Optimizing the selection of fillers in police lineups

Colloff, Melissa; Wilson, Brent; Seale-Carlisle, Travis; Wixted, John T

DOI:
10.1073/pnas.2017292118

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.

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

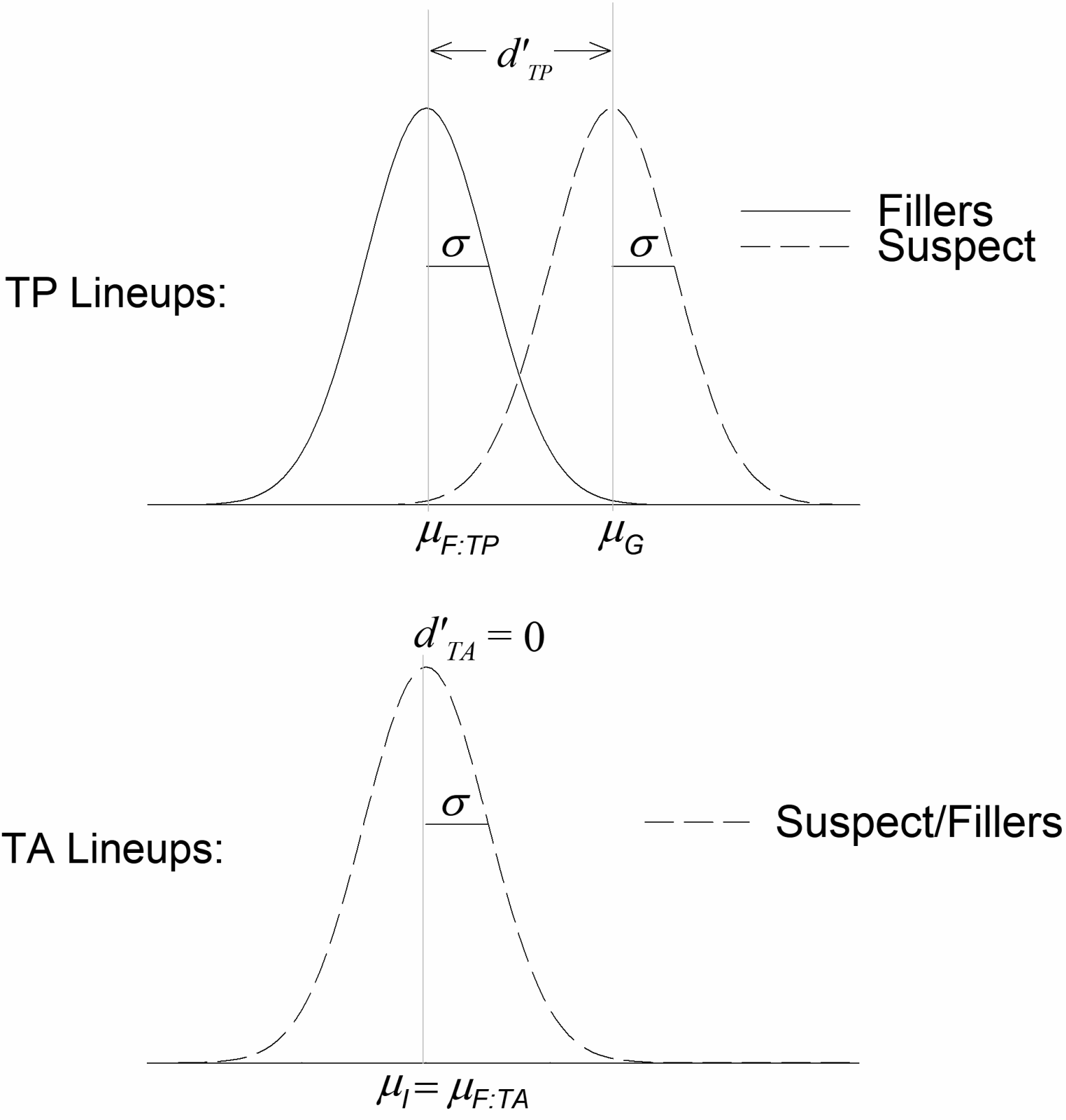
Figure 1. d'_{TP} is the difference between the mean of the TP filler distribution (e.g., $\mu_{F:TP} = 8$) and the guilty suspect distribution (e.g., $\mu_G = 20$) in standard deviation units (e.g., $\sigma = \sqrt{20}$). Here, $d'_{TP} = \frac{\mu_G - \mu_{F:TP}}{\sigma} = \frac{20-8}{\sqrt{20}} = 2.68$. 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.

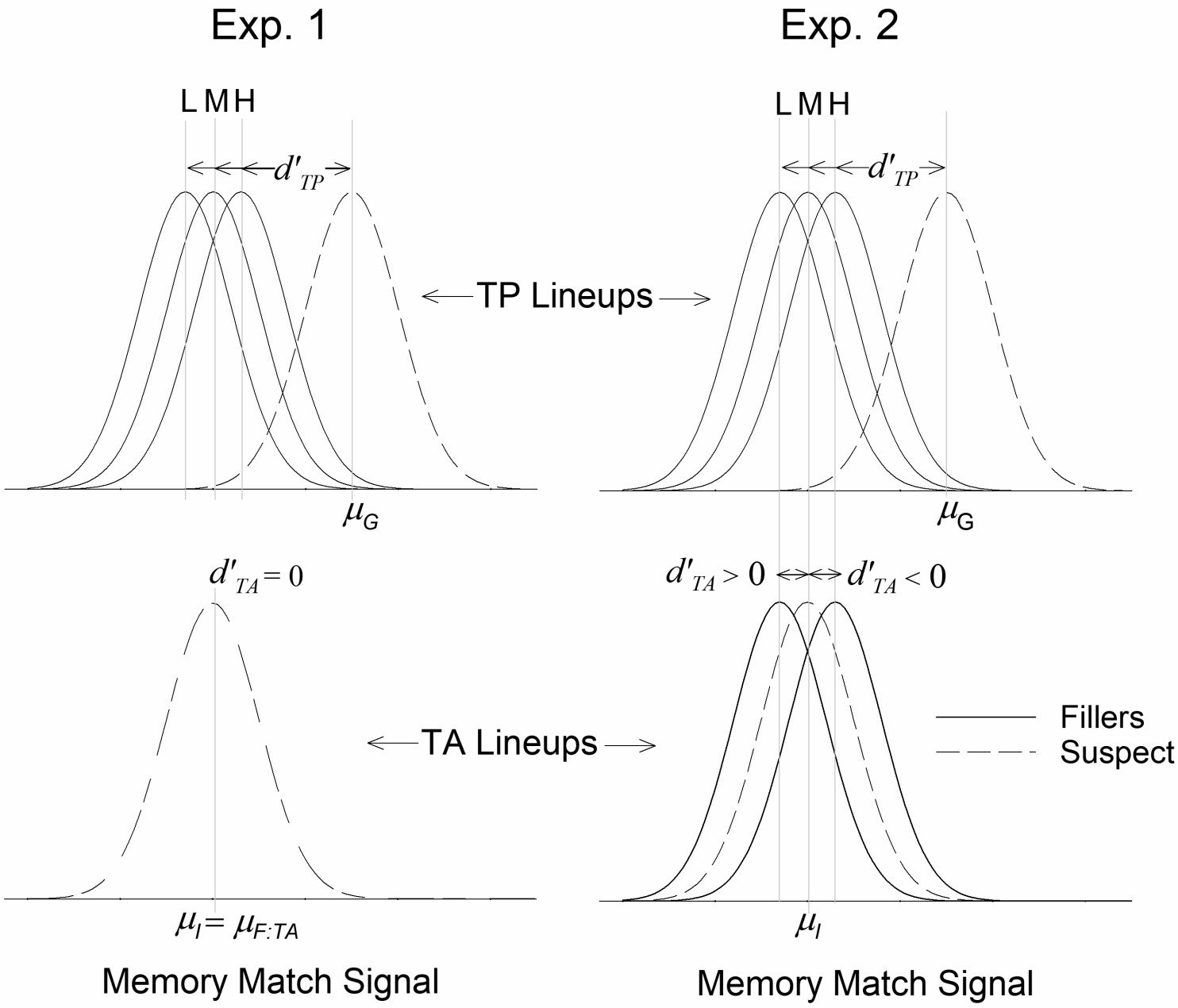
Figure 2. Exp. 1: d'_{TP} increases as filler similarity to the suspect varies from high (H) to medium (M) to low (L). By contrast, d'_{TA} theoretically remains equal to 0 because varying filler similarity to the innocent suspect in a TA lineup should not affect the degree to which those fillers match 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 ($d'_{TA} < 0$).

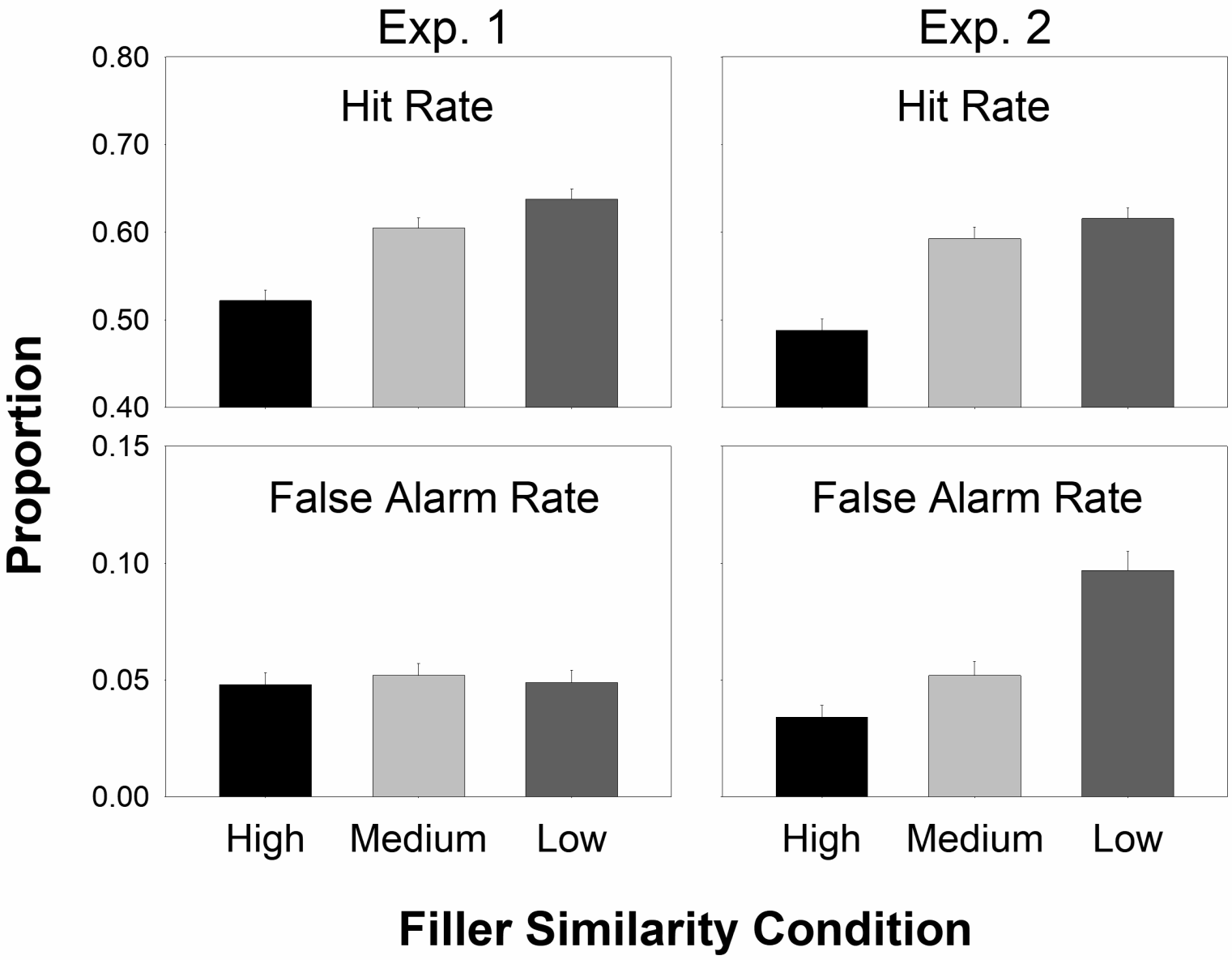
Figure 3. Exp. 1: The hit rate in the low-similarity condition was significantly higher than the hit rate in both the medium-similarity condition ($z = 2.02$, $p = .043$) and high-similarity condition ($z = 6.99$, $p < .001$). The hit rate in the medium-similarity condition was also higher than that 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-

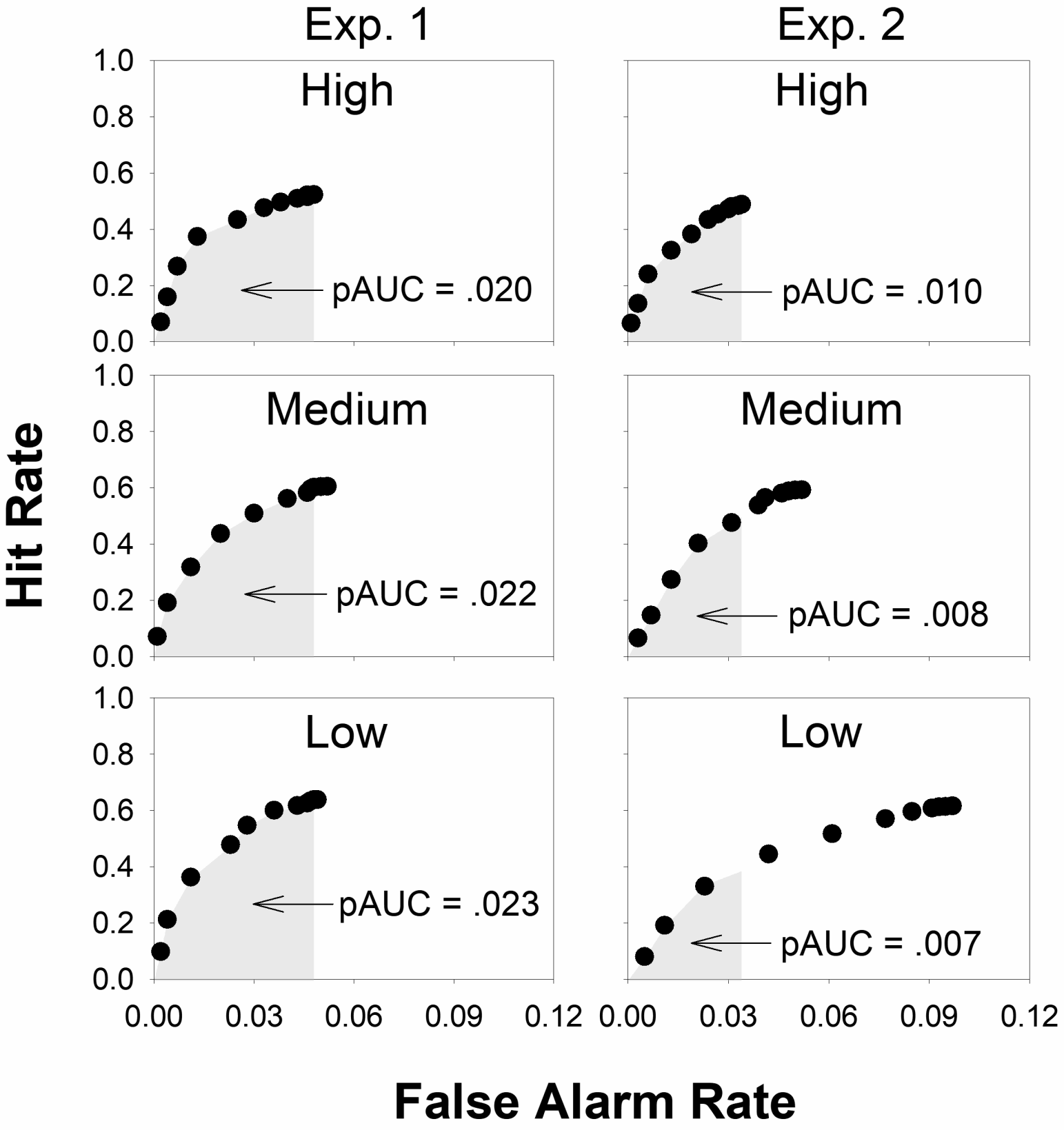
significantly higher than the hit rate in the medium-similarity condition ($z = 0.74$, $p = .461$) and significantly higher than the hit rate in the high-similarity condition ($z = 5.80$, $p < .001$). The hit rate in the medium-similarity condition was also significantly higher than that of the high-similarity 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