Optimizing the selection of fillers in police lineups

Article *in* Proceedings of the National Academy of Sciences · February 2021

DOI: 10.1073/pnas.2017292118

4 authors, including:

Melissa Colloff
University of Birmingham
37 PUBLICATIONS **278** CITATIONS
SEE PROFILE

Travis M Seale-Carlisle
Duke University
19 PUBLICATIONS **170** CITATIONS
SEE PROFILE

Section: Social Sciences, Psychological and Cognitive Sciences

Title: Optimizing the Selection of Fillers in Police Lineups

Authors: Melissa F. Colloff[a], Brent M. Wilson[b], Travis M. Seale-Carlisle[c], & John T. Wixted[b1]

[a]School of Psychology, University of Birmingham, U.K.
[b]University of California San Diego, Department of Psychology
[c]Wilson Center for Science and Justice, Duke University

[1]**Corresponding author**

John T. Wixted, PhD
Department of Psychology, University of California, San Diego, La Jolla, CA 92093
Email: jwixted@ucsd.edu
Phone: 858-534-3956

The authors declare no conflict of interest.

This paper has been published:

Colloff, M.F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups, PNAS, 118 (8), e2017292118. https://doi.org/10.1073/pnas.2017292118

Abstract

A typical police lineup contains a photo of one suspect (who is innocent in a target-absent lineup and guilty in a target-present lineup) plus photos of five or more fillers who are known to be innocent. To create a fair lineup in which the suspect does not stand out, two filler selection methods are commonly used. In the first, fillers are selected if they are similar in appearance to the suspect. In the second, fillers are selected if they possess facial features included in the witness's description of the culprit (e.g., "20-year-old White male"). The police sometimes use a combination of the two methods by preferentially selecting description-matched fillers whose appearance is also similar to that of the suspect in the lineup. Decades of prior research on which approach is better remains unsettled. Based on predictions made by a formal signal-detection-based feature-matching model, we tested a counterintuitive prediction: from a pool of acceptable description-matched photos, selecting fillers whose appearance is otherwise *dissimilar* to the suspect should increase the hit rate without affecting the false alarm rate (increasing discriminability). In Experiment 1, we confirmed this prediction using a standard mock-crime paradigm. In Experiment 2, the effect on discriminability was reversed (as also predicted by the model) when fillers were matched on similarity to the perpetrator in both target-present and target-absent lineups. These findings suggest that signal-detection theory offers a useful theoretical framework for understanding eyewitness identification.

Keywords: eyewitness identification; signal detection theory; filler similarity

Significance Statement

Eyewitness misidentifications have contributed to many wrongful convictions later overturned by DNA evidence. In response, many useful reforms have been introduced to protect the innocent. However, some police practices designed to protect the innocent also protect the guilty. We investigated a method for selecting fillers in a police lineup that protects the innocent while also making it easier to identify guilty suspects. Intuitively, fair lineups are created by choosing fillers who are similar in appearance to the suspect. However, we found that choosing fillers who match the description of the perpetrator but who are otherwise *dissimilar* to the suspect's appearance yield fair lineups and enhance eyewitness identification performance. It does so by imperilling the guilty while protecting the innocent.

Lineups are routinely administered to eyewitnesses globally to help determine whether a police suspect is the perpetrator of a crime. During a lineup test, a witness typically views photos of one suspect among photos of multiple "fillers" who physically resemble the suspect but are known to be innocent. The suspect may be guilty, in which case it is a "target-present" (TP) lineup, or may be innocent, in which case it is a "target-absent" (TA) lineup. The inclusion of fillers in a lineup offers protection to an innocent suspect when a witness is inclined to make a positive identification even when guessing. In that case, there is only a 1 / *k* chance of mistakenly identifying an innocent suspect, where *k* is the number of photos in the lineup.

A lineup offers protection to an innocent suspect only if the fillers are selected in such a way that the suspect does not stand out—that is, only if the lineup is fair (1). Traditionally, two methods have been used to create a fair lineup. The first method is to select fillers because they are judged by the investigating officer to be physically similar to the suspect (2). This is the most common method used by police in the U.S., and as many as one-third of U.S. police departments strive to ensure that fillers "look as much like the suspect as possible" (3). The second method is to select fillers who match the description of the perpetrator provided by the eyewitness (4,5) or, in the absence of an adequate description, to match on some basic default characteristics, such as race, gender, age, and facial hair (6). Using this approach, a filler need not look very similar to the suspect besides matching on the (usually small number of) features included in the witness's description.

Which approach is better? Despite decades of research (4-12), the answer remains unknown: "The net result of these complex problems is that the science has not yet been able to specify what the optimal level of similarity of fillers to the suspect ought to be and thus, at this time, there is no single strategy or formula for selecting fillers to be used in a lineup" (1, p. 18). In practice, researchers and police sometimes use a combination of these two methods

by first creating a pool of description-matched photos and then, from that pool, selecting fillers who are similar to the suspect (13, 14). Intuitively, this combined approach results in a lineup that is as fair as possible, but a longstanding concern is that choosing similar description-matched fillers will only serve to confuse the eyewitness (15).

The combined approach heavily emphasizes the protection of innocent suspects, but it also protects guilty suspects simply by making the task more difficult. Here, we investigate a counterintuitive alternative strategy—one grounded in a formal signal detection model—that simultaneously protects innocent suspects while imperilling guilty suspects. The alternative strategy is as follows: from a pool of acceptable description-matched photos, select fillers who are *dissimilar* to the suspect.

Lineup memory as a signal detection problem

Consider a highly simplified model that can be used to think through the effect of manipulating filler similarity on a witness's ability to discriminate innocent from guilty suspects. Suppose that a face is defined by $n = 20$ features (features $f_1 \rightarrow f_{20}$) and that each feature has 5 possible settings (i.e., $m = 5$). As an example, if feature 1 = race/ethnicity, the 5 possible settings for f_1 might be (1) = Caucasian, (2) = African American, (3) = Hispanic, (4) = Asian, and (5) = Pacific Islander. We consider only low-level physical features for simplicity, but higher-level feature conjunctions and even holistic signals could also be represented as features for modelling purposes.

After witnessing a crime, assume the witness has encoded all 20 features of the perpetrator's face. Because the guilty suspect (G) and the perpetrator (P) correspond to the same face, assume that the feature settings of the guilty suspect's face (features $G_1 \rightarrow G_{20}$) all match the settings of the corresponding features of the perpetrator's face ($P_1 \rightarrow P_{20}$) stored in memory. The number of matching feature settings between the guilty suspect's face and the memory of the perpetrator, n_{GP}, is therefore equal to n (i.e., $n_{GP} = n = 20$) in the simplest

case. By contrast, for fillers and innocent suspects, who are not guilty ($\hat{G}$), the number of features that match the corresponding settings in memory will be less than n (i.e., $n_{\hat{G}P} < n$).

Of the n encoded features of the perpetrator's face, some number of them will be included in the description of the perpetrator provided to the police (n_D). Assume that $n_D = 5$, corresponding to the settings of $P_1 \rightarrow P_5$ in memory. In a description-matched lineup, photos are selected for inclusion in a lineup precisely because they match these features in the witness's description. Therefore, the feature settings of everyone in the lineup will necessarily match the settings in memory for $P_1 \rightarrow P_5$.

Because the settings of features $f_1 \rightarrow f_5$ are shared by everyone in the lineup, these features are non-diagnostic of guilt. By contrast, features $f_6 \rightarrow f_{20}$ are potentially diagnostic because their settings for the guilty suspect's face ($G_6 \rightarrow G_{20}$) are more likely to match memory of the perpetrator ($P_6 \rightarrow P_{20}$) than the corresponding settings for the innocent suspect or fillers ($\hat{G}_6 \rightarrow \hat{G}_{20}$). Although these 15 settings for the guilty suspect's face match memory with probability 1.0, the corresponding settings for non-guilty innocent suspects and fillers match memory by chance alone. Because each feature has $m = 5$ possible settings, the probability of a chance match to the corresponding feature of the perpetrator's face in memory is $p = 1/m = 1/5 = .2$. Thus, assuming independence, $n_{\hat{G}P} = n_D + p(n - n_D) = 5 + .20(20 - 5) = 5 + 3 = 8$, on average. In other words, for the innocent suspect and the fillers, 8 of the 20 feature settings will match the corresponding features settings of the perpetrator in memory (5 by design, 3 by chance).

The overall memory-match signal for a given face is assumed to equal the sum of the memory-match signals generated by the 20 features. For convenience, the mean and variance of the memory signal generated by a matching feature are both set to 1, whereas the mean and variance of the memory signal generated by a mismatching feature are set to 0 and 1, respectively. Across many lineups, the mean of the summed memory signal for guilty

suspects would be 20, and, because variances sum, the standard deviation of the summed memory signal would be $\sqrt{20}$. For non-guilty lineup members, the mean of the summed memory signal would be 8. However, because variances sum whether or not the feature matches, the standard deviation would still be $\sqrt{20}$ (Fig. 1).

Manipulating filler similarity

Consider selecting fillers in a lineup from a pool of description-matched photos who also happen to look similar to the suspect. This involves selecting fillers who most resemble the guilty suspect in TP lineups, but two different ways of selecting similar fillers have been used for TA lineups: (1) selecting fillers who most resemble the innocent suspect, or (2) selecting fillers who most resemble the perpetrator. Previous filler-similarity experiments have often used the second approach even though the police are not in a position to do that (i.e., the police do not know what the perpetrator looks like when, unbeknownst to them, their suspect happens to be innocent). Nevertheless, this approach is useful for testing theoretical accounts of lineup memory. Here, we investigate the first method of manipulating filler similarity in TA lineups in Experiment 1 and the second in Experiment 2. The method used for TP lineups was the same for both experiments.

Filler similarity relative to the guilty suspect in TP Lineups. In a TP lineup, the faces in the pool of potential fillers (*F*) are already matched to the guilty suspect on the features that were included in the witness's description ($F_1 \rightarrow F_5 = G_1 \rightarrow G_5$). Thus, choosing high-similarity fillers from that pool involves choosing fillers whose remaining features ($F_6 \rightarrow F_{20}$) match some or all of the guilty suspect's feature settings that were *not* included in the witness's description ($G_6 \rightarrow G_{20}$). This will increase the number of matching features over and above those that already match due to chance. As a result, the overall memory-match signal generated by a similar TP filler ($\mu_{F:TP}$) will increase, thereby decreasing d'_{TP} (Fig. 2,

Exp.1). Thus, the hit rate should decrease because high-similarity fillers will compete with the guilty suspect (and be mistakenly identified) to a greater extent compared to when description-matched fillers are selected without regard to similarity.

The opposite effect on the hit rate is expected using the alternative strategy of selecting low-similarity fillers from a pool of description-matched photos (i.e., faces who appear dissimilar to the guilty suspect in a TP lineup). This approach decreases the probability that a diagnostic feature will match a feature of the filler's face, thereby increasing d'_{TP} (Fig. 2, Exp. 1). Thus, the hit rate should increase because fewer low-similarity fillers will compete with the guilty suspect compared to when description-matched fillers are selected without regard to similarity.

Filler similarity relative to the innocent suspect in TA Lineups. In Experiment 1, we manipulated filler similarity in TA lineups relative to the innocent suspect. The innocent suspect (*I*) and the potential fillers (*F*) are already matched to the suspect on the description-matched features. Thus, choosing similar fillers involves choosing fillers whose remaining features settings ($F_6 \rightarrow F_{20}$) match some or all of the innocent suspect's corresponding feature settings ($I_6 \rightarrow I_{20}$). The key intuition is that choosing fillers to be similar or dissimilar to the corresponding features of the innocent suspect should not affect how likely these remaining features will match the features of the memory of the perpetrator ($P_6 \rightarrow P_{20}$). Instead, the features that happen to coincidentally match the perpetrator will change, without changing the number that match (*SI Appendix*, Figs. S1 and S2). Thus, choosing a filler for a TA lineup who matches the description of the perpetrator but who is otherwise dissimilar to the innocent suspect should not affect the degree to which that filler matches the memory of the perpetrator. Because $\mu_{F\text{-}TA}$ would therefore remain constant across manipulations of filler similarity, it should still be the case that $d'_{TA} = 0$ (Fig. 2, Exp. 1). Thus, the false alarm rate should not vary as a function of filler similarity, consistent with prior results (10,16). Because

the hit rate should increase but the false alarm rate should remain constant as filler similarity decreases, the receiver operating characteristic (ROC) should reflect an improved ability to discriminate innocent from guilty suspects (17).

Filler similarity relative to the perpetrator in TA Lineups. In Experiment 2, we manipulated filler similarity in TA lineups relative to the perpetrator. Now, theoretically, decreasing filler similarity to the perpetrator should not only cause the guilty suspect in TP lineups to stand out in memory but should also cause the innocent suspect in TA lineups to stand out in memory (Fig. 2, Exp. 2). Thus, the model predicts that the low-similarity condition will be associated with both an increased hit rate and an increased false alarm rate.

In a conceptually related study (18), the perpetrator in the crime video had a distinctive feature (a black eye). This feature was always present on both the guilty suspect in TP lineups and the innocent suspect in TA lineups. In the high-similarity (fair) condition, all fillers shared that feature, but in the low-similarity (unfair) condition, none did. The hit rate and false alarm rate were both higher—and discriminability was lower—in the low-similarity condition. Theoretically, the discriminability advantage in the high-similarity condition occurred because witnesses in that condition discounted the black eye that was shared by everyone in the lineup. Relying on a non-diagnostic feature adds nothing but noise to the memory signal, reducing discriminability (19).

Experiment 2 here also involved fillers who were lower in similarity to the perpetrator than the suspect was, even in TA lineups. Therefore, not only should the hit and false alarm rates be highest in that condition, if participants discount shared (i.e., non-diagnostic) features, the pattern of discriminability across filler-similarity conditions should be the reverse of that predicted for Experiment 1 (*SI Appendix,* Figs. S3 and S4).

Results

For both experiments, we first analysed the hit and false alarm rate data. The hit rate is the proportion of TP lineups resulting in a correct ID of the guilty suspect, and the false alarm rate is the proportion of TA lineups resulting in an incorrect ID of the innocent suspect. The data were similar across replications, so we present the results aggregated over replications for both experiments (see *SI Appendix*, Tables S1-4 and Fig. S5 for each experiment analysed individually). All our data are available (https://osf.io/uzk48/; https://osf.io/c36bf/).

The trends in the hit and false alarm rates (Fig. 3) correspond to the predictions made by the feature-matching model presented earlier (Fig. 2). That is, when filler-similarity was manipulated relative to the suspect in TP and TA lineups (Experiment 1), the hit rate increased as filler similarity decreased, but there is no apparent trend in the corresponding false alarm rate data. By contrast, again consistent with the feature-matching model (Fig. 2), when filler-similarity was manipulated relative to the perpetrator in both TP and TA lineups (Experiment 2), the hit rate and the false alarm rate both increased as filler similarity decreased.

The confidence-based identification ROC curves for the low-, medium- and high-similarity conditions also exhibited the predicted trends (Fig. 4). These are partial ROCs because the maximum false alarm rate for a lineup is less than 1.0 (see 20, 21). Overall, the data suggest that when choosing fillers from a pool of description-matched photos, discriminability is enhanced by choosing dissimilar fillers (Fig. 4, Exp. 1), but the opposite result is obtained when fillers for both TP and TA lineups are selected based on similarity to the perpetrator (Fig. 4, Exp. 2).

Discussion

To create a fair police lineup, the fillers need to be similar to the suspect, but if they are too similar, the lineup task becomes impossibly difficult (4). Yet the police often choose

fillers based on similarity to the suspect, which raises a question that has bedevilled the field for decades: what is the optimal level of similarity (12)? Most prior work on this question has not been guided by formal models. Indeed, with a few notable exceptions (e.g., 22,23), efforts to improve lineups have been largely untethered to what basic scientists have learned about memory, perception, and decision-making (24,25). Here, using a feature-matching model of face memory in conjunction with signal detection theory, we investigated a counterintuitive strategy that was predicted to yield a favourable outcome: from a pool of description-matched photos, choose fillers who are dissimilar to (not similar to) the suspect.

The use of dissimilar fillers in Experiment 1 increased the hit rate without affecting the false alarm rate (Fig. 3), thereby increasing the ability of witnesses to discriminate innocent from guilty suspects (Fig. 4). By contrast, when fillers for both TP and TA lineups were dissimilar to the perpetrator in Experiment 2, the false alarm rate instead increased and the observed effect on discriminability was reversed. This reversal, which was predicted by diagnostic feature-detection theory (19), reinforces the results of related studies that manipulated similarity using distinctive features (18, 26).

While the results of Experiment 2 are theoretically informative, the results of Experiment 1 are more pertinent to police practices. Our results suggest that choosing dissimilar fillers from a pool of acceptable description-matched photos, the hit rate can be increased by ~10% while leaving the false alarm rate largely unchanged. However, our investigation is a first step and does not have immediate policy implications. For example, we used the median-similarity filler as our innocent suspect because our model-based simulations suggest that the results would be representative of results obtained using a wide range of similarities. More specifically, when the innocent suspect happens to be similar to the perpetrator (an innocent lookalike), the use of low-similarity fillers should increase the false alarm rate, and when the innocent suspect happens to be dissimilar to the perpetrator,

the use of low-similarity fillers should decrease the false alarm rate (*SI Appendix*, Fig. S6). Overall, the risk to innocent suspects should remain unchanged, as it was here in Experiment 1 using the median-similarity filler (Fig. 3). Whether these predictions are confirmed by future research remains to be seen.

The increased risk to innocent lookalikes when low-similarity fillers are used sounds alarming, but that effect should be observed no matter how discriminability is enhanced, such as conducting lineups in bright rather than dim light (24). Using bright light, the guilty suspect in a TP lineup will stand out from the fillers as providing the best match to memory of the perpetrator, but the innocent lookalike in a TA lineup will also stand out from the fillers for the same reason. Even so, no one would advocate routinely conducting lineups in dim light to protect rare lookalikes.

Although our findings do not have immediate implications for real-world policy, they do have immediate implications for basic and applied scientists. Specifically, for the first time in the long history of research on this topic, our signal-detection-based model provides guidance concerning the optimal selection of fillers for a police lineup. From a broader perspective, our research is an example of how basic (theory-driven) memory research can be put to effective use in tackling important applied questions (25).

Materials and Methods

Design. For both experiments, we used a 3 (suspect-filler similarity: low, medium, high) × 2 (target: present, absent) between-subjects design. Our data-collection stopping rule was to recruit at least 3,000 participants, 500 in each of the between-subject conditions. This pre-planned sample size yielded sufficient power to detect predicted trends in hit rates and false alarms rates, but when broken down by confidence, the data were too noisy to conduct informative ROC analyses. We therefore directly replicated both experiments twice each and

analysed the collapsed data. The research was approved by the University of California, San Diego Institutional Review Board for research involving human subjects, and all participants provided informed consent prior to participation.

Participants. For Experiment 1, we recruited 3,877 (Experiment 1), 3,395 (Replication 1), and 3,530 (Replication 2) from Amazon Mechanical Turk who completed the study for 50 cents. For Experiment 2, we recruited 3,425 (Experiment 2), 2,561(Replication 1), and 3,520 (Replication 2). We excluded participants who incorrectly answered an attention check question about the number of people in the video, yielding final samples of 3,778, 3,344, and 3,437, respectively for Experiment 1 (combined $N = 10,559$); and 3,331, 2,496, and 3,346 for Experiment 2 (combined $N = 9,173$).

Materials. We used a mock-crime video depicting a white male perpetrator stealing a laptop from an office and then created a pool of 328 description-matched fillers from an initially larger pool, eliminating photos depicting individuals who did not fit the description, who had prominent distinctive features like scars, bruises, or tattoos, or were not facing the camera. The median-similar filler was then selected to serve as the designated innocent suspect in target-absent lineups. We then asked Amazon Mechanical Turk participants to rate the similarity of the remaining 327 fillers to the perpetrator and, separately, to the innocent suspect (*SI Appendix*, Fig. S7). For Experiment 1, we then divided the fillers into three sets of 109 fillers that had high, medium, or low similarity to the perpetrator (target-present filler group) and into three sets of 109 fillers that had high, medium, or low similarity to the innocent suspect (target-absent filler group). For Experiment 2, we used the target-present filler group for both TP and TA lineups (*SI Appendix*, Fig. S8).

Procedure. Participants first watched the mock-crime video. Next, participants saw a lineup composed of two rows of three photos; the photos displayed depended on to which of the six experimental condition the participant had been randomly assigned. In TP lineups, the

perpetrator was presented alongside five fillers selected randomly from the pool of low-, medium-, or high- similarity target-present filler group. In TA lineups, the innocent suspect was presented alongside five fillers who were selected at random from either the low-, medium-, or high-similarity target-absent filler group (Experiment 1) or from the low-, medium-, or high-similarity target-present filler group (Experiment 2). Participants were asked to make an identification by clicking on either the person they believed to be the perpetrator or on an option underneath the lineup labelled "Not Present" and to then provide a confidence rating using a 11-point scale. Additional details of experimental methods are available in *SI Appendix*.

Acknowledgments

This work was supported by the College of Life and Environmental Sciences, University of Birmingham (to M.F.C), Laura and John Arnold Foundation (to J.T.W).

References

1. G. L. Wells, M.B. Kovera, A. B. Douglass, N. Brewer, C. A. Meissner, J. T. Wixted, Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law. Hum. Behav.* **44**, 3-36 (2020).

2. M. S. Wogalter, R. S. Malpass, D. E. McQuiston, A national survey of U.S. police on preparation and conduct of identification lineups. *Psychol. Crime. Law.* **10**, 69 – 82 (2004).

3. Police Executive Research Forum, A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies (2013). Retrieved October 14, 2020, from http://www.policeforum.org/

4. C. A. E. Luus, G. L. Wells, Eyewitness identification and the selection of distracters for lineups. *Law. Hum. Behav.* **15**, 43–57 (1991).

5. G. L. Wells, S. M. Rydell, E. P. Seelau, The selection of distractors for eyewitness lineups. *J. Appl. Psychol.* **78**, 835–844 (1993).

6. R. C. L. Lindsay, R. Martin, L. Webber, Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy. *Law. Hum. Behav.* **18**, 527–541 (1994).

7. C.A. Carlson, A.R. Jones, J. E. Whittington, et al. Lineup fairness: propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cogn. Res. Princ. Implic.* **4**, 20 (2019). https://doi.org/10.1186/s41235-019-0172-5

8. S. Darling, T. Valentine, A. Memon, Selection of lineup foils in operational contexts. *Appl. Cogn. Psychol.* **22**, 159–169 (2008).

9. R. J. Fitzgerald, H. L. Price, C. Oriet, S. D. Charman, The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychol. Pub. Pol. Law.* **19**, 151–164 (2013).

10. R. J., Fitzgerald, C. Oriet, H. L. Price, Suspect filler similarity in eyewitness lineups: a literature review and a novel methodology. *Law. Hum. Behav.* **39**, 62–74 (2015).

11. P. Juslin, N. Olsson, A. Winman, Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 1304-1316 (1996).

12. J. L. Tunnicliff, S. E. Clark, Selecting foils for identification lineups: Matching suspects or descriptions? *Law. Hum. Behav.* **24**, 231–258 (2000).

13. R. S. Malpass, C. G. Tredoux, D. McQuiston-Surrett, Lineup construction and lineup fairness. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol. 2. Memory for people* (p. 155–178). Lawrence Erlbaum Associates Publishers (2007).

14. R. J. Fitzgerald, E. Rubínová, S. Junca, Eyewitness identification around the world. In A. M. Smith, M. P. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks*. Taylor and Francis (in press).

15. Wells, G. L. What do we know about eyewitness identification? *Am. Psy*. **48**, 553-571 (1993).

16. Oriet, C., Fitzgerald, R. J. The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law. Hum. Behav.* **42**, 1-12 (2018).

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

18. M. F. Colloff, K. A. Wade, & D. Strange Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychol. Sci.* **27***,* 1227–1239 (2016).

19. J. T. Wixted, L. Mickes, A signal-detection-based diagnostic feature-detection model of eyewitness identification. *Psychol. Rev.* **121**, 262-276 (2014).

20. S. D. Gronlund, J. T. Wixted, L. Mickes, Evaluating eyewitness identification procedures using ROC analysis. *Curr. Dir. Psychol. Sci.* **23**, 3-10 (2014).

21. L. Mickes, H. D. Flowe, J. T. Wixted, Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *J. Exp. Psychol. Appl.* **18**, 361-376 (2012).

22. S. E. Clark, A memory and decision model for eyewitness identification. *Appl. Cogn. Psychol.* **17**, 629–654 (2003).

23. S. Gepshtein, Y. Wang, F. He, D. Diep, T. D. Albright, A perceptual scaling approach to eyewitness identification. *Nat. Commun.* **11**, 3380 (2020).

24. National Research Council, *Identifying the Culprit: Assessing Eyewitness Identification*. Washington, DC: The National Academies Press (2014).

25. T. D. Albright, J. S. Rakoff, The impact of the National Academy of Sciences report on eyewitness identification. *Judicature* **104**, 21-29 (2020).

26. M. F. Colloff, K. A. Wade, D. Strange, J. T. Wixted, Filler-Siphoning Theory Does Not Predict the Effect of Lineup Fairness on the Ability to Discriminate Innocent From Guilty Suspects: Reply to Smith, Wells, Smalarz, and Lampinen (2018). *Psychol. Sci.* **29**, 1552-1557 (2018).

Figure Legends

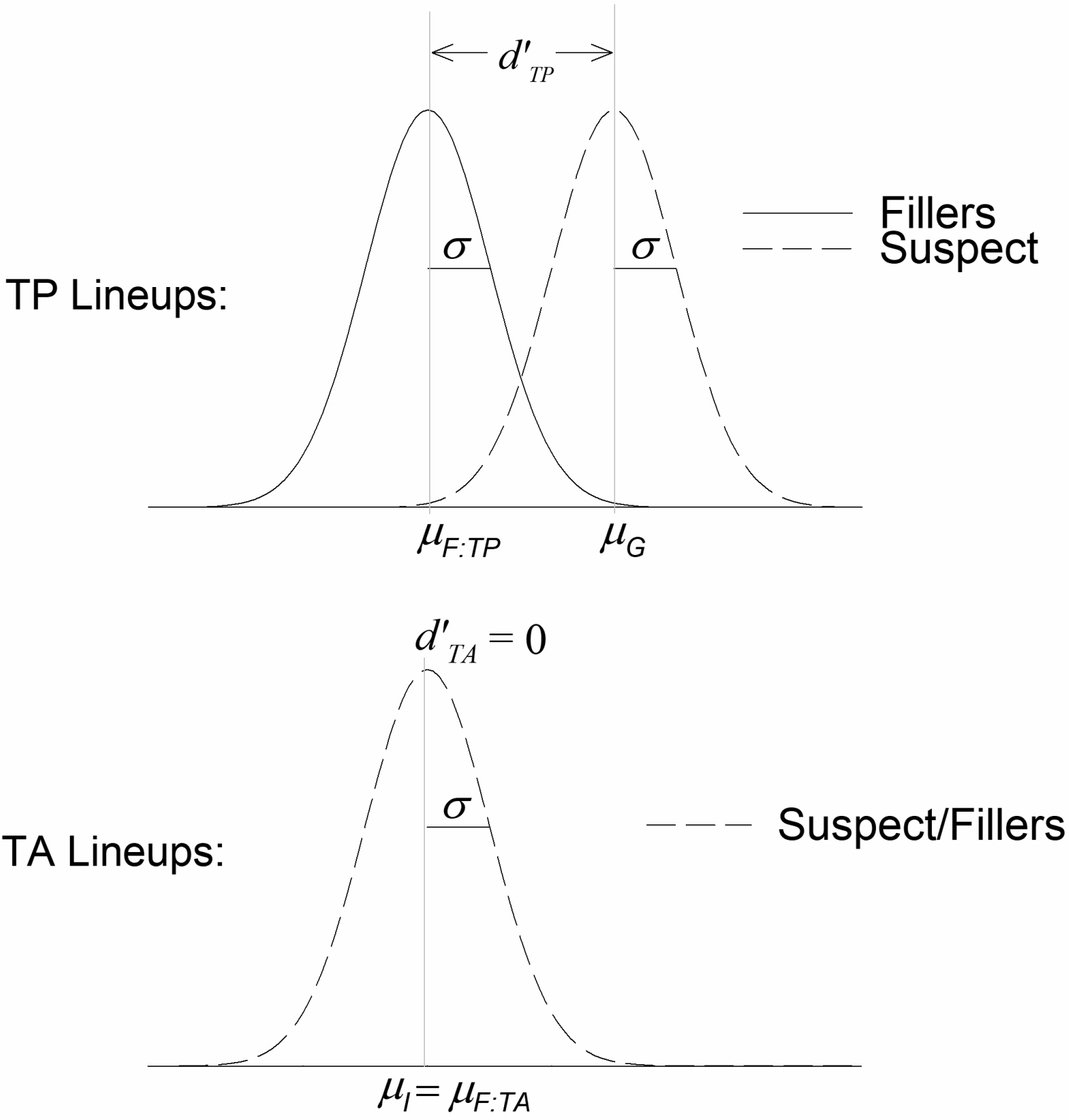
Figure 1. d'_{TP} is the difference between the mean of the TP filler distribution (e.g., $\mu_{F:TP} = 8$) and the guilty suspect distribution (e.g., $\mu_G = 20$) in standard deviation units (e.g., $\sigma = \sqrt{20}$). Here, $d'_{TP} = \frac{\mu_G - \mu_{F:TP}}{\sigma} = \frac{20-8}{\sqrt{20}} = 2.68$. d'_{TA} is the standardized difference between the TA filler distribution (e.g., $\mu_{F:TA} = 8$) and the innocent suspect distribution (e.g., $\mu_I = 8$). Because $\mu_I = \mu_{F:TA}$, $d'_{TA} = 0$. The witness's decision is theoretically based on a criterion (not shown). If the face that generates the strongest memory signal exceeds the criterion, it is identified. Otherwise, the lineup is rejected.

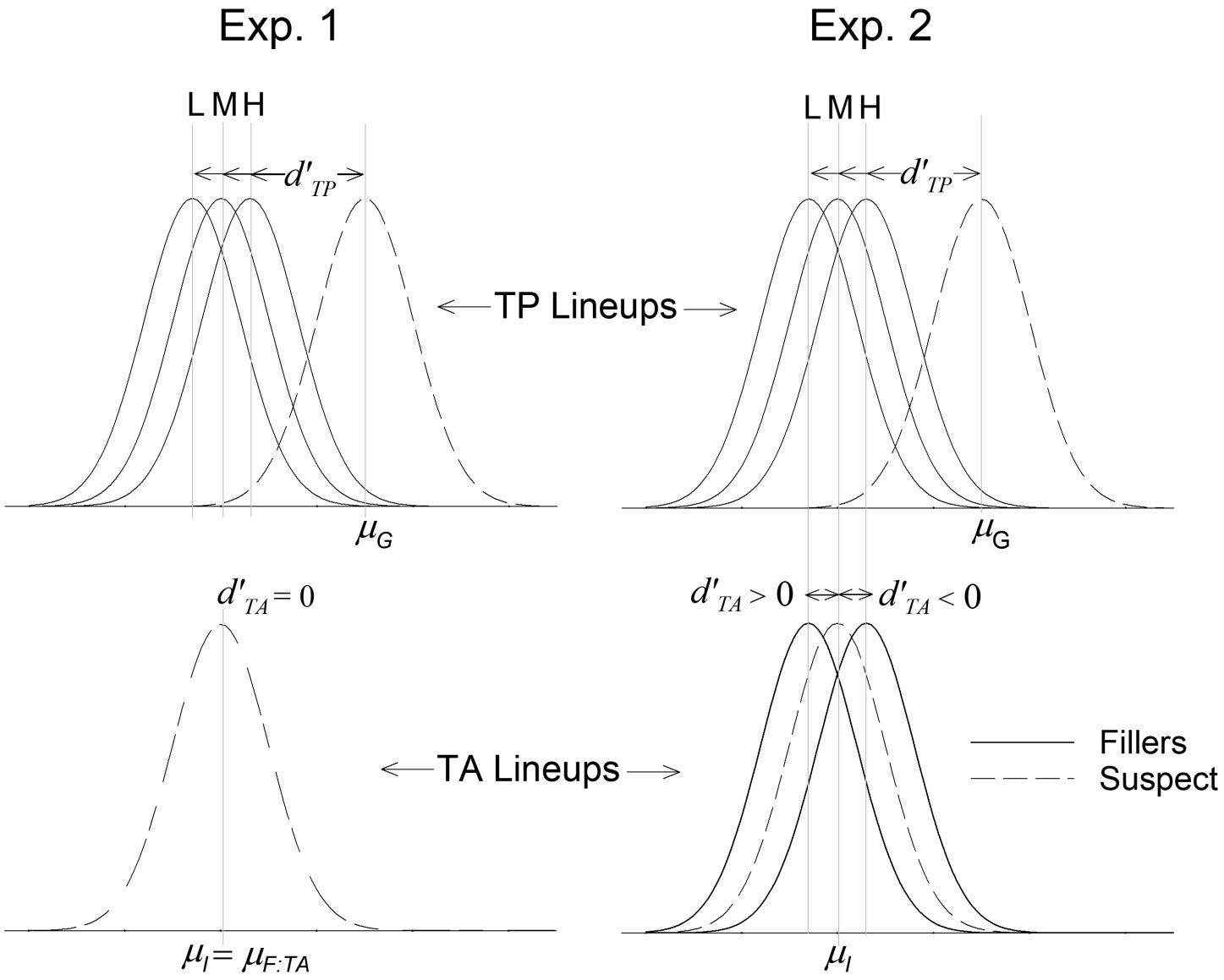
Figure 2. Exp. 1: d'_{TP} increases as filler similarity to the suspect varies from high (H) to medium (M) to low (L). By contrast, d'_{TA} theoretically remains equal to 0 because varying filler similarity to the innocent suspect in a TA lineup should not affect the degree to which those fillers match memory of the perpetrator. Exp. 2: In a TP lineup, the situation is identical to Exp. 1. However, when filler similarity is varied with respect to the perpetrator in a TA lineup, d'_{TA} should now vary with filler similarity in such a way that the innocent suspect stands out when low-similarity fillers are used ($d'_{TA} > 0$) and should be protected when high-similarity fillers are used ($d'_{TA} < 0$).

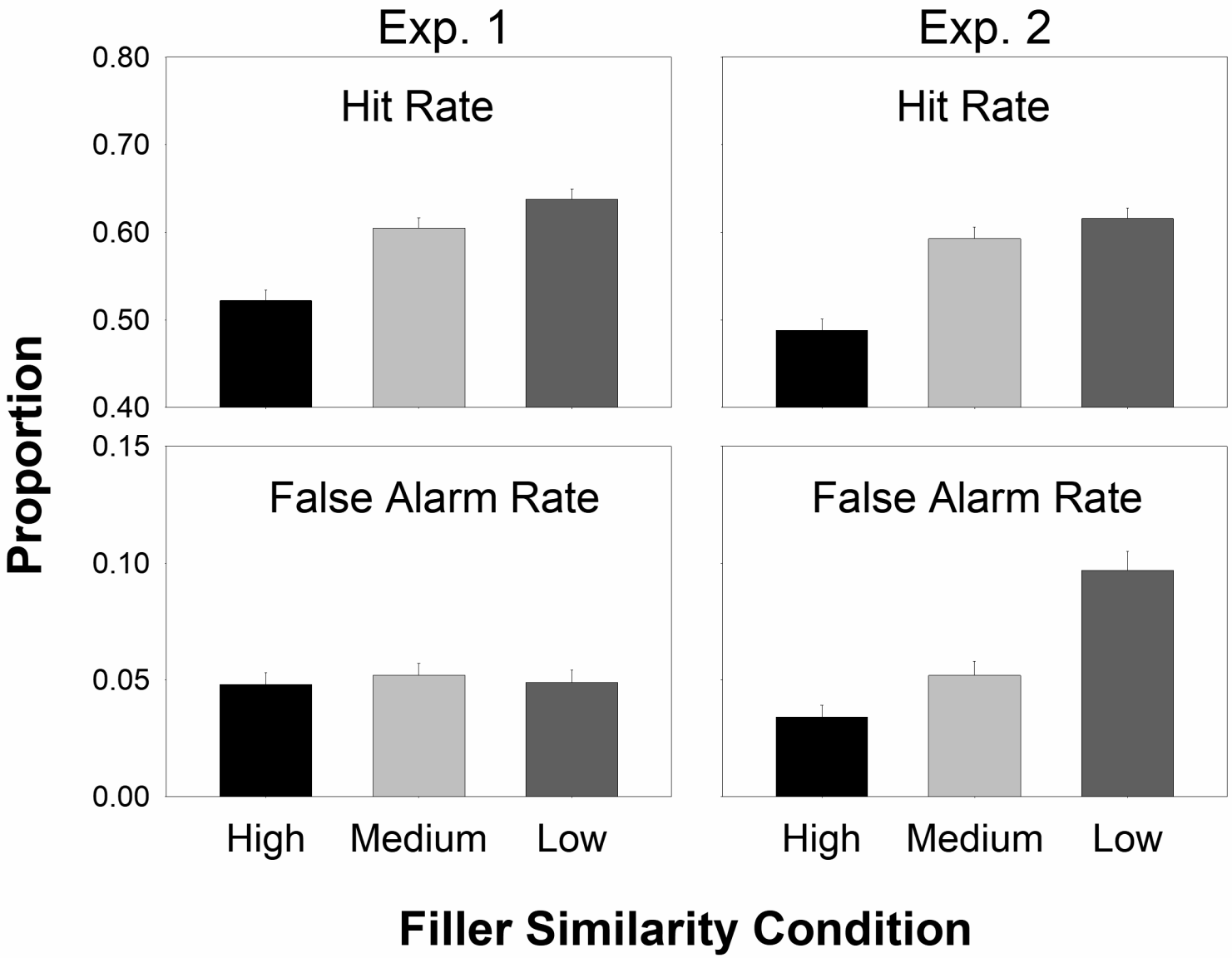
Figure 3. Exp. 1: The hit rate in the low-similarity condition was significantly higher than the hit rate in both the medium-similarity condition ($z = 2.02$, $p = .043$) and high-similarity condition ($z = 6.99$, $p < .001$). The hit rate in the medium-similarity condition was also higher than that of the high-similarity condition ($z = 4.97$, $p < .001$). The corresponding comparisons for the false alarm rates did not approach significance ($z = 0.35$, $p = .726$, $z = 0.16$, $p = .874$, and $z = 0.51$, $p = .610$). Exp. 2: The hit rate in the low-similarity condition was non-

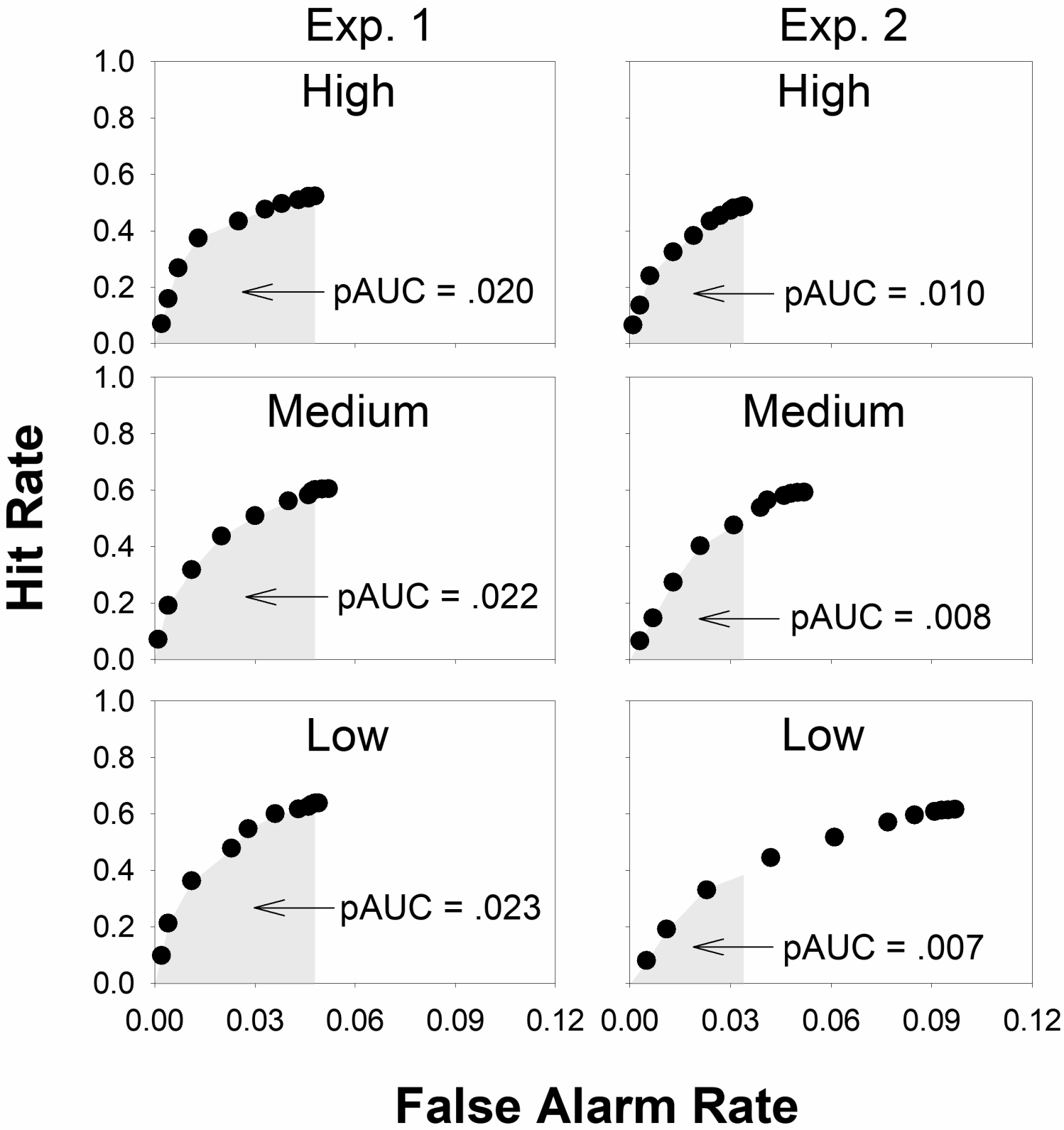
significantly higher than the hit rate in the medium-similarity condition ($z = 0.74$, $p = .461$) and significantly higher than the hit rate in the high-similarity condition ($z = 5.80$, $p < .001$). The hit rate in the medium-similarity condition was also significantly higher than that of the high-similarity condition ($z = 5.02$, $p < .001$). The false alarm rate in the low-similarity condition was significantly higher than the false alarm rate in both the medium-similarity ($z = 3.48$, $p = < .001$) and high-similarity ($z = 5.39$, $p = < .001$) conditions, and the false alarm rate in the medium-similarity condition was non-significantly higher than the false alarm rate in the high-similarity ($z = 1.78$, $p = .075$).

Figure 4. Discriminability is measured using partial area under the curve (pAUC), using a common false alarm rate across the three conditions (21). Exp. 1: The low-similarity pAUC was significantly larger than the high-similarity pAUC ($p = .023$, one-tailed, per our pre-registration). Exp. 2: The low-similarity pAUC was significantly smaller than the high-similarity pAUC ($p = .01$, one-tailed, per our pre-registration).


d'_{TP}
TP Lineups:
σ
σ
Fillers
Suspect
$\mu_{F:TP}$
μ_G
$d'_{TA} = 0$
TA Lineups:
σ
Suspect/Fillers
$\mu_I = \mu_{F:TA}$
Memory Match Signal


Exp. 1
Exp. 2
L M H
L M H
d'_{TP}
d'_{TP}
TP Lineups
μ_G
μ_G
$d'_{TA} = 0$
$d'_{TA} > 0$
$d'_{TA} < 0$
TA Lineups
Fillers
Suspect
$\mu_I = \mu_{F:TA}$
μ_I
Memory Match Signal
Memory Match Signal


Exp. 1
Exp. 2
Hit Rate
Hit Rate
False Alarm Rate
False Alarm Rate
0.80
0.70
0.60
0.50
0.40
0.15
0.10
0.05
0.00
Proportion
High
Medium
Low
High
Medium
Low
Filler Similarity Condition


Exp. 1
Exp. 2
High
High
Medium
Medium
Low
Low
pAUC = .020
pAUC = .010
pAUC = .022
pAUC = .008
pAUC = .023
pAUC = .007
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.03
0.06
0.09
0.12
Hit Rate
False Alarm Rate

Supporting Information for:

Optimizing the Selection of Fillers in Police Lineups

Melissa F. Colloff, Brent Wilson, Travis M. Seale-Carlisle, & John T. Wixted

Corresponding author: John T. Wixted

Email: jwixted@ucsd.edu

This file includes:

Supplementary text

Figures S1 to S8

Tables S1 to S4

SI References

Supporting Information

Manipulating filler similarity

Filler similarity relative to the innocent suspect in TA Lineups. Using our feature-matching model, Figure S1 illustrates the predicted effect of choosing fillers who are similar or dissimilar to the innocent suspect, using a single diagnostic feature (f_6 = blue eyes). Because this feature was not included in the witness's description of the perpetrator, the eyes of the innocent suspect will match this diagnostic feature by chance (left side of the tree in Figure S1) or will mismatch it by chance (right side of the tree in Figure S1). Given that each feature has 5 potential settings ($m = 5$), the probability that the innocent suspect's face will match the diagnostic feature by chance is, as noted earlier, $p = 1/m = .20$. If it does match (i.e., if the innocent suspect has blue eyes), then the probability that a similar filler selected to match the blue eyes of the innocent suspect will also have blue eyes is, of course, 1.0. Conversely, the probability that a dissimilar filler selected *not* to match the blue eyes of the innocent suspect will have blue eyes is 0. In that case, the dissimilar filler's eyes will be one of the remaining $m - 1$ non-blue colors.

Next, consider the right side of the tree in Figure S1. The probability that a diagnostic feature such as blue eyes will *not* match a feature of the innocent suspect's face by chance is $1 - p = .80$. If it does not match (i.e., if the suspect has brown eyes), then the probability that a similar filler selected to match the brown eyes of the innocent suspect will have blue eyes is, of course, 0 (i.e., the similar filer will have brown eyes, too). By contrast, the probability that a dissimilar filler selected *not* to match the brown eyes of the innocent suspect will have some chance of having blue eyes (thereby matching a diagnostic feature in memory). Excluding brown eyes, there are $m - 1$ eye colours left to choose from. Thus, the probability that the dissimilar filler will have blue eyes given that the innocent suspect has non-blue eyes is $1 / (m - 1) = p/(1 - p) = 1/4 = .25$.

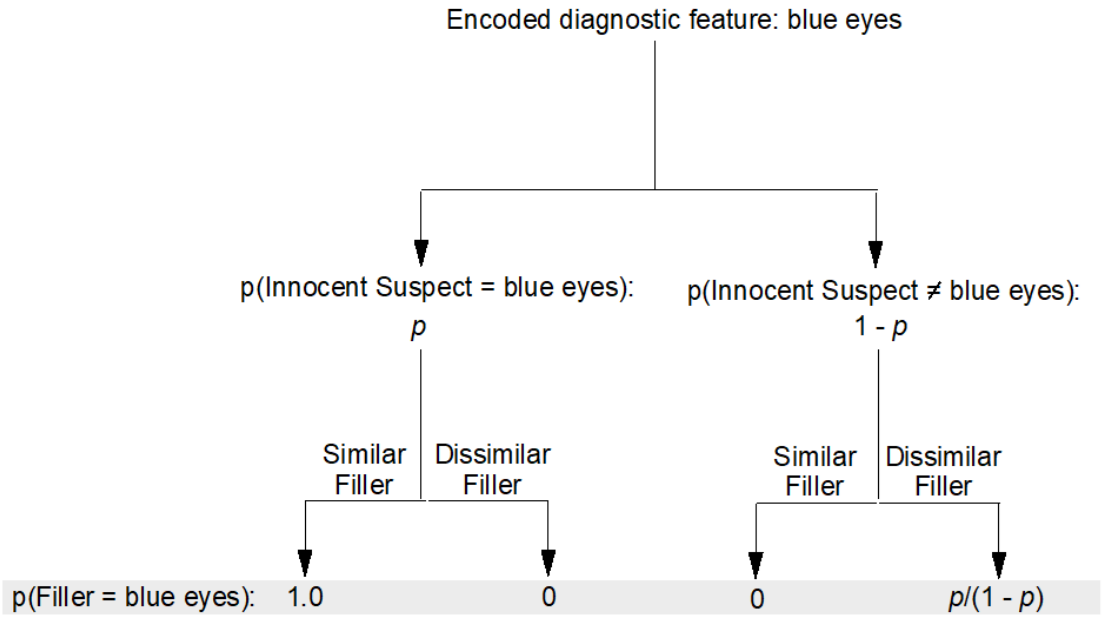


Figure S1. Conditional probabilities that a filler will match an encoded diagnostic feature (blue eyes), when fillers are chosen to be similar or dissimilar to an innocent suspect who either does or does not have blue eyes.

With the conditional probabilities in Figure S1 specified, we can now directly compute the probability that a TA filler will have blue eyes (matching a diagnostic feature in the memory of the eyewitness) depending on whether the filler was selected to be similar or dissimilar to the innocent suspect. The probability that a similar filler selected to have the same eye colour as the innocent suspect will match the blue eyes of the perpetrator stored in the witness's memory (the two "similar filler" paths in Fig. S1) is equal to the probability that the innocent suspect has blue eyes by chance (p) times 1.0 plus the probability that the innocent suspect has other-than-blue eyes by chance ($1 - p$) times 0. The resulting probability comes to p:

$$P(\text{Filler} = \text{blue eyes}|\text{Similar}) = (p)(1.0) + (1 - p)(0) = p$$

Similarly, the probability that a dissimilar filler selected to mismatch the eye colour of the innocent suspect will match the blue eyes of the perpetrator stored in the witness's memory

(the two "dissimilar filler" paths in Fig. S1) is equal to the probability that the innocent suspect has blue eyes by chance (p) times 0 plus the probability that the innocent suspect has other-than-blue eyes by chance ($1 - p$) times the probability that, of the remaining $m - 1$ feature settings for eye colour (excluding the non-blue eye colour of the suspect), the filler ends up with blue eyes. That probability is, of course, $1 / (m - 1)$. As noted above, is $1 / (m - 1) = p/(1 - p)$. Thus, the probability that the innocent suspect will not have blue eyes and a dissimilar filler will have blue eyes is $(1 - p)(p / [1 - p])$. Overall, the probability of a filler selected to be dissimilar to the innocent suspect comes to:

$$P(\text{Filler = blue eyes}|\text{Disimilar}) = (p)(0) + (1 - p)(\text{p}/[1 - p]) = p$$

This is the same probability we obtained when fillers are selected to be similar to the innocent suspect. Thus, according to this simple feature-matching model, everyone in a TA lineup—innocent suspect, similar fillers and dissimilar fillers alike—all have the same chance of matching the perpetrator's blue eyes (namely, p). Because, none of the fillers chosen to be similar or dissimilar to the innocent suspect will look more like the perpetrator (as encoded in the memory of the eyewitness) than the innocent suspect does, $\mu_{F\text{-}TA}$ remains the same across manipulations of filler similarity, so the false alarm rate should remain unchanged.

Filler similarity relative to the perpetrator in TA Lineups. Now consider choosing fillers for TA lineups who are similar or dissimilar to the *perpetrator*, again using a single diagnostic feature (f_6 = blue eyes). In this case, the terms "similar filler" and "dissimilar filler" in lower part of Figure S2 refer to the filler's similarity to the perpetrator (not to the innocent suspect). On the left, as before, we assume the innocent suspect happens to have blue eyes, matching the corresponding feature of the perpetrator in memory. Thus, for this feature, whether we are choosing a filler in order to match a feature to the innocent suspect or

to the perpetrator, everything remains the same. That is, the probability that a similar filler selected to match the blue eyes of the perpetrator will also have blue eyes is 1.0, and the probability that a dissimilar filler selected to mismatch the blue eyes of the perpetrator will have blue eyes is 0.

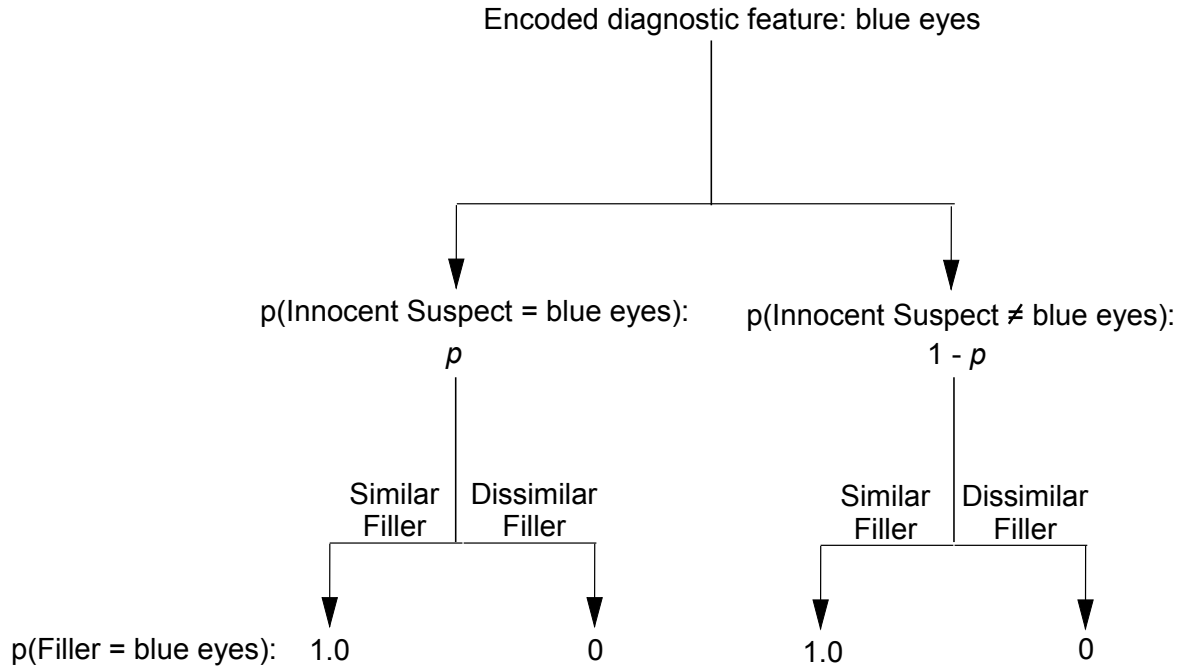


Figure S2. Conditional probabilities that a filler will match an encoded diagnostic feature (blue eyes), when fillers are chosen to be similar or dissimilar to the perpetrator depending on whether the innocent suspect either does or does not have blue eyes. Now, the eyes of the innocent suspect are irrelevant because similar and dissimilar fillers are chosen with respect to the perpetrator, without regard for the innocent suspect.

On the right, the innocent suspect happens to have non-blue eyes, mismatching the corresponding feature of the perpetrator in memory. Regardless, the probability that a similar filler selected to match the blue eyes of the perpetrator will also have blue eyes is still 1.0, not 0. Similarly, the probability that a dissimilar filler selected to mismatch the blue eyes of the perpetrator will have blue eyes is still 0, not $p / (1 - p)$. In other words, because the innocent suspect was not taken into consideration when selecting fillers, whether or not the innocent suspect has blue eyes is irrelevant. To complete this argument using equations parallel to those used above, the probability that a similar filler selected to have the same eye colour as the perpetrator will match the blue eyes of the perpetrator stored in the witness's memory is

equal to the probability that the innocent suspect has blue eyes by chance (p) times 1.0 plus the probability that the innocent suspect has other-than-blue eyes by chance ($1 - p$) times 0. The resulting probability comes to 1.0:

$$P(\text{Filler = blue eyes}|\text{Similar}) = (p)(1.0) + (1 - p)(1.0) = 1.0$$

Similarly, the probability that a dissimilar filler selected to mismatch the eye colour of the perpetrator will match the blue eyes of the perpetrator stored in the witness's memory is equal to the probability that the innocent suspect has blue eyes by chance (p) times 0 plus the probability that the innocent suspect has other-than-blue eyes by chance ($1 - p$) times 0:

$$P(\text{Filler = blue eyes}|\text{Dissimilar}) = (p)(0) + (1 - p)(0) = 0$$

Keep in mind that the probability that the innocent suspect's eye colour coincidentally matches memory of the perpetrator, namely p, falls between these two values. Thus, similar fillers are more likely to match the memory of the perpetrator than the innocent suspect is, so the false alarm rate should now decrease rather than staying the same. In other words, the innocent suspect will be protected by what has sometimes been referred to as "filler siphoning" (Smith, Wells, Smalarz, & Lampinen, 2018). Conversely, dissimilar fillers (selected because they are dissimilar to the perpetrator) are now *less* likely to match the memory of the perpetrator than the innocent suspect is. A lineup biased in this manner would result in a higher false alarm rate because filler siphoning would happen to a lesser degree.

The point is that, in contrast to selecting fillers based on similarity to the guilty suspect in TP lineups and based on similarity to the innocent suspect in TA lineups, when fillers in both TP lineups and TA lineups are selected based on similarity to the perpetrator, the hit rate

and the false alarm rate should increase as filler similarity decreases (stretching the ROC to the right). In more formal terms, instead of d'_{TP}, selectively increasing as filler similarity decreases (with d'_{TA} remaining fixed at 0), both d'_{TP} and d'_{TA} will increase as filler similarity decreases. This is another way of saying that both the guilty suspect and the innocent suspect will increasingly stand out in the lineup as filler similarity to the perpetrator decreases.

Shifting the ROC to the right (i.e., increasing both the hit rate and the false alarm rate) as filler similarity decreases does not mean that the ability to discriminate innocent from guilty suspects will necessarily change. As noted by Colloff et al. (2018), the filler-siphoning effect is neutral with respect to that issue. However, diagnostic feature-detection theory specifically predicts that not only will the ROC shift to the right as filler similarity to the perpetrator decreases, the ability to discriminate innocent from guilty suspects will also decrease (Colloff et al., 2016). Thus, diagnostic feature-detection theory predicts that the ordering of the ROCs in the high-, medium-, and low-similarity conditions should be the opposite of what is observed when TA fillers are matched to the innocent suspect. Instead of the low-similarity condition yielding the highest level of discriminability, it should now yield the lowest level of discriminability. The basis for this prediction is illustrated in Figure S3.

A **Similar Fillers**

	In Description					Not in Description															
f	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ	σ^2
TP Guilty	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	20
TP Filler	1	1	1	1	1	0	1	1	0	0	1	0	1	0	1	1	0	1	1	0	13	20
TA Innocent	1	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	8	20
TA Filler	1	1	1	1	1	0	1	1	0	0	1	0	1	0	1	1	0	1	1	0	13	20

B **Dissimilar Fillers**

	In Description					Not in Description															
f	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ	σ^2
TP Guilty	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	20
TP Filler	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	20
TA Innocent	1	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	8	20
TA Filler	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	20

Figure S3. Example illustrating the settings of features when fillers in both TP and TA lineups are selected based on their similarity to the perpetrator.

Figure S3A illustrates a high-similarity condition and Figure S3B illustrates a low-similarity condition. The entries refer to whether or not a feature is present (1 = present, 0 = absent). The 5 features included in the witness's description are always present because we assume that these are description-matched lineups. For the remaining potentially diagnostic features, the innocent suspect will coincidentally match of few of the perpetrator's feature settings in memory (features 9, 11, and 17 in this example).

In the high-similarity condition, TP and TA fillers alike are chosen to match additional features of the perpetrator. Imagine that 8 of the remaining 15 features are chosen to match the perpetrator. Thus, the relevant means come to 20 and 8 for the guilty and innocent suspects (as was true of our earlier examples), but now come to 13 for the fillers in both TP and TA lineups. Because it is still the case that 20 features were summed in all cases, the standard deviation of memory signal is still $\sigma = \sqrt{20}$. Thus,

$$d'_{TP} = \frac{\mu_G - \mu_{F:TP}}{\sigma} = \frac{20 - 13}{\sqrt{20}} = 1.57$$

In addition,

$$d'_{TA} = \frac{\mu_I - \mu_{F:TA}}{\sigma} = \frac{8 - 13}{\sqrt{20}} = -1.12$$

In other words, in the high-similarity condition, the innocent suspect now generates a memory signal that is smaller than that of the fillers.

In the low-similarity condition (Figure S3B), fillers chosen because they are dissimilar to the perpetrator will match on none of the remaining features. The mean memory-match signal comes to 20 and 8 for the guilty and innocent suspects (as before), but come to only 5 for the fillers in both TP and TA lineups. Because 20 features were summed in all cases, the standard deviation of memory signal is $\sigma = \sqrt{20}$. Thus,

$$d'_{TP} = \frac{\mu_G - \mu_{F:TP}}{\sigma} = \frac{20 - 5}{\sqrt{20}} = 3.35$$

In addition,

$$d'_{TA} = \frac{\mu_I - \mu_{F:TA}}{\sigma} = \frac{8-5}{\sqrt{20}} = 0.67$$

In other words, in the low-similarity condition, the innocent suspect now generates a memory signal that is greater than that of the fillers.

These calculations merely formalize the point that the innocent suspect stands out when low-similarity fillers are used and is effectively concealed when high-similarity fillers are used. Interestingly, however, the ability to discriminate innocent from guilty suspects across TP and TA lineups (d'_{IG}), remains unchanged. Note that $d'_{IG} = d'_{TP} - d'_{TA}$. For high-similarity lineups, $d'_{IG} = 1.57 - (-1.12) = 2.68$. For low-similarity lineups, $d'_{IG} = 3.35 - 0.67 = 2.68$. Thus, the predicted filler siphoning that will occur in the high-similarity condition would not be expected to affect the ability to discriminate innocent from guilty suspects.

Now consider what should happen if witnesses discount non-diagnostic features, as illustrated in Figure S4A (high-similarity condition) and Figure S4B (low-similarity condition). If non-diagnostic features are discounted, any features that happen to match in the TP lineup (namely, all of the description-matched features in both similarity conditions and some of the remaining features in the high-similarity condition) are no longer taken into consideration because they are non-diagnostic of guilt. Basically, all features in Figure S3 where both are set to 1 for the filler and suspect in TP lineups and where both are set to 1 for the filler and suspect in TA lineups are removed from consideration, as if they do not exist. As illustrated next, discounting non-diagnostic features enhances discriminability, amplifying both d'_{TP} and d'_{TA} (compared to when they are not discounted). More interestingly, now, d'_{IG} should be affected by filler similarity as well.

A **Similar Fillers**

	In Description					Not in Description															
f	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ	σ^2
TP Guilty						1			1	1		1		1			1			1	7	7
TP Filler						0			0	0		0		0			0			0	0	7
TA Innocent						0	0	0	1	0		0	0	0	0	0	1	0	0	0	2	14
TA Filler						0	1	1	0	0		0	1	0	1	1	0	1	1	0	7	14

B **Dissimilar Fillers**

	In Description					Not in Description															
f	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ	σ^2
TP Guilty						1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	15
TP Filler						0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15
TA Innocent						0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	3	15
TA Filler						0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15

Figure S4. Example illustrating the settings of features when fillers in both TP and TA lineups are selected based on their similarity to the perpetrator and when features are disregarded when they have the same setting.

After discounting non-diagnostic features, in the high-similarity condition (Figure S4A), the means come to 7 and 0 for the guilty suspects and fillers in TP lineups, respectively, and to 2 and 7 for innocent suspects and fillers in TA lineups, respectively. Because 7 features are summed in TP lineups, the standard deviation of the memory signal for faces in a TP lineup is $\sigma = \sqrt{7}$. Thus,

$$d'_{TP} = \frac{\mu_G - \mu_{F:TP}}{\sigma} = \frac{7 - 0}{\sqrt{7}} = 2.65$$

In addition, because 14 features are summed in TA lineups, the standard deviation of the memory signal for faces in a TP lineup is $\sigma = \sqrt{14}$. Thus,

$$d'_{TA} = \frac{\mu_I - \mu_{F:TA}}{\sigma} = \frac{2 - 7}{\sqrt{14}} = -1.34$$

In the low-similarity condition (Figure S4B), the means come to 15 and 0 for the guilty suspects and fillers in TP lineups, respectively, and to 3 and 0 for innocent suspects and

fillers in TA lineups, respectively. Because 15 features are summed in TP lineups, the standard deviation of the memory signal for faces in a TP lineup is $\sigma = \sqrt{15}$. Thus,

$$d'_{TP} = \frac{\mu_G - \mu_{F:TP}}{\sigma} = \frac{15 - 0}{\sqrt{15}} = 3.87$$

In addition,

$$d'_{TA} = \frac{\mu_I - \mu_{F:TA}}{\sigma} = \frac{3 - 0}{\sqrt{15}} = 0.77$$

Now, the ability to discriminate innocent from guilty suspects, which is captured by d'_{TP} - d'_{TA}, is lower in the low-similarity condition. In the low-similarity condition, d'_{TP} - d'_{TA} = 3.87 – 0.77 = 3.10, and in the high-similarity condition, d'_{TP} - d'_{TA} = 2.65 – (-1.34) = 3.98. Thus, when features are discounted, the high-similarity condition is now expected to increase the ability to discriminate innocent from guilty suspects (similar to the effect reported by Colloff et al., 2016). Note that this is exactly the opposite of the filler-similarity prediction that is made when description-match TA fillers are selected on the basis of their similarity to the innocent suspect rather than to the perpetrator.

SI Results

Identification Responses

Experiment 1. Table S1 presents the proportions (and frequencies) of response outcomes (suspect ID, filler ID, or No ID) for TP and TA lineups across the three levels of filler similarity for Experiment 1 and the two replications. It is clear that, as predicted, the hit rate increased as filler similarity decreased, whereas the false alarm rate exhibited no systematic trends. As expected, in TP lineups, the filler ID rate increased as the hit rate decreased with increasing similarity (i.e., fillers who were more similar to the guilty suspect were more attractive than dissimilar fillers). Unexpectedly, in TA lineups, the filler ID rate was consistently higher in the high-similarity condition relative to the other two similarity conditions. Thus, for reasons unknown, the high-similarity fillers were slightly more

attractive than the fillers in the other conditions. Because of that effect, the TA lineup rejection rate was lower in the high-similarity condition. However, this trend had no apparent effect on the false alarm rate, which remained stable across the three filler-similarity conditions.

Table S1
Proportion (and Frequencies) of Suspect, Filler, and Reject (No ID) Identification Responses in Low, Medium, and High Similarity Target-Present and Target-Absent Lineups in Experiment 1

Experiment and Similarity Condition	Target-present			Target-absent		
	Suspect	Filler	No ID	Suspect	Filler	No ID
Experiment 1						
Low	**0.63** (404)	0.09 (56)	0.28 (182)	**0.05** (33)	0.32 (204)	0.62 (394)
Medium	**0.58** (361)	0.14 (86)	0.28 (175)	**0.05** (36)	0.29 (189)	0.66 (434)
High	**0.51** (303)	0.19 (110)	0.30 (177)	**0.04** (24)	0.39 (249)	0.57 (361)
Replication 1						
Low	**0.64** (364)	0.09 (50)	0.27 (157)	**0.05** (27)	0.29 (153)	0.65 (339)
Medium	**0.61** (345)	0.10 (56)	0.29 (166)	**0.04** (23)	0.27 (153)	0.69 (388)
High	**0.53** (297)	0.18 (101)	0.29 (162)	**0.05** (29)	0.37 (206)	0.58 (328)
Replication 2						
Low	**0.65** (377)	0.09 (53)	0.26 (151)	**0.04** (25)	0.33 (191)	0.63 (363)
Medium	**0.63** (360)	0.14 (79)	0.23 (133)	**0.06** (35)	0.31 (187)	0.63 (372)
High	**0.52** (291)	0.24 (135)	0.24 (132)	**0.06** (31)	0.41 (227)	0.53 (295)
Combined data						
Low	**0.64** (1145)	0.09 (159)	0.27 (490)	**0.05** (85)	0.32 (548)	0.63 (1096)
Medium	**0.61** (1066)	0.13 (221)	0.27 (474)	**0.05** (94)	0.29 (529)	0.66 (1194)
High	**0.52** (891)	0.20 (346)	0.28 (471)	**0.05** (84)	0.39 (682)	0.56 (984)

Experiment 2. Table S2 presents the proportions (and frequencies) of response outcomes (suspect ID, filler ID, or No ID) for TP and TA lineups across the three levels of filler similarity for Experiment 2 and the two replications. It is clear that, as predicted, the hit rate and the false alarm rate increased as filler similarity decreased. As expected, in both TP and TA lineups, the filler ID rate increased as the hit rate decreased with increasing similarity (i.e., fillers who were more similar to the guilty suspect were more attractive than dissimilar fillers). That is, with increasingly similar fillers, the guilty and innocent suspect are protected

by what has sometimes been referred to as "filler siphoning" (Smith, Wells, Smalarz, & Lampinen, 2018).

Table S2
Proportion (and Frequencies) of Suspect, Filler, and Reject (No ID) Identification Responses in Low, Medium, and High Similarity Target-Present and Target-Absent Lineups in Experiment 2

Experiment and Similarity Condition	Target-present			Target-absent		
	Suspect	Filler	No ID	Suspect	Filler	No ID
Experiment 2						
Low	**0.62** (349)	0.09 (49)	0.29 (162)	**0.09** (53)	0.23 (127)	0.68 (378)
Medium	**0.61** (328)	0.11 (59)	0.28 (150)	**0.04** (24)	0.30 (170)	0.66 (371)
High	**0.48** (259)	0.22 (118)	0.30 (158)	**0.04** (22)	0.38 (217)	0.59 (337)
Replication 1						
Low	**0.59** (258)	0.06 (25)	0.35 (151)	**0.09** (33)	0.20 (77)	0.72 (278)
Medium	**0.57** (239)	0.13 (54)	0.30 (125)	**0.06** (26)	0.25 (109)	0.69 (300)
High	**0.48** (191)	0.22 (88)	0.30 (117)	**0.03** (12)	0.41 (173)	0.56 (240)
Replication 2						
Low	**0.63** (351)	0.12 (66)	0.26 (144)	**0.11** (58)	0.25 (136)	0.64 (349)
Medium	**0.59** (343)	0.15 (86)	0.26 (150)	**0.05** (31)	0.38 (216)	0.56 (320)
High	**0.50** (264)	0.23 (121)	0.28 (148)	**0.03** (19)	0.51 (285)	0.46 (259)
Combined data						
Low	**0.62** (958)	0.09 (140)	0.29 (457)	**0.10** (144)	0.23 (340)	0.67 (1005)
Medium	**0.59** (910)	0.13 (199)	0.28 (425)	**0.05** (81)	0.32 (495)	0.63 (991)
High	**0.49** (714)	0.22 (327)	0.29 (423)	**0.03** (53)	0.43 (675)	0.53 (836)

Figure S5 summarizes the findings of primary interest for both Experiments 1 and 2, namely the hit and false alarm rates across the three similarity conditions. Considering Experiment 1 (left column), it is clearly apparent that, for all three runs of the experiment, the hit rate increases in orderly fashion as filler similarity decreases. The increase in each case is > .10 in every case, so the effect is nontrivial in terms of potential real-world impact. The corresponding false alarm rates from the three experiments are similar across filler similarity conditions, but no apparent trends are evident. However, because false alarms were relatively rare, the data are noisy, making it hard to rule out the possibility that a trend exists. Conversely, considering Experiment 2 (right column), it is clearly apparent that, for all three

runs of the experiment, both the hit rate and false alarm increase in orderly fashion as filler similarity decreases.

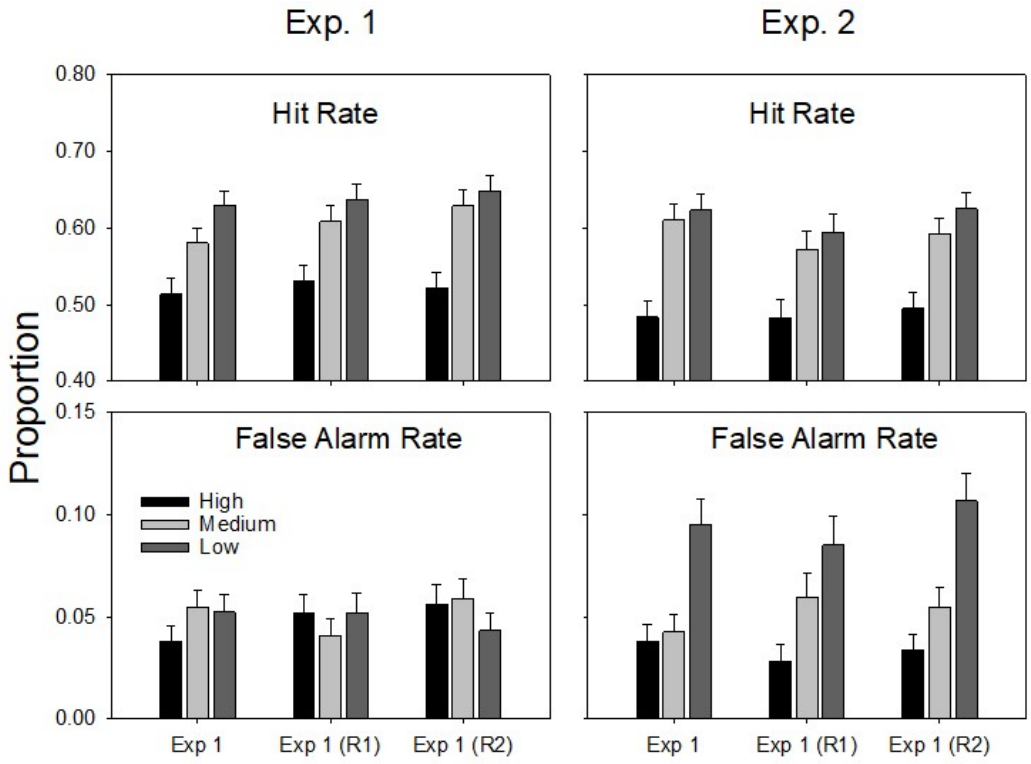


Figure S5. In Experiment 1, the hit rate in the low-similarity condition was consistently higher than the hit rate in both the medium-similarity condition and high-similarity condition. By contrast, the false alarm rate did not vary systematically across filler similarity conditions.

Empirical discriminability

We constructed empirical identification partial ROC curves. To construct these ROC curves, we used the 11-point confidence scale, ranging from 100% to 0%, and plotted the cumulative correct ID rate (number of guilty suspect IDs ÷ total number of target-present lineups) against the cumulative false ID rate (number of innocent suspect IDs ÷ total number of target-absent lineups) over decreasing levels of confidence. The leftmost point on the ROC represents the correct and incorrect suspect IDs made with the highest level of confidence (100% sure), the next point represents the correct and incorrect suspect IDs made with the second-highest level of confidence (100% or 90% sure), and, continuing along the curve, the rightmost point represents all suspect IDs made with any level of confidence (100% - 0%).

To statistically compare the *p*ROC curves we used the *p*ROC statistical package to calculate the partial Area Under the Curve (*p*AUC) and *D,* a measure of effect size (D = (AUC1 – AUC2)/s, where *s* is the standard deviation of the difference between the two AUCs and is estimated using bootstrapping (Robin et al., 2011). In all *p*AUC analyses, we defined the specificity as 1 – FAR using the smallest false alarm rate (FAR) range in each experiment.

Experiment 1. ROCs plot two dependent measures against each other (hit rate vs. false alarm, rate), both of which are independently associated with measurement error. Moreover, for each experiment considered individually, the false alarm rate data are particularly noisy (shown in Figure S5). As such, the difference between the filler similarity conditions were not statistically significant according to the *p*AUC analysis. It is clear, however, from the *p*AUC statistics (Table S3) that the predicted pattern of results for Experiment 1 (low similarity > medium similarity > high similarity) was observed in 2 out of the 3 experiments (Replication 1 and 2). In the remaining experiment (Experiment 1), although the low-similarity condition once again yielded the best discriminability as predicted, the other two conditions were not ordered as predicted. The probability of obtaining ordered results as good or better than this is $p = .047$.

Table S3
Identification ROC Analysis Partial Area Under the Curve (pAUC) Statistics [and 95% Confidence Intervals] for Experiment 1, Replication 1 and 2, and Combined Data

Similarity Condition	Experiment 1	Replication 1	Replication 2	Combined data
Low	0.016 [0.014, 0.019]	0.019 [0.015, 0.022]	0.017 [0.014, 0.021]	0.023 [0.021, 0.026]
Medium	0.015 [0.013, 0.018]	0.017 [0.012, 0.021]	0.016 [0.012, 0.019]	0.022 [0.019, 0.024]
High	0.016 [0.011, 0.020]	0.015 [0.011, 0.018]	0.013 [0.010, 0.017]	0.020 [0.018, 0.022]

Note. We used the FAR range of the least extensive curve in each analysis to set specificity (1 – FAR) to .96 for Experiment 1, Replications 1 and 2, and to .95 for the combined data analysis.

Experiment 2. It is clear from the *p*AUC statistics (Table S4) that the predicted pattern of results for Experiment 2 (high similarity > medium similarity > low similarity) was observed in 2 out of the 3 experiments (Experiment 2, Replication 1). In the remaining experiment (Replication 2), although the low-similarity condition once again yielded the poorest discriminability as predicted, the other two conditions were not ordered as predicted. The probability of obtaining ordered results as good or better than this is $p = .047$.

Table S4
Identification ROC Analysis Partial Area Under the Curve (pAUC) Statistics [and 95% Confidence Intervals] for Experiment 2, Replication 1 and 2, and Combined Data

Similarity Condition	Experiment 2	Replication 1	Replication 2	Combined data
Low	0.012 [0.010, 0.015]	0.005 [0.001, 0.009]	0.007 [0.005, 0.009]	0.007 [0.005, 0.008]
Medium	0.014 [0.010, 0.019]	0.006 [0.003, 0.010]	0.010 [0.007, 0.013]	0.008 [0.006, 0.011]
High	0.015 [0.012, 0.018]	0.011 [0.009, 0.014]	0.008 [0.005, 0.011]	0.010 [0.008, 0.011]

Note. We used the FAR range of the least extensive curve in each analysis to set specificity (1 – FAR) to .96 for Experiment 2, and to .97 for Replication 1 and 2 and the combined data analysis.

Vary similarity between the innocent suspect and the perpetrator

The experiments we conducted used the median-similarity filler from a large pool of description-matched fillers as the designated innocent suspect. However, in the real world, innocent suspects will sometimes be more similar to the perpetrator and sometimes less similar to the perpetrator. Here, we illustrate what our feature-matching model predicts across the full range of similarity between the innocent suspect and perpetrator.

For these simulations, the mean of the guilty suspect distribution was always set to 2, and the mean of the innocent suspect distribution varied from -2 (innocent suspect and perpetrator are extremely dissimilar) to 2 (innocent suspect and perpetrator are identical) in 9 steps. For the middle step (step 5), the mean of the innocent suspect distribution was equal to 0, and this simulated condition corresponds to the use of the median-similarity filler as the innocent suspect. For target-present lineups, the means of the low-, medium-, and high-similarity filler distributions were always set to 1, 0.5, and 0, respectively.

For the target-absent lineups, the means of the low-, medium-, and high-similarity filler distributions differed depending on the mean of the innocent suspect distribution, μ_I. Specifically, the means of the filler distributions were set to the values used for target-present lineups multiplied by (μ_I/μ_G). Thus, for the extreme case in which the innocent suspect was maximally dissimilar to the perpetrator, μ_I = -2.0, μ_G = 2.0, so μ_I/μ_G = -1.0. Thus, in that case, the means of the low-, medium-, and high-similarity filler distributions were set to -1.0(1.0), -1.0(0.5), and -1.0(0), or -1.0. -0.5 and 0, respectively. For the opposite extreme in which the innocent suspect was identical to the perpetrator, μ_I = 2.0, μ_G = 2.0, so μ_I/μ_G = 1.0. Thus, in that case, the means of the low-, medium-, and high-similarity filler distributions were set to 1.0(1.0), 1.0(0.5), and 1.0(0), or 1.0. 0.5 and 0, respectively. Finally, for the median-similarity case, μ_I = 0, μ_G = 2.0, so μ_I/μ_G = 0. Thus, in that case, the means of the low-, medium-, and high-similarity filler distributions were set to 0(1.0), 0(0.5), and 0(0), which is to say that they were all set to 0.

The results of the simulation are presented in Fig. S6. Each graph shows a predicted ROC, with the rightmost point of each ROC representing the overall hit and false alarm rate. The three columns correspond to the three filler-similarity conditions (low to high). The nine rows correspond to varying degrees of similarity between the innocent suspect and the perpetrator (extremely low to extremely high).

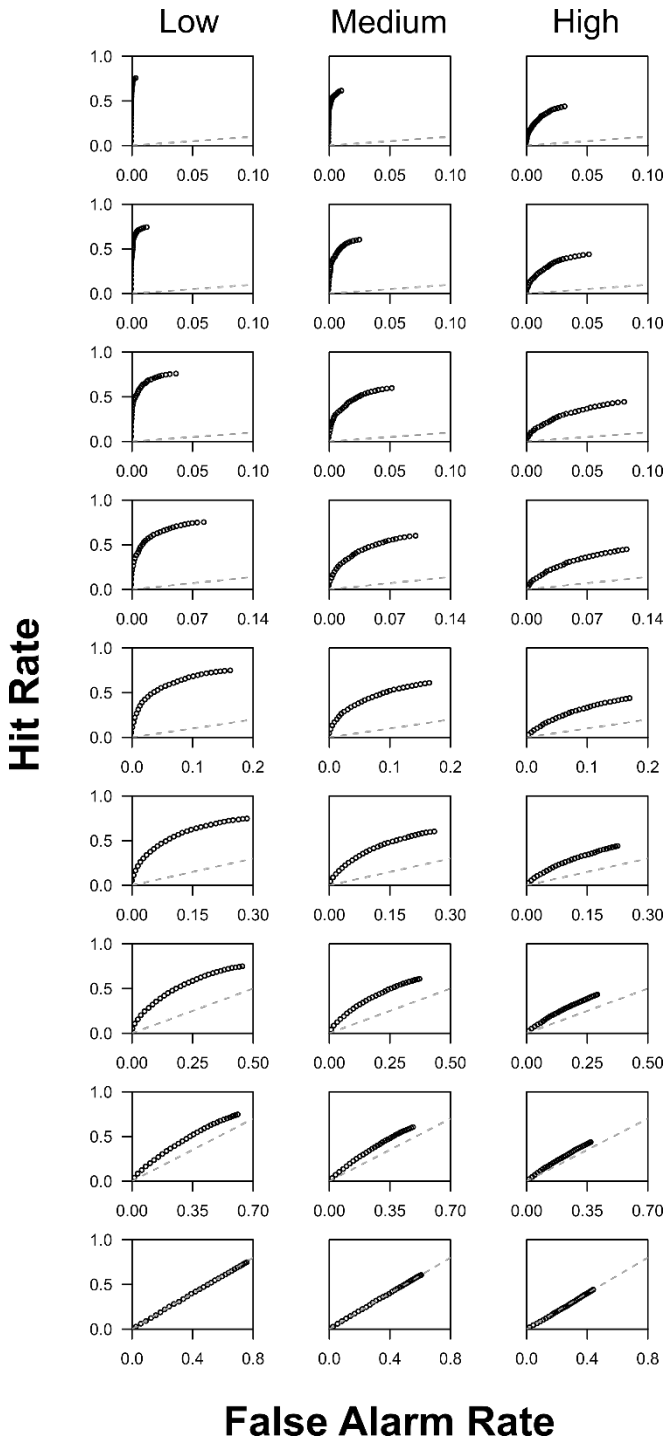


Figure S6. Model-based predictions of manipulating filler similarity from low to high (left to right) as the similarity between the innocent suspect and the perpetrator ranges from low to high (top to bottom). The middle row corresponds to the se of the median-similarity filler we used for Experiment 1. The rightmost point of each ROC represents the overall hit and false alarm rate for a given condition. When the innocent suspect and the perpetrator are maximally dissimilar (top row), the use of low-similarity fillers reduces the false alarm rate. By contrast, when the innocent suspect and the perpetrator are maximally similar such that they are identical twins (bottom row), the use of low-similarity fillers increases the false alarm rate.

As is evident in Fig. S6, the feature-matching model predicts that when the innocent suspect happens to be dissimilar to the perpetrator (rows 1 through 4, with row 1 corresponding to maximum dissimilarity), the use of low-similarity fillers should decrease the false alarm rate. By contrast, when the innocent suspect happens to be similar to the perpetrator (rows 6 through 9, with row 9 corresponding to maximum similarity), the use of low-similarity fillers should increase the false alarm rate. Overall, the risk to innocent suspects should remain unchanged, as it was here in Experiment 1 using the median-similarity filler (corresponding to row 5 in Fig. S6).

Fig. S6 also shows that, except in the extreme case where the innocent suspect and perpetrator are identical twins (discriminability = 0, row 9), the model further predicts that the use of low-similarity fillers should enhance discriminability across the board (i.e., regardless of how similar the innocent suspect is to the perpetrator).

SI Materials and Methods

Design

We used a 3 (suspect-filler similarity: low, medium, high) × 2 (target: present, absent) between-subjects design. We pre-registered our design and analyses before we collected data (Experiment 1: https://osf.io/s4fq6/?view_only=0cae62f2cc744acd880f91053723a75a; Experiment 2: https://osf.io/5sr9j/?view_only=e58b9c72abff45e4bd2fad79287a32a4).

Sample

Experiment 1. We recruited 3,877 participants from Amazon Mechanical Turk who completed the study for 50 cents. We excluded participants who incorrectly answered an attention check question about the number of people in the video ($N = 99$). The final sample was 3,778 participants (aged: 16 – 83, $M_{age} = 34.17$; gender: 52% female, 47% male, <1% other or prefer not to say; ethnicity: 65% White, 14% Asian, 7% Black, 6% Hispanic, 3% Mixed, 2% Native American, 3% other or prefer not to say).

Experiment 1, Replication 1. We recruited 3,395 new participants from Amazon Mechanical Turk who completed the study. We excluded participants who incorrectly answered an attention check question about the number of people in the video ($N = 51$). The final sample was 3,344 participants (aged: 16 – 76, $M_{age} = 32.93$; gender: 53% female, 47% male, <1% other or prefer not to say; ethnicity: 62% White, 13% Asian, 8% Black, 9% Hispanic, 3% Mixed, 1% Native American, 3% other or prefer not to say).

Experiment 1, Replication 2. We recruited 3,530 new participants from Amazon Mechanical Turk who completed the study. We excluded participants who incorrectly answered an attention check question about the number of people in the video ($N = 93$). The final sample was 3,437 participants (aged: 16 – 80, $M_{age} = 34.58$; gender: 52% female, 48% male, <1% other or prefer not to say; ethnicity: 58% White, 20% Asian, 7% Black, 7% Hispanic, 3% Mixed, 1% Native American, 5% other or prefer not to say).

Experiment 2. We recruited 3,425 participants from Amazon Mechanical Turk who completed the study. We excluded participants who incorrectly answered an attention check question about the number of people in the video ($N = 94$). The final sample was 3,331 participants (aged: 16 – 75, $M_{age} = 31.53$; gender: 48% female, 51% male, 1% other or prefer not to say; ethnicity: 59% White, 12% Asian, 7% Black, 12% Hispanic, 4% Mixed, 2% Native American, 4% other or prefer not to say).

Experiment 2, Replication 1. We recruited 1,822 new participants from Amazon Mechanical Turk who completed the study and 739 students from UCSD who completed the study for course credit (total $N = 2,561$). We excluded participants who incorrectly answered an attention check question about the number of people in the video ($N = 64$), and a participant had completed the study a second time ($N = 1$). The final sample was 2,496 participants (aged: 16 – 77, $M_{age} = 27.60$; gender: 51% female, 48% male, 1% other or prefer

not to say; ethnicity: 44% White, 24% Asian, 6% Black, 15% Hispanic, 5% Mixed, 1% Native American, 5% other or prefer not to say).

Experiment 2, Replication 2. We recruited 3,530 new participants from Amazon Mechanical Turk who completed the study. We excluded participants who incorrectly answered an attention check question about the number of people in the video (N = 174). The final sample was 3,346 participants (aged: 16 – 77, M_{age} = 33.25; gender: 46% female, 53% male, 1% other or prefer not to say; ethnicity: 57% White, 11% Asian, 11% Black, 9% Hispanic, 3% Mixed, 4% Native American, 5% other or prefer not to say).

Procedure

Participants first watched the mock-crime video. They were instructed to pay close attention because they would be asked questions about it later. After the video ended, subjects played Tetris as a filler task for 5 min. Next, participants were told that they would view a lineup of six people and the perpetrator may or may not be present. Participants saw a lineup composed of two rows of three photos; the photos displayed depended on to which of the six experimental condition the participant had been randomly assigned. In TP lineups, the perpetrator was presented alongside five fillers selected randomly from the pool of low-, medium-, or high- similarity target-present filler group. In TA lineups, the innocent suspect was presented alongside five fillers who were selected at random from either the low, medium-, or high-similarity target-absent filler group. The position of lineup members in the array was randomly determined for each participant. They were asked to make an identification by clicking on either the person they believed to be the perpetrator or on an option underneath the lineup labelled "Not Present." Next, we asked participants to use an 11-point Likert-type scale (0% = *guessing*, 100% = *completely certain*) to rate their confidence in their decision. Finally, subjects answered multiple-choice attention-check

questions (e.g., “How many people were in the video?”), and answered a number of demographic questions.

Materials

Stimuli Creation

We presented participants (N = 103) from Amazon Mechanical Turk with our mock-crime video depicting a male perpetrator stealing a laptop from an office. After a 4 min filler task, participants were asked to describe the appearance of the perpetrator in the video, as if they were describing that person to a police investigator. We removed data from 12 participants who did not describe the appearance of the perpetrator, and then formed a general description of the perpetrator that was consistent with the descriptions from the 91 remaining participants (e.g., white, male). Using the description, we selected potential fillers from a pool of 529 images that had previously been downloaded from online prison databases in the US (e.g., Florida Department of Corrections). From the pool, we removed 201 photos depicting individuals who did not fit the description, who had prominent distinctive features like scars, bruises, or tattoos, or were not facing the camera. This resulted in a final pool of 328 description-matched fillers for use in our experiments. We edited the filler images and the perpetrator’s image to remove visible clothing, and to ensure that the background colour and dimensions were consistent across images.

Next, we collected similarity ratings in two stages. In stage one, participants (N = 315) from Amazon Mechanical Turk were presented with an image of the perpetrator alongside a filler image, and were asked to rate how physically similar the two individuals were on a Likert-type scale from 1 (*not at all physically similar*) to 7 (*very physically similar*). Each participant rated the similarity of the perpetrator to 50 fillers randomly selected from the pool. On average, each filler received 43 ratings. Across all participants and ratings, the mean similarity rating was 2.94. We selected the filler face with the mean rating (2.94), to be the

designated innocent suspect and removed him from the pool. In stage two, a new group of participants (N = 352) rated the similarity of the same fillers to the innocent suspect, using the same pairwise procedure.

Filler similarity manipulation

Experiment 1. For each photo in our pool of 327 potential fillers, we obtained an average similarity rating to the perpetrator (stage 1) and an average similarity rating to the innocent suspect (stage 2). Ideally, these two ratings would be completely unrelated to each other across the 327 faces (correlation = 0). If the correlation were 0, then a filler chosen because the face was rated as being dissimilar to the innocent suspect would not, on average, also be dissimilar to the perpetrator. Similarly, a filler chosen because the face was rated as being similar to the innocent suspect would not, on average, also be similar to the perpetrator.

Figure S7 shows the scatterplot of the average similarity ratings to the innocent suspect vs. the average similarity ratings to the perpetrator for the pool of 327 filler photos. Each point reflects the average ratings (to the perpetrator on the x-axis and to the innocent suspect on the y-axis) separately for each of the photos. The data indicate that the similarity ratings are largely, but not entirely, independent. The regression line exhibits a positive slope (ideally, it would be flat), and the R^2 is .051 (ideally, it would be 0).

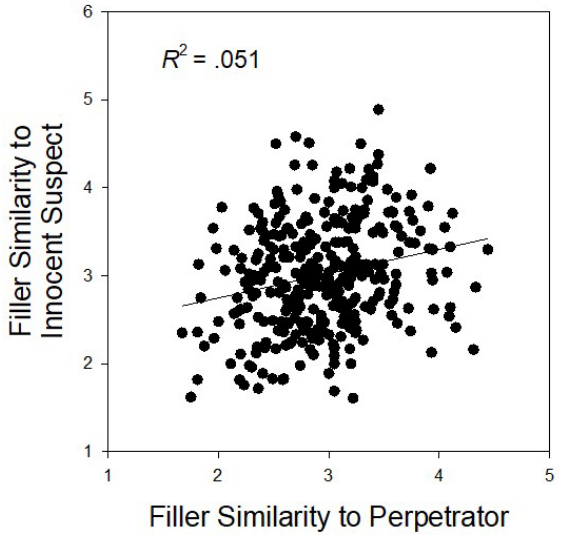


Figure S7. Average similarity ratings to the innocent suspect versus the average similarity ratings to the perpetrator for the 327 filler photos.

To create the three different filler similarity conditions in Experiment 1, we divided the ratings into thirds. For the TA lineup, the high-similarity fillers were the one-third of faces (n=109) that had the highest average similarity ratings to the innocent suspect (the upper third of points in the scatterplot in Figure S5; range: 3.30-4.89), the medium-similarity fillers were drawn from the middle third (range: 2.70-3.30), and the low-similarity fillers were drawn from the lowest third (range: 1.61-2.69). Despite the small positive correlation between the similarity ratings to the innocent suspect and the perpetrator, the data in Figure S5 indicate that filler photos in each TA similarity category (low, medium or high) spanned the full range of similarity to the perpetrator. For the TP lineup, the high-similarity fillers were the one-third of faces that had the highest average similarity ratings to the perpetrator (the rightmost third of points in the scatterplot in Figure S7; range: 3.15-4.43), the medium-similarity fillers were drawn from the middle third (range: 2.70-3.15), and the low-similarity fillers were drawn from the lowest third (the leftmost third of points in the scatterplot in Figure S7; range: 1.66-2.70).

Figure S8 summarizes the filler similarity rating data depicted in Figure S7. The first three bars in Figure S6 show the average similarity ratings to the innocent suspect for the TA fillers used in the three conditions. The middle three bars in Figure S8 show the average similarity ratings to the perpetrator for the TP fillers used in the three conditions. The last three bars show the average similarity ratings to the *perpetrator* for the TA fillers used in the three conditions. Ideally, as fillers become less similar to the innocent suspect, they would not also become less similar to the perpetrator. However, there is a small trend in that direction, reflecting the positive correlation in the scatterplot shown in Figure S7.

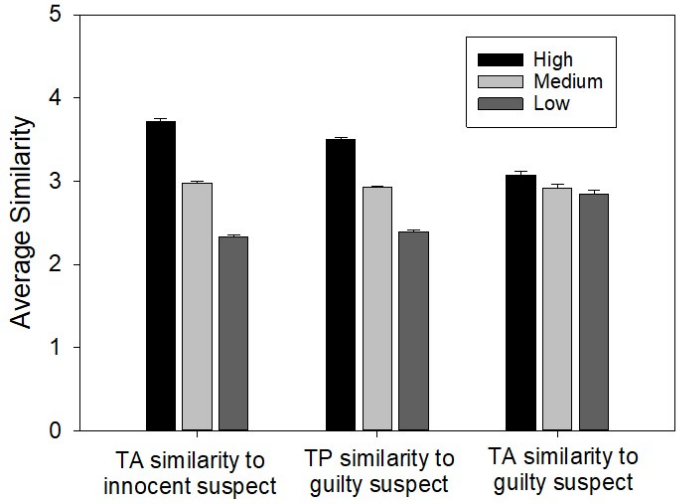


Figure S8. Average filler similarity ratings for high-, medium- and low- similarity fillers in target absent (TA) and target present (TP) conditions.

This undesirable trend suggests that our strategy of choosing dissimilar fillers from a pool of description-matched photos might result in a slight increased risk to the innocent suspect (because a filler who is dissimilar to the innocent suspect is also slightly dissimilar to the guilty and so should match memory of the perpetrator to a slightly lesser extent). Overall, however, the data suggest that a much greater increased risk to the guilty suspect, who should stand our fairly conspicuously in the low-similarity condition.

Experiment 2. For both TP and TA lineups, we used the TP filler categories (low, medium or high) from Experiment 1. Thus, in both TP and TA lineups, fillers were matched on similarity to the perpetrator.

References

M. F. Colloff, K. A. Wade, & D. Strange Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychol. Sci.* **27***,* 1227–1239 (2016).

M. F. Colloff, K. A. Wade, D. Strange, J. T. Wixted, Filler-Siphoning Theory Does Not Predict the Effect of Lineup Fairness on the Ability to Discriminate Innocent From Guilty Suspects: Reply to Smith, Wells, Smalarz, and Lampinen (2018). *Psychol. Sci.* **29**, 1552-1557 (2018).

A. M. Smith, G. L. Wells, L. Smalarz, L., J. M. Lampinen, 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). *Psychol. Sci.* **29**, 1548–1551 (2018).