101.

Brown, S. D. & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153 – 178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Cervantes, V. H. & Benjamin, A. S. (2023). Models of unforced choice. *Unpublished preprint.*

Cha, J. & Dobbins, I. G. (2015). Critical tests of the continuous dual-process model of recognition. *Cognition, 215,* 104827. https://doi.org/10.1016/j.cognition.2021/104827

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior, 35*(6), 479 – 500. https://doi.org/10.1007/s10979-010-9261-1

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17,* 629 – 654. https://doi.org/10.1002/acp.891

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35*(5), 364–380. https://doi.org/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. https://doi.org/10.1177/0956797616655789

Dobbins, I. G. (2023). Recognition receiver operating characteristic asymmetry: Increased noise or information? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(2), 216 – 229. https://doi.org/10.1037/xlm0001224

Duncan, M. (2006). *A signal detection model of compound decision tasks.* (*Tech Note DRDC TR 2006 – 256).* Defence Research and Development Canada.

Dunning, D. & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identifications. *Journal of Applied Psychology, 87*(5), 951-962. https://doi.org/10.1037/0021-9010.87.5.951

Dunning, D. & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology, 67*(5), 818-835. https://doi.org/10.1037/0022-3514.67.5.818

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19,* 151 – 164. https://doi.org/10.1037/a0030618.

Fitzgerald, R. J., Rubinova, E., Ribbers, E., & Juncu, S. (2023, August). Initial testing of a stimulus database for eyewitness identification research. Paper presented at the 14th Biennial Meeting of the Society for Applied Research in Memory and Cognition, Nagoya, Japan.

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

Hanczakowski, M., Butowska, E., Beaman, C. P., Jones, D. M., & Zawadzka, K. (2021). The dissociations of confidence from accuracy in forced-choice recognition judgments. *Journal of Memory and Language, 117*, 104189. https://doi.org/10.1016/j.jml.2020.104189.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 598 – 615. https://doi.org/10.1037/0278-7393.10.4.598.

Horry, R. & Brewer, N. (2016). How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145*(12), 1615 – 1634. https://doi.org/10.1037/xge0000227.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138*(2), 291 – 306. https://doi.org/10.1037/a0015525.

Kellen, D. & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review, 122*(3), 542 – 557. https://doi.org/10.1037/a0039251.

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review, 128*(6), 1022-1050. https://doi.org/10.1037/rev0000288

Lakens, D. & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1), Article 2515245920951503. https://doi.org/10.1177/2515245920951503.

Lindsay, R. C. L. & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4,* 303 – 313. https://doi.org/10.1007/bf01040622.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavioral Research Methods, 49*(2), 433–442. http://dx.doi.org/10.3758/s13428-016-0727-z.

Ma, Q., Starns, J. J., & Kellen, D. (2022). Bias effects in a two-stage recognition paradigm: A challenge for "pure" threshold and signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(10), 1484-1506. https://doi.org/10.1037/xlm0001107

Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral science: Methodological issues* (pp. 21-57). Lawrence Erlbaum Associates, Inc.

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

Meyer-Grant, C. G. & Klauer, K. C. (2022). Disentangling different aspects of between-item similarity unveils evidence against the ensemble model of lineup memory. *Computational Brain & Behavior, 5,* 509 – 526. https://doi.org/10.1007/s42113-022-00135-4.

Palmer, M. A. & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36*(3), 247 – 255. https://doi.org/10.1037/h0093923.

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/.

Sakamoto, Y., & Miyoshi, K. (2024). A confidence framing effect: Flexible use of evidence in metacognitive monitoring. *Consciousness and Cognition, 118,* 103636. Advance online publication. https://doi.org/10.1016/j.concog.2024.103636.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137*(3)*,* 528–547. http://dx.doi.org/10.1037/a0012712.

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., Ying, R. C., Goldstein, A. R., & Fitzgerald, R. J. (2023). *Absolute and relative judgment models of eyewitness decision-making.* https://osf.io/t34pr/?view_only=75b9a275488b4b388f42101fece18efd.

Smith, A. M., Ayala, N. T., & Ying, R. C. (2023). The rule out procedure: A signal-detection-informed approach to the collection of eyewitness identification evidence. *Psychology, Public Policy, and Law, 29*(1), 19 – 31. https://doi.org/10.1037/law0000373.

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

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior, 41*(2), 127–145. http://dx.doi.org/10.1037/lhb0000219.

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. https://doi.org/10.1177/0956797617698528.

Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full Receiver

Operating Characteristic curves of lineup identification procedures. *Perspectives on Psychological Science, 15*(3)*,* 589–607. http://dx.doi.org/10.1177/174569162-9-2426

Smith, P. L. & Vickers D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematic Psychology, 32*(2), 135 – 168. https://doi.org/10.1016/0022-2496(88)90043-0.

Starns, J. J., Chen, T., & Staub, A. (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language, 93,* 55 – 66. https://doi.org/10.1016/j.jml.2016.09.001.

Starns, J. J., Cohen, A. L., & Rotello, C. M. (2023). A complete method for assessing the effectiveness of eyewitness identification procedures: Expected information gain. *Psychological Review, 130*(3), 677 - 719. http://dx.doi.org/10.1037/rev0000332.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 582 – 600. https://doi.org/10.1037/0278-7393.26.3.582.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13*(1), 37 – 58. https://doi.org/10.1080/00140137008931117.

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14*(2), 89 – 103. https://doi.org/10.1111/j.1559-1816.1984.tb0223.x

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*(5), 553 – 571. https://doi.org/10.1037/0003-066X.48.5.553

Wetmore, S. A., McAdoo, R. M., Gronlund, S. D., Neuschatz, J. S. (2017). The impact of fillers on lineup performance. *Cognitive Research: Principles and Implications, 2*(1), 48. https://doi.org/10.1186/s41235-017-0084-1

Wickens, T. D. (2002). *Elementary Signal Detection Theory.* Oxford University Press.

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

Windschitl, P. D. & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 198 – 215. https://doi.org/10.1037/0278-7393.30.1.198.

Wixted, J. T. & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition, 4*(4), 329 – 334. https://doi.org/10.1016/j.jarmac.2015.08.007.

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.

Yang, Y. & Burke, J. (2022). A multidimensional signal detection model for eyewitness identification [unpublished manuscript]. Department of Psychology, University of Nevada, Reno.

Zawadzka, K., Higham, P. A., Hanczakowski, M. (2017). Confidence in forced-choice recognition: What underlies the ratings? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 552 – 564. https://doi.org/10.1037/xlm0000321.

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).

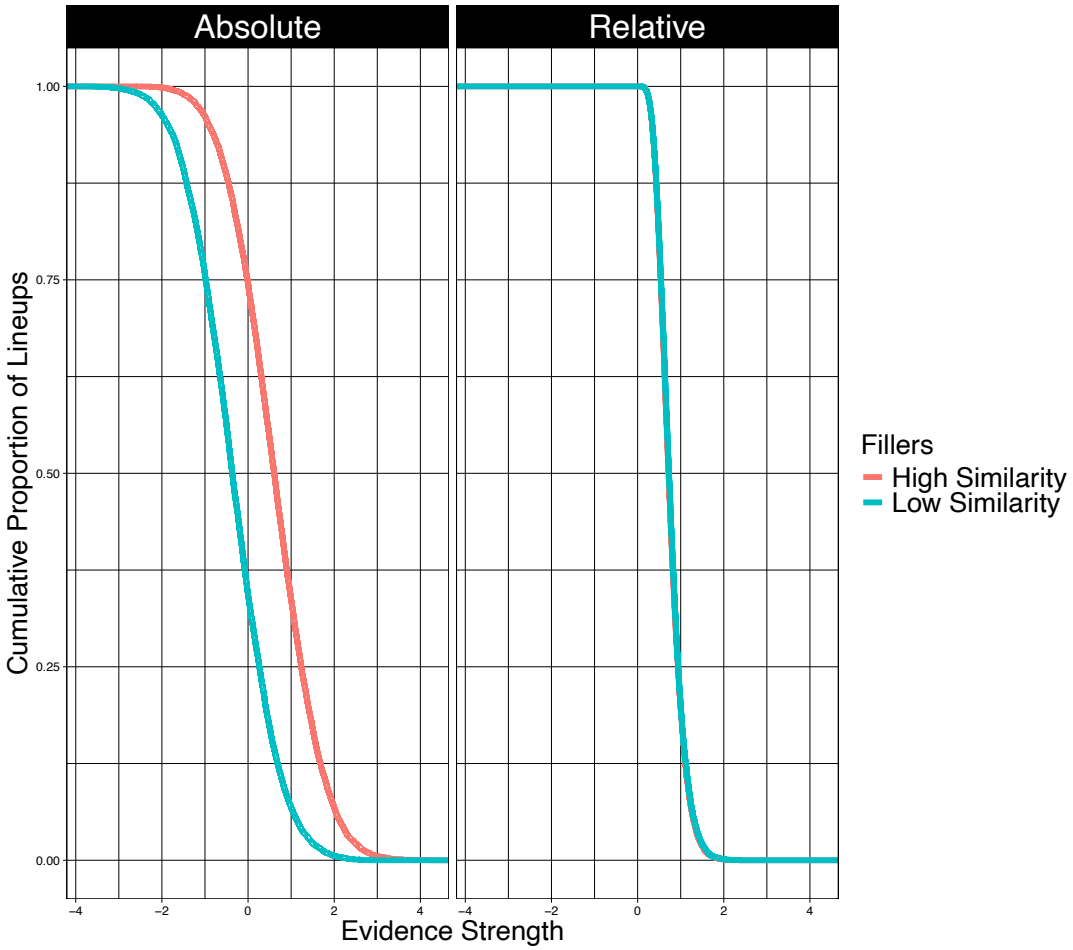
Rejection Rates for Low-Similarity Versus High-Similarity Culprit-Absent Lineups Across Various Levels of Signal Correlation (Experiment 1 Predictions). We examined what impact assuming correlated memory signals among lineup members had on the predictions of both the absolute (MAX) and relative (BEST-REST) 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***

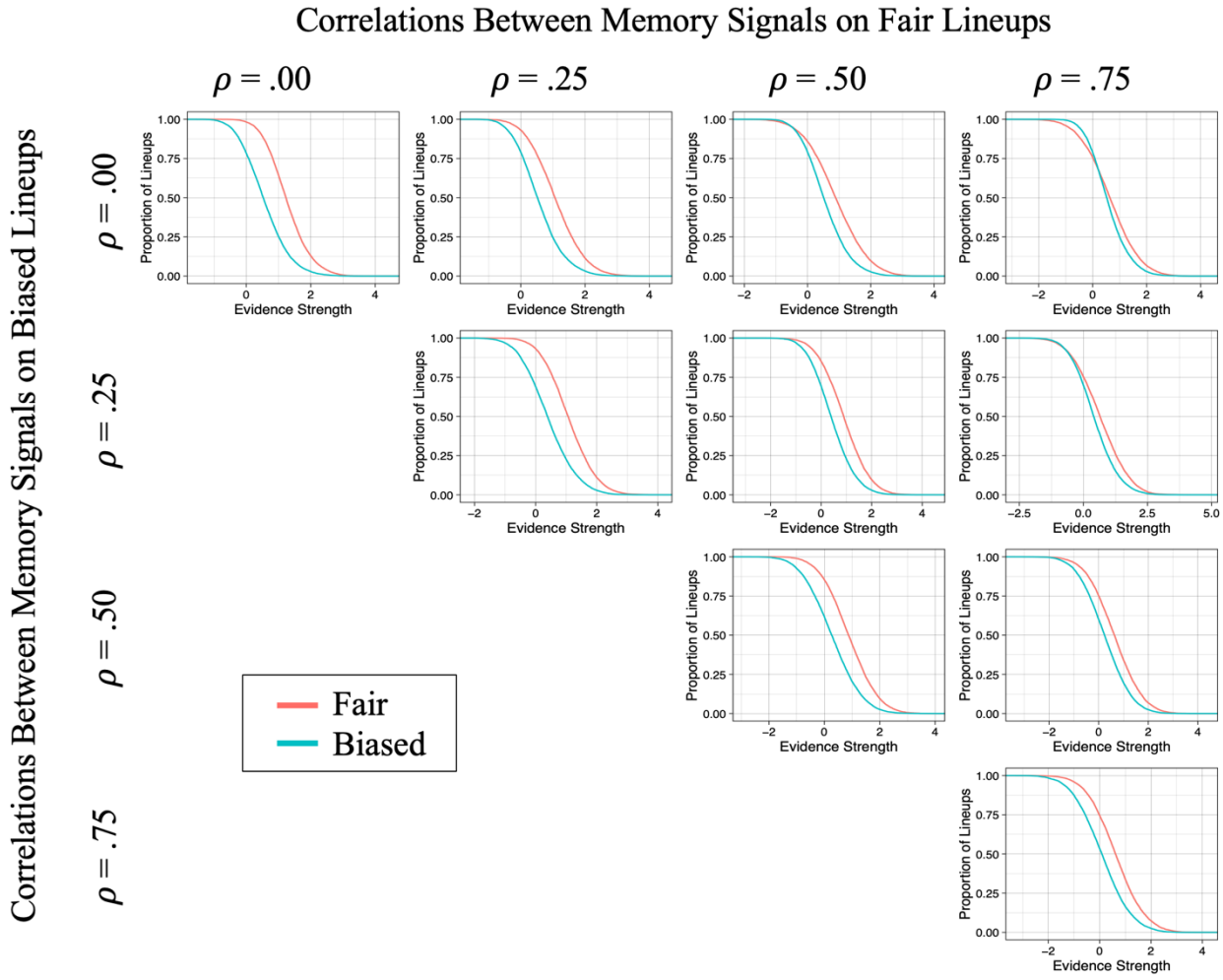


Figure S3 shows the predictions of the absolute model for culprit-present lineups. The absolute model consistently predicts either a slight increase in rejections of biased culprit-absent lineups compared to biased culprit-present lineups or no change in rejection rates.

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

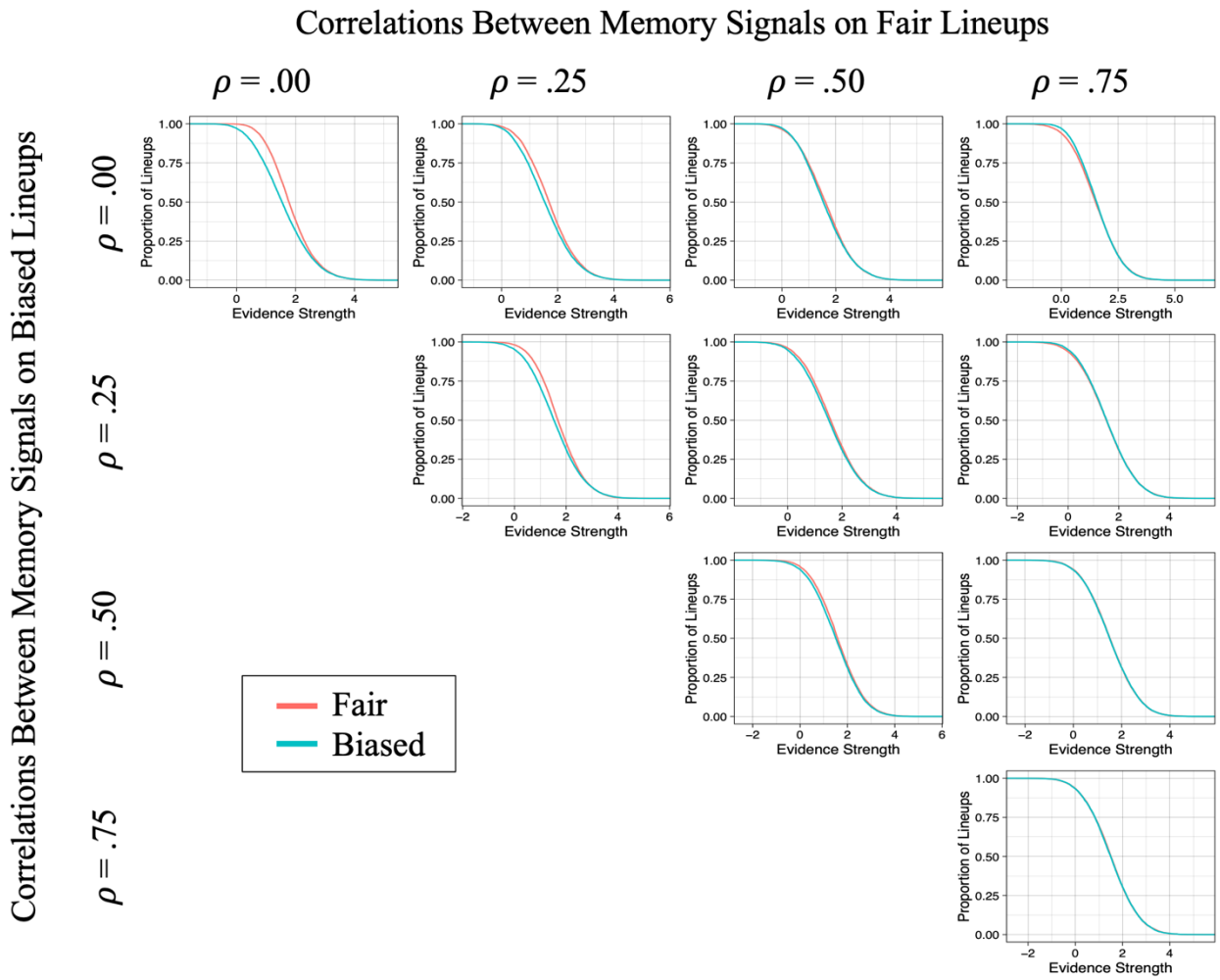


Figure S4 and Figure S5 show the predictions of the relative model for culprit-absent and culprit-present lineups, respectively. For both culprit-absent and culprit-present lineups, the relative model consistently predicts more rejections of fair lineups than biased lineups. This is evident from the 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***

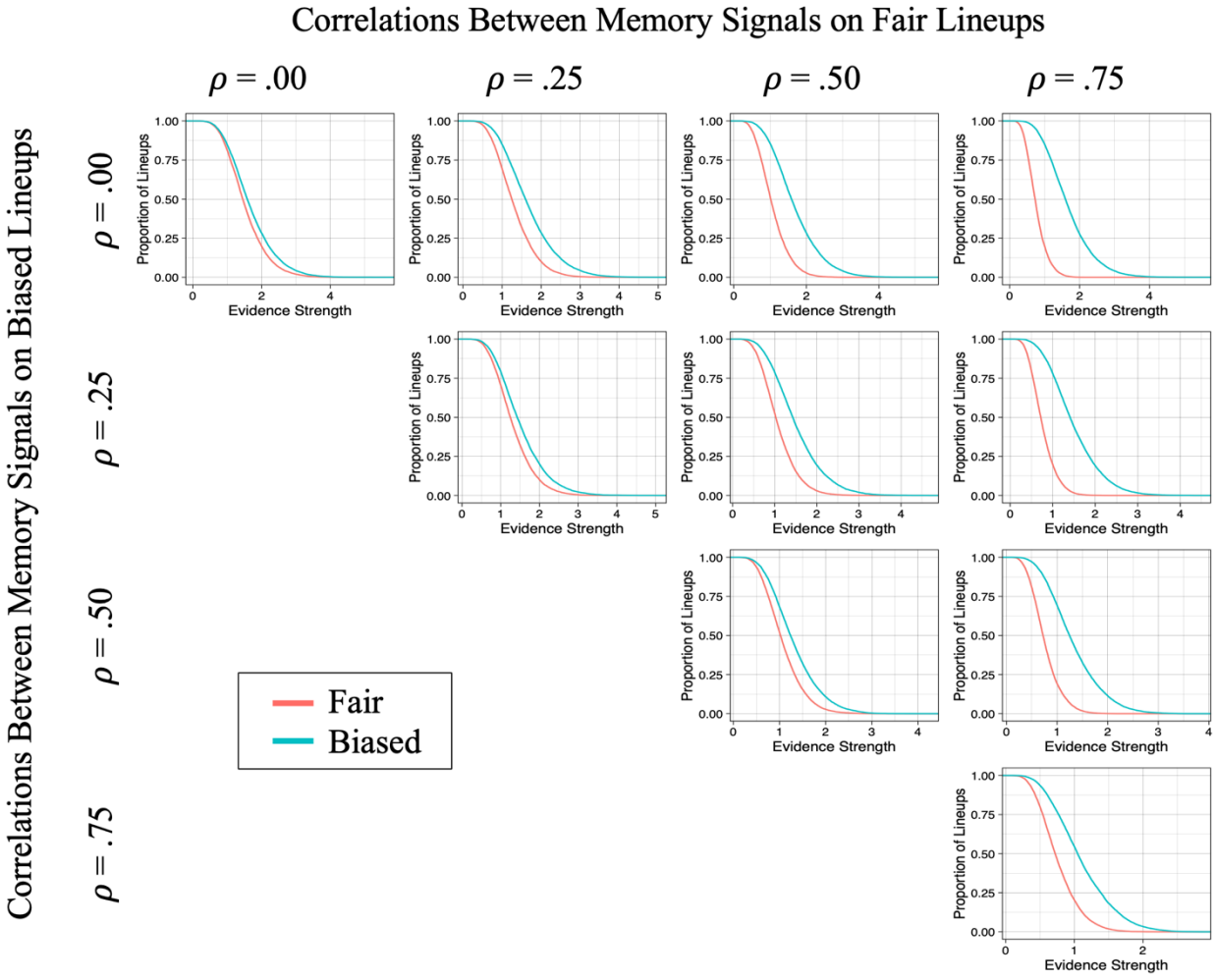
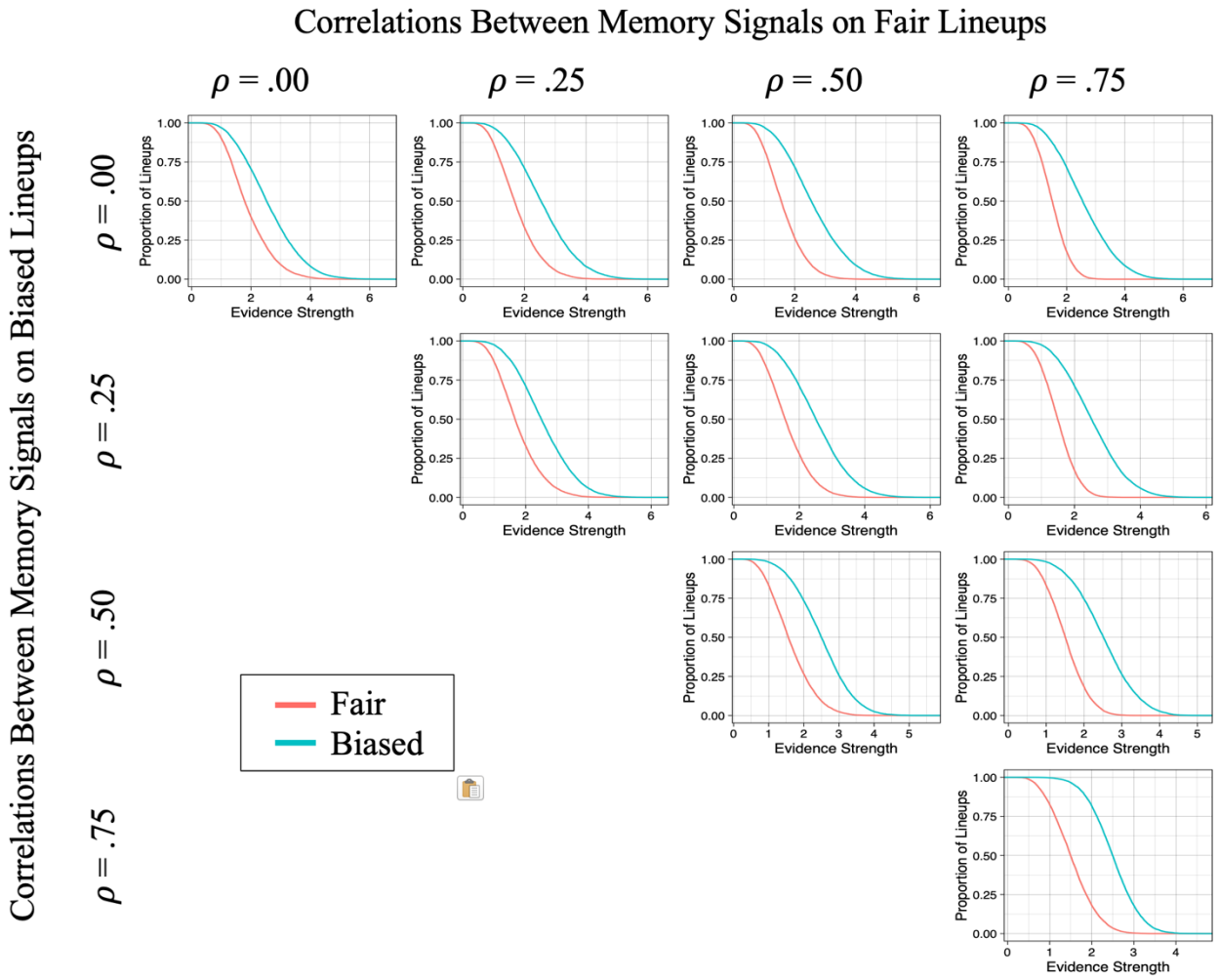


Figure S5: ***Evidence Strength Distributions Predicted by the Relative Model on Fair and Biased Culprit-Present Lineups***



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.