Optimizing the selection of fillers in police lineups

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

DOI:
10.1073/pnas.2017292118

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Colloff, M, Wilson, B, Seale-Carlisle, T & Wixted, JT 2021, 'Optimizing the selection of fillers in police lineups', *Proceedings of the National Academy of Sciences*, vol. 118, no. 8, e2017292118. https://doi.org/10.1073/pnas.2017292118

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 $1 / (m – 1) = p/(1 – p) = 1/4 = .25$.

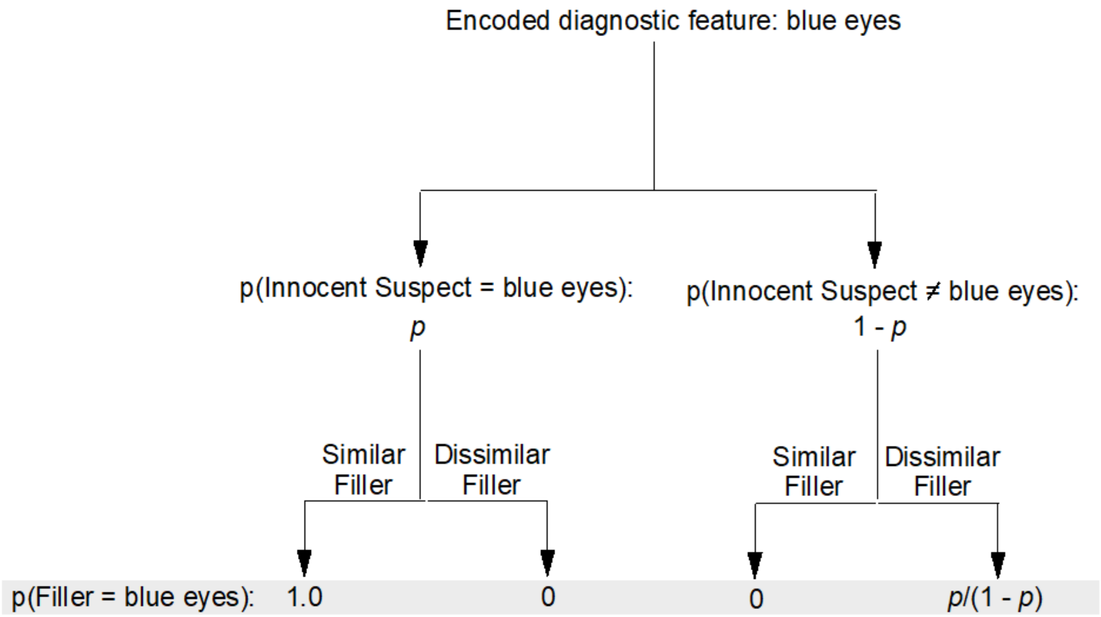

Figure S1. Conditional probabilities that a filler will match an encoded diagnostic feature (blue eyes), when fillers are chosen to be similar or dissimilar to an innocent suspect who either does or does not have blue eyes.

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*:

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

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:

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

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.

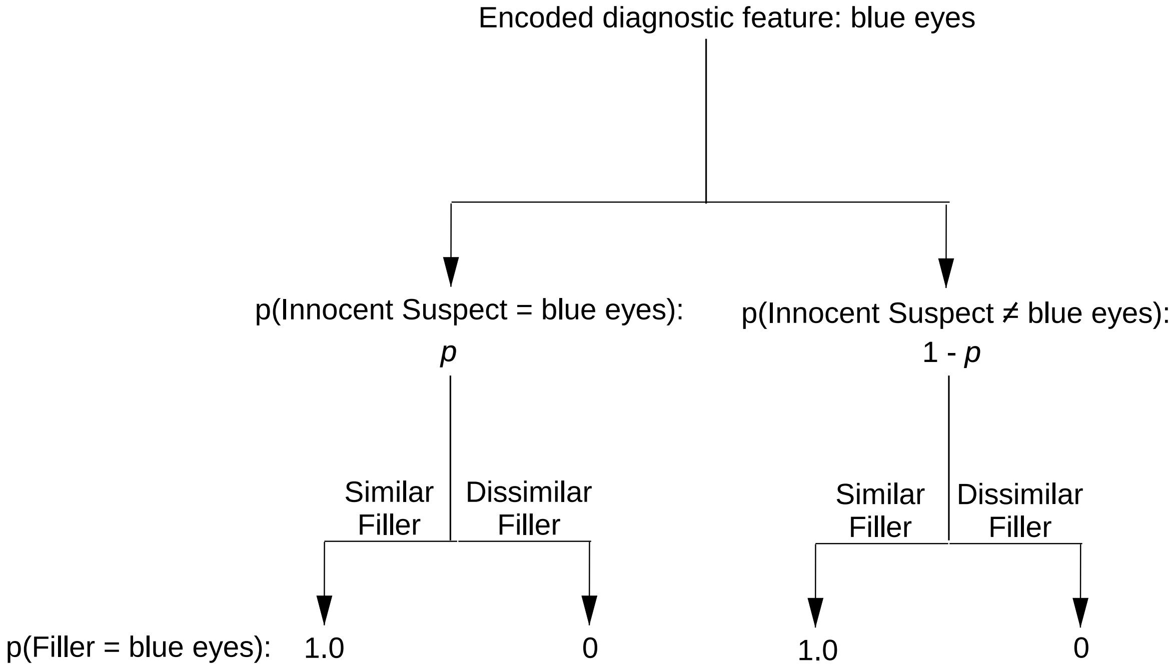
Encoded diagnostic feature: blue eyes
p(Innocent Suspect = blue eyes):
p
p(Innocent Suspect ≠ blue eyes):
1 - p
Similar Filler
Dissimilar Filler
Similar Filler
Dissimilar Filler
p(Filler = blue eyes):
1.0
0
1.0
0

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:

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

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:

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

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.

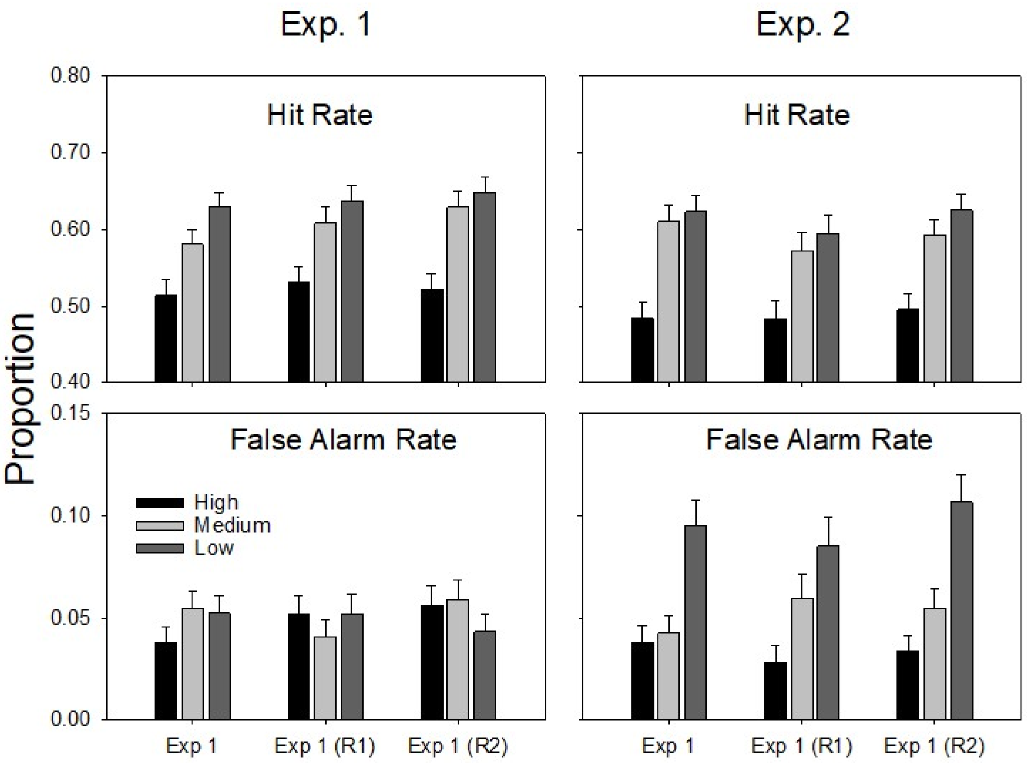

Figure S5. In Experiment 1, the hit rate in the low-similarity condition was consistently higher than the hit rate in both the medium-similarity condition and high-similarity condition. By contrast, the false alarm rate did not vary systematically across filler similarity conditions.

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).

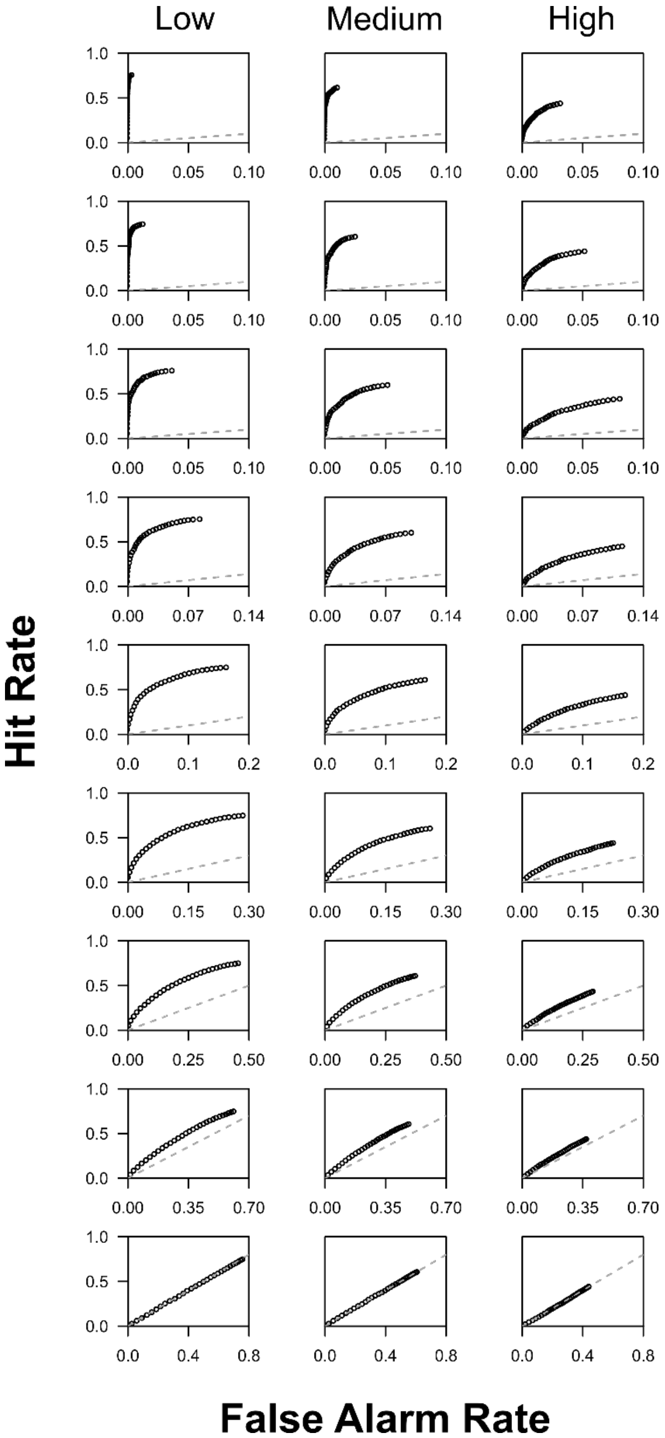

Figure S6. Model-based predictions of manipulating filler similarity from low to high (left to right) as the similarity between the innocent suspect and the perpetrator ranges from low to high (top to bottom). The middle row corresponds to the se of the median-similarity filler we used for Experiment 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).

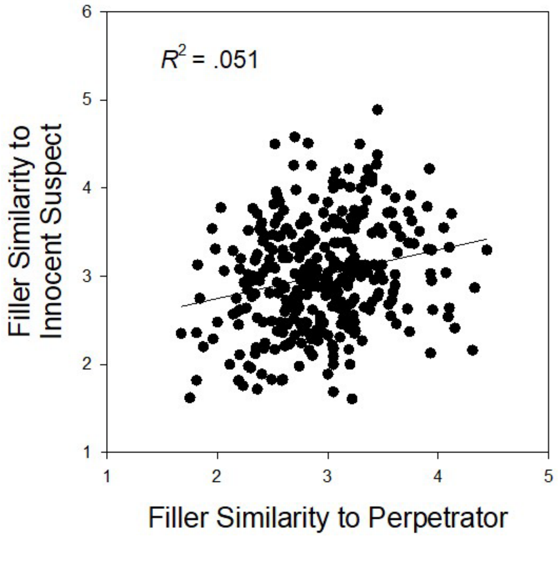

Figure S7. Average similarity ratings to the innocent suspect versus the average similarity ratings to the perpetrator for the 327 filler photos.

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.

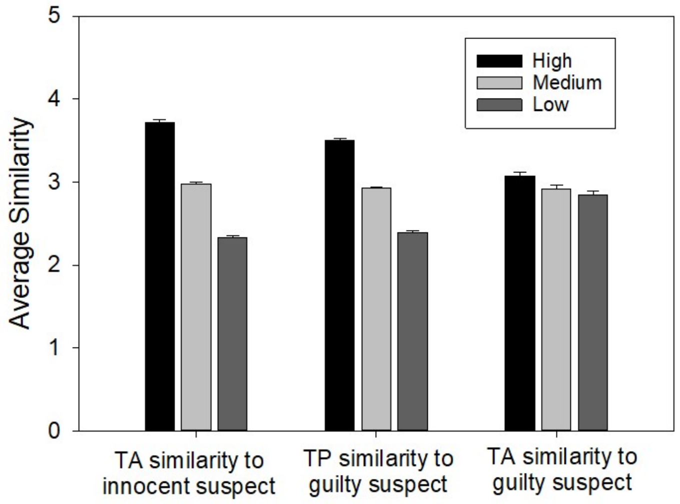

Figure S8. Average filler similarity ratings for high-, medium- and low- similarity fillers in target absent (TA) and target present (TP) conditions.

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).