# ROC Analysis Measures Objective Discriminability for any Eyewitness Identification Procedure

John T. Wixted[1] & Laura Mickes[2]

[1]University of California, San Diego

[2]Royal Holloway, University of London

## Author Note

John T. Wixted, Department of Psychology, University of California, San Diego. Laura Mickes, Department of Psychology, Royal Holloway, University of London.

This work was supported in part by the National Science Foundation [SES-1456571] to John T. Wixted and the Economic and Social Research Council [ES/L012642/1] to Laura Mickes and John T. Wixted. The content is solely the responsibility of the authors and does not necessarily reflect the views of the National Science Foundation or the Economic and Social Research Council.

Correspondence concerning this article should be addressed to John T. Wixted (jwixted@ucsd.edu).

Abstract

Which eyewitness identification procedure better enables eyewitnesses to discriminate between innocent and guilty suspects? In other words, which procedure better enables eyewitnesses to sort innocent and guilty suspects into their correct categories? The answer to that objective, theory-free question is what policymakers need to know, and it is precisely the information that ROC analysis provides. Wells et al. largely ignore that question and focus instead on whether ROC analysis accurately measures underlying (theoretical) discriminability for lineups. They argue that the apparent discriminability advantage for lineups over showups is an illusion caused by "filler siphoning." Here, we demonstrate that, both objectively and theoretically, the ability of eyewitnesses to discriminate innocent from guilty suspects is higher for lineups compared to showups, just as the ROC data suggest. Intuitions notwithstanding, filler siphoning does not account for the discriminability advantage for lineups. An actual theory of discriminability is needed to explain that interesting phenomenon.

Keywords: Eyewitness Identification; ROC Analysis; Discriminability; Bayesian Analysis; Filler Siphoning

## ROC Analysis Measures Objective Discriminability for any Eyewitness Identification Procedure

The two competing claims in this debate could not be clearer:

1. Wells, Smarlarz, and Smith (in press) claim that ROC analysis does not measure discriminability when lineups are used (because lineups have fillers) and so cannot be used to evaluate the diagnostic accuracy of that eyewitness identification procedure; instead, in their view, a Bayesian analysis, based on the joint consideration of the diagnosticity ratio and base rates, offers a better way to measure the diagnostic accuracy of lineups.

2. We claim that ROC analysis *does* measure discriminability when lineups are used (despite the presence of fillers) and is the only definitive way to measure diagnostic accuracy; moreover, just as in diagnostic medicine, a Bayesian analysis has *no bearing whatsoever* on the diagnostic accuracy of a lineup procedure.

It is hard to imagine a more urgent issue for the field to resolve because only one of these arguments can be correct, yet both approaches (ROC analysis and Bayesian analysis) are being used to adjudicate important applied questions, such as whether or not simultaneous lineups are diagnostically superior to sequential lineups. A National Academy of Sciences committee on eyewitness identification recently endorsed ROC analysis over the longstanding Bayesian approach based on the diagnosticity ratio (National Research Council, 2014). We believe they made the right call.

As shown in Figure 1, the fair lineup condition from Wetmore et al. (2015) yielded a higher ROC curve (i.e., higher discriminability) than the showup condition. Wells et al. (in press)

used the overall correct and false ID rates from those two conditions to illustrate their claim that, theoretically, lineups do not yield higher discriminability than showups – contrary to what ROC analysis suggests.

We agree with Wells et al. (in press) that the Wetmore et al. (2015) data can be used to conclusively settle the debate about what ROC analysis actually measures, so we focus much of our response on those data. We first consider objective (theory-free) discriminability, which is the only concern of policymakers. We then focus on theoretical discriminability, which is of concern to theoreticians (not policymakers) yet was the main focus of the Wells et al. critique.

**Objective Discriminability**

Imagine a group of 100 innocent and 100 guilty suspects. Wetmore et al. (2015) found that the overall correct ID rate for the showup procedure was .61. Thus, using a showup, 61 out of the 100 guilty suspects would be correctly classified as guilty. The overall false ID rate for the showup procedure was .42. Thus, using the showup, 42 out of the 100 innocent suspects would be incorrectly classified as guilty. For the lineup, the correct and false ID rates were .67 and .10, respectively, so 67 out of the 100 guilty suspects would be correctly classified as guilty and 10 of the 100 innocent suspects would be incorrectly classified as guilty. Thus, using the lineup, more of the 100 innocent suspects *and* more of the 100 guilty suspects would be correctly classified. No theoretical model – and no consideration of filler IDs – is needed to appreciate the fact that the lineup yields higher objective discriminability in that it more accurately classifies both innocent and guilty suspects than the showup. When the question concerns which procedure more accurately discriminates innocent from guilty suspects, filler IDs are simply irrelevant.

The superiority of the lineup would not change if the base rates of innocent and guilty suspects were no longer equal. Imagine, for example, a mixture of 100 guilty suspects and 1000

innocent suspects. Using the showup, 61 of the 100 guilty suspects would be correctly classified as guilty, and 420 of the 1000 innocent suspects would be incorrectly classified as guilty. Using the lineup, 67 of the 100 guilty suspects would be correctly classified as guilty, but only 100 of the 1000 innocent suspects would be incorrectly classified as guilty. Thus, no matter what the base rates, the lineup procedure more accurately classifies both innocent and guilty suspects than the showup does. This example illustrates the fact that Bayesian considerations play *no role whatsoever* in determining which eyewitness identification procedure better classifies innocent and guilty suspects into their proper categories.

*ROC Analysis Measures Objective Discriminability.* What does all of this have to do with ROC analysis? Wells et al. (in press) focused on the overall correct and false ID rates from each procedure used by Wetmore et al. (2015; namely, the rightmost ROC point for each procedure), but the same logic applies to all of the correct and false ID rates that can be achieved by either eyewitness identification procedure. A 6-person fair lineup can achieve false ID rates in the range of 0 to .167 (because always choosing from a fair target-absent lineup would result in the innocent suspect being identified 1/6 = .167 of the time). In that range, consider the achievable correct ID rates for the showup in Figure 1 (indicated by the smooth curve) and choose the point that, according to your subjective values, most appropriately balances the costs of a false ID and the benefits of a correct ID. No matter which showup point you pick, now consider the fact that the lineup – because it yields a higher ROC – can achieve a higher correct ID rate and, at the same time, a lower false ID rate than your preferred showup point. The fact that the lineup can achieve a superior outcome remains true no matter what the base rates of target-present and target-absent lineups might be and no matter what the filler ID rate might be for the lineup. As a general rule (not just for the Wetmore et al. data), the procedure that yields the higher ROC can

simultaneously achieve a higher correct ID rate *and* lower false ID rate than the procedure that yields the lower ROC. That is precisely why the procedure that yields a higher ROC is, objectively (and, we would add, unarguably), the diagnostically superior procedure.

*ROC Analysis vs. Bayesian Analysis.* Wells et al. (in press) mistakenly assert that "… ROC analysis assumes a 50/50 base rate…", but each ROC point is independent of the base rate, just as the overall correct and false ID rates are. They also say that "…a Bayesian analysis generates curves that examine posterior probabilities that the suspect is guilty across the entire range of possible base rates." However, a Bayesian analysis merely quantifies the posterior odds of guilt for a particular suspect who has been identified by an eyewitness: posterior odds = prior odds times the diagnosticity ratio (cf. Zweig & Campbell, 1993). The posterior odds of guilt can be high or low even when using an inferior diagnostic procedure, depending on how conservative or liberal the decision criterion is. ROC analysis, by contrast, tells you which diagnostic procedure does a better job of sorting innocents and guilty suspects into their correct categories. Thus, ROC analysis and Bayesian analysis *address different questions*; they are in no way competing methods for identifying the diagnostically more accurate eyewitness identification procedure. Only ROC analysis can do that.

**Underlying (Theoretical) Discriminability**

Wells et al. (in press) focus mainly on theoretical discriminability even though it is not relevant to the debate over the applied utility of ROC analysis. For example, they make the following claim: "The fact that fillers are known by the legal system to be innocent (and, hence, are not prosecuted) has nothing to do with *underlying discriminability*" (p. 7, emphasis added). Underlying discriminability is theoretical discriminability, which is of interest to theoreticians

(e.g., Wixted & Mickes, 2014; Wixted & Mickes, 2010; Mickes, Wixted, & Wais, 2007), not to policymakers.

Wells et al. (in press) argue that theoretical discriminability is not higher for lineups than showups. In their view, the apparently higher discriminability for lineups in Figure 1 is an illusion caused by the many filler IDs that lineups occasion. However, Wells et al. relied on intuition alone to analyze this theoretical issue. We now analyze the same data using signal-detection theory – the standard theory of recognition memory for more than half a century (Egan, 1958).

*A Simple Signal-Detection Model for Lineups*. According to the simplest signal-detection model (Figure 2), memory strength values for fillers, innocent suspects and guilty suspects are distributed according to Gaussian distributions with means of $\mu_{Filler}$, $\mu_{Innocent}$, and $\mu_{Guilty}$, respectively. A 6-member target-present lineup is conceptualized as 5 random draws from the Filler distribution and 1 random draw from the Guilty distribution; a 6-member target-absent lineup is conceptualized as 5 random draws from the Filler distribution and 1 random draw from the Innocent distribution. If a fair target-absent lineup is used, then $\mu_{Filler} = \mu_{Innocent}$, in which case the model reduces to a 2-distribution model. Of primary interest is the ability of eyewitnesses to collectively discriminate between innocent and guilty suspects, and that ability is represented by the distance between the means of the $\mu_{Innocent}$ and $\mu_{Guilty}$ distributions.

*Model Fits of the Wetmore et al. (2015) Data.* To fit the model, we first collapsed the Wetmore et al. (2015) data to a 3-point scale by combining confidence ratings of 6 and 7 (high confidence), 3, 4 and 5 (medium confidence), and 1 and 2 (low confidence). We collapsed the data in this manner to keep the number of parameters to be estimated reasonably low. In this model, there are three decision criteria: $c_{Low}$, $c_{Medium}$, and $c_{High}$. Using the simplest decision rule,

an ID is made if the most familiar person in a lineup exceeds the lowest decision criterion, $c_{Low}$. The corresponding confidence rating is determined by the highest confidence criterion that is exceeded.

With $\mu_{Innocent}$ set to 0 as a reference point and the standard deviations for all three distributions set to 1 for the sake of simplicity, the model has 5 parameters ($\mu_{Guilty}$, $\mu_{Filler}$, $c_{Low}$, $c_{Medium}$, and $c_{High}$). When a fair target-absent lineup is used, as was true of the Wetmore et al. (2015) study, $\mu_{Filler}$ is also equal to 0, and the model reduces to a 4-parameter model. This 4-parameter model for a fair lineup ($\mu_{Guilty}$, $c_{Low}$, $c_{Medium}$, and $c_{High}$) is the one that also applies to the showup. The theoretical discriminability score in this case is simply equal to the estimated value of $\mu_{Guilty}$, which is a $d'$ score. The parameters were estimated by adjusting them until the chi square comparing observed and predicted observations was minimized. The fit was adequate in both cases: $\chi^2_{\text{Showup}}(6) = 4.18$; $\chi^2_{\text{Lineup}}(12) = 12.54$.

Table 1 shows the parameter estimates. The most important parameter is $\mu_{Guilty}$ (which is a $d'$ discriminability estimate) because it shows the model's estimate of the ability of eyewitnesses to discriminate between innocent and guilty suspects. Critically, the estimate is much higher for the lineup. This result corresponds to what one would immediately infer by examining the objective ROC data in Figure 1 and contradicts the claim by Wells et al. (in press) that theoretical discriminability is not higher for lineups. Indeed, the smooth curves drawn through the empirical data in Figure 1 represent the predictions of this best-fitting signal-detection model.

Table 2 shows the observed values and model-based predicted values for target-present lineups, target-absent lineups, and showups. These data show that the model correctly predicts the frequent occurrence of filler IDs when lineups are used. Wells et al. (2015) call this

phenomenon "filler siphoning," and they are under the mistaken impression that it accounts for the apparent discriminability advantage for lineups in the ROC data reported by Wetmore et al. (2015). But the model very naturally predicts filler siphoning while at the same time *requiring* a discriminability advantage for lineups to adequately fit the data. To illustrate this point, we next generated hypothetical ROC data from showup and lineup models that assume identical theoretical discriminability between innocent and guilty suspects (as depicted in Figure 3). For both models, $\mu_{Guilty}$ was set to the average of the two values in Table 1 (i.e., for both, $\mu_{Guilty}$ = 1.167). Thus, the *only* difference between the two theoretical conditions is that the lineup has fillers, which results in filler siphoning. With the criterion set as shown in Figure 3, the overall correct and false ID rates (and filler ID rates) predicted by these models are presented in Table 3. Both are lower for the lineup due to filler siphoning. As shown in Figure 4, the predicted ROCs fall essentially atop one another throughout most of the false ID rate range from 0 to .167 (the obtainable range for a fair lineup). By contrast, the objective data (Figure 1) show a large discriminability advantage for lineups throughout that same range. Thus, filler siphoning cannot explain the large objective discriminability advantage for lineups, but, as noted by Wetmore et al., the diagnostic feature-detection model can (Wixted & Mickes, 2014).

*Model Fits of the Hypothetical Data in Wells et al. (in press).* Similar considerations apply to another example advanced by Wells et al. in their effort to show why ROC analysis does not measure discriminability. In their Table 2, Wells et al. present hypothetical data – reproduced here as "observed" data in Table 4 – from two lineup conditions that they believe would yield identical ROC curves despite wildly different filler ID rates. These hypothetical data show only one correct and false ID rate per procedure (identical for both), which means that their respective ROCs would have to intersect at that point but not necessarily overlap completely. Still, it seems

fair to assume that the two ROCs would be similar, and the presumably close correspondence in discriminability in Panel A and Panel B is what Wells et al. seem certain cannot be true. It cannot be true, in their view, because the filler ID rates are so much higher in Panel B than Panel A. From their perspective, this means that the data in Panel A indicate good discriminability, whereas the data in Panel B indicate poor discriminability.

To appreciate the mistake being made here, consider a simple question: Discriminability between what and what? Wells et al. (in press) do not say, and therein lies the problem with their argument. To clarify what is being discriminated from what, we specified the two signal detection lineup models (Figure 5A and 5B) that correspond to their hypothetical data. The model in Figure 5A (Model A) represents an unfair lineup in which the fillers resemble the guilty suspect to a much lesser degree than the innocent suspects do. The model in Figure 5B (Model B), by contrast, represents a nearly fair lineup in which the fillers would now be more difficult to discriminate from guilty suspects. Thus, "filler siphoning" would occur much more often in a situation represented by Model B compared to a situation represented by Model A. The data predicted by these models closely match the hypothetical "observed" data (Table 4).

If one were inclined to plot it for some reason, an ROC analysis comparing guilty suspect ID rates computed from target-present lineups to *filler* ID rates computed from target-absent lineups would show vastly higher discriminability for Model A than Model B. Indeed, Figure 6A shows the predicted filler-vs.-guilty ROCs. This is the discriminability comparison that Wells et al. appear to be focused on for reasons that make no sense to us. A much more relevant ROC comparing innocent suspect ID rates to guilty suspect ID rates actually is similar for both models (Figure 6B). Note that those two ROCs are not identical (as Wells et al. mistakenly assume must be true), but they are similar, and they intersect where they should. Thus, when Wells et al. assert

that "… ROC analysis would have us believe that discriminability is the same…" it is important to understand that they did not specify what is being compared to what. Discriminability between *innocent and guilty suspects* actually is about the same for the two situations. It is the ability to discriminate *fillers* from guilty suspects that differs wildly across the two situations.

Wells et al. inaccurately refer to "discriminability" as if it were an amorphous concept that applies to a comparison involving three distributions. However, of the three possible pairwise discriminations (fillers vs. guilty suspects, innocent suspects vs. guilty suspects, and fillers vs. innocent suspects), only one is of interest to the legal system, namely, innocent suspects vs. guilty suspects. Wells et al. should use a formal model of discriminability if they wish to theoretically conceptualize performance on eyewitness identification procedures. Then again, there really is no need to use any model to answer the applied question of whether or not the Wetmore et al. data show a lineup advantage. Objectively, they unarguably do.

References

Egan, J. P. (1958). *Recognition memory and the operating characteristic.* (Tech Note AFCRC-TN-58-51). Bloomington, IN: Indiana University, Hearing and Communication Laboratory.

Mickes, L., Wixted, J. T., & Wais, P. (2007). A direct test of the unequal variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858-865.

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

Wells, G. L., Smalarz, L., Smith, A. M. (in press). ROC Analysis of Lineups Does Not Measure Underlying Discriminability and Has Limited Value. *Journal of Applied Research in Memory and Cognition.*

Wetmore, S., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A. & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition, 4,* 8-14.

Wixted, J. T. & Mickes, L. (2010). A Continuous Dual-Process Model of Remember/Know Judgments. *Psychological Review, 117*, 1025-1054.

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

Zweig, M. H. & Campbell, G. (1993). Receiver Operating Characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*, 561-577.

*Table 1. Best-fitting parameter estimates for a fit of the signal-detection model in Figure 2 to the Wetmore et al. (2015) data in Figure 1.*

| Parameter estimate | Showup | Lineup |
|---|---|---|
| $\mu_{Guilty}$ *(d′)* | 0.70 | 1.63 |
| $c_{Low}$ | 0.35 | 1.09 |
| $c_{Medium}$ | 0.46 | 1.25 |
| $c_{High}$ | 1.45 | 2.26 |

*Table 2. Best-fitting parameter estimates for a fit of the signal-detection model in Figure 2 to the Wetmore et al. (2015) data in Figure 1.*

**Target-Present Lineup**

**A**

| | Correct ID | | Filler ID | | No ID | |
|---|---|---|---|---|---|---|
| Confidence | Observed | Predicted | Observed | Predicted | Observed | Predicted |
| 6-7 | 20 | 15.4 | 0 | 2.9 | | |
| 4-5 | 19 | 18.4 | 5 | 11.3 | 13 | 8.4 |
| 1-3 | 2 | 1.7 | 1 | 1.8 | | |

**Target-Absent Lineup**

**B**

| | False ID | | Filler ID | | No ID | |
|---|---|---|---|---|---|---|
| Confidence | Observed | Predicted | Observed | Predicted | Observed | Predicted |
| 6-7 | 4 | 1.3 | 4 | 6.7 | | |
| 4-5 | 7 | 8.3 | 48 | 41.3 | 42 | 48.9 |
| 1-3 | 1 | 1.9 | 12 | 9.6 | | |

**Showup**

**C**

| | Correct ID | | False ID | |
|---|---|---|---|---|
| Confidence | Observed | Predicted | Observed | Predicted |
| 6-7 | 17 | 16.6 | 16 | 14.1 |
| 4-5 | 25 | 27.4 | 58 | 47.9 |
| 1-3 | 4 | 2.9 | 7 | 7.6 |
| no ID | 28 | 27.1 | 112 | 123.5 |

Table 3. Predicted suspect ID, filler ID and no ID rates for the lineup and showup models depicted in Figure 3. Note the substantial filler siphoning that occurs in the case of the lineup, which reduces both suspect IDs and no IDs for the lineup compared to the showup.

| | Lineup | | | Showup | |
|---|---|---|---|---|---|
| **Suspect Status** | **p(Suspect ID)** | **p(Filler ID)** | **p(No ID)** | **p(Suspect ID)** | **p(No ID)** |
| Target Present | 0.44 | 0.38 | 0.18 | 0.57 | 0.43 |
| Target Absent | 0.11 | 0.54 | 0.35 | 0.16 | 0.84 |

Table 4. Hypothetical observed lineup data from Table 2 of Wells et al. and corresponding data predicted by the two signal-detection models in Figure 5. The predicted data in Panel A correspond to the model in Figure 5A, and the predicted data in Panel B correspond to the model in Figure 5B.

Panel A

| | Observed | | | Predicted | | |
|---|---|---|---|---|---|---|
| **Suspect Status** | **p(Suspect ID)** | **p(Filler ID)** | **p(No ID)** | **p(Suspect ID)** | **p(Filler ID)** | **p(No ID)** |
| Target Present | 0.68 | 0.01 | 0.31 | 0.67 | 0.01 | 0.32 |
| Target Absent | 0.10 | 0.01 | 0.89 | 0.10 | 0.01 | 0.88 |

Panel B

| | Observed | | | Predicted | | |
|---|---|---|---|---|---|---|
| **Suspect Status** | **p(Suspect ID)** | **p(Filler ID)** | **p(No ID)** | **p(Suspect ID)** | **p(Filler ID)** | **p(No ID)** |
| Target Present | 0.68 | 0.31 | 0.01 | 0.67 | 0.33 | 0.001 |
| Target Absent | 0.10 | 0.89 | 0.01 | 0.11 | 0.88 | 0.01 |

Figure 1. Lineup and showup ROC data from Wetmore et al. (2015). The smooth curves indicate fits to a theoretical model, which is discussed in a later section entitled "Underlying (Theoretical) Discriminability." The rightmost point on each ROC represents the overall correct and false ID rates. The dashed gray line indicates chance performance.

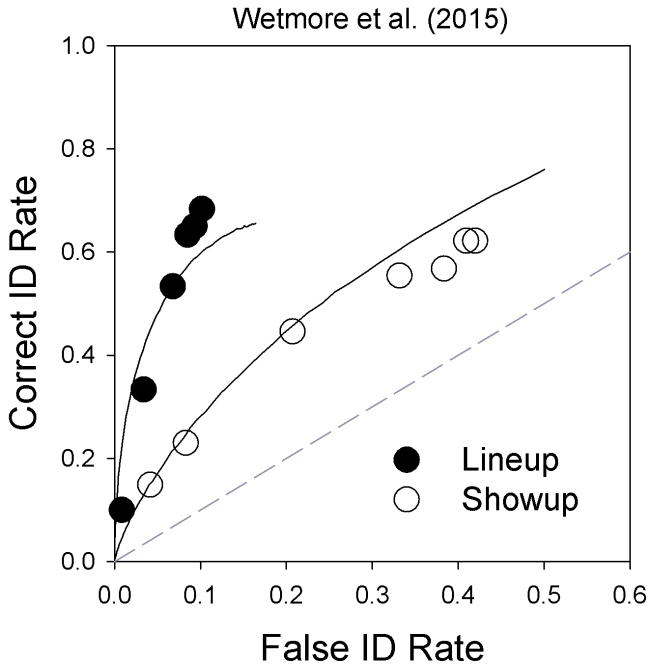

Figure 2. (**A**) Basic signal-detection model for a lineup. The model has three distributions (one each for fillers, innocent suspects and guilty suspects). For a fair lineup, the filler and innocent suspect distributions would be the same (i.e., $\mu_{Filler} = \mu_{Innocent}$). (**B**) Same model but with a decision criterion added. The simplest decision rule holds that the filler or suspect (from a target-present or target-absent lineup) who generates the strongest memory signal is identified if that memory signal exceeds c. If low, medium or high confidence ratings are taken, there would be 3 decision criteria ($c_{Low}$, $c_{Medium}$ and $c_{High}$) arranged in ascending order on the memory strength axis.

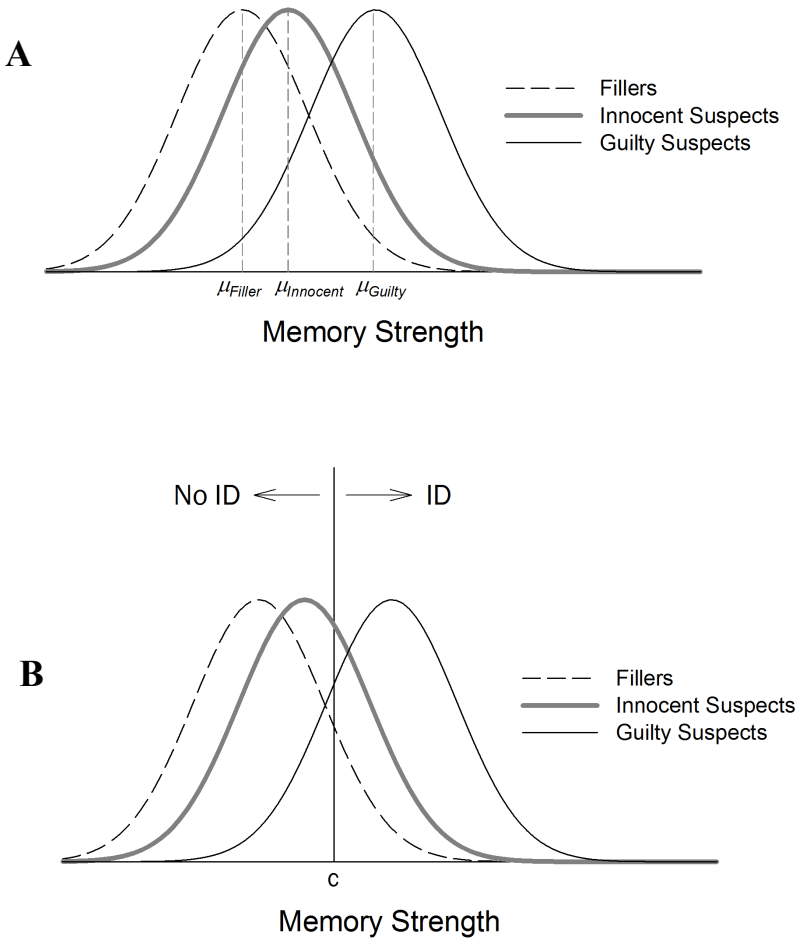

Figure 3. Signal-detection models for a showup (A) and lineup (B) with identical discriminability between innocent and guilty suspects. The dashed filler distribution for the lineup model is slightly left-shifted to make it visible. The mean of the innocent suspect distribution was set to 0, and the criterion (c) was placed at 1.0 on the memory strength axis (i.e., one standard deviation above the mean of the innocent suspect distribution).

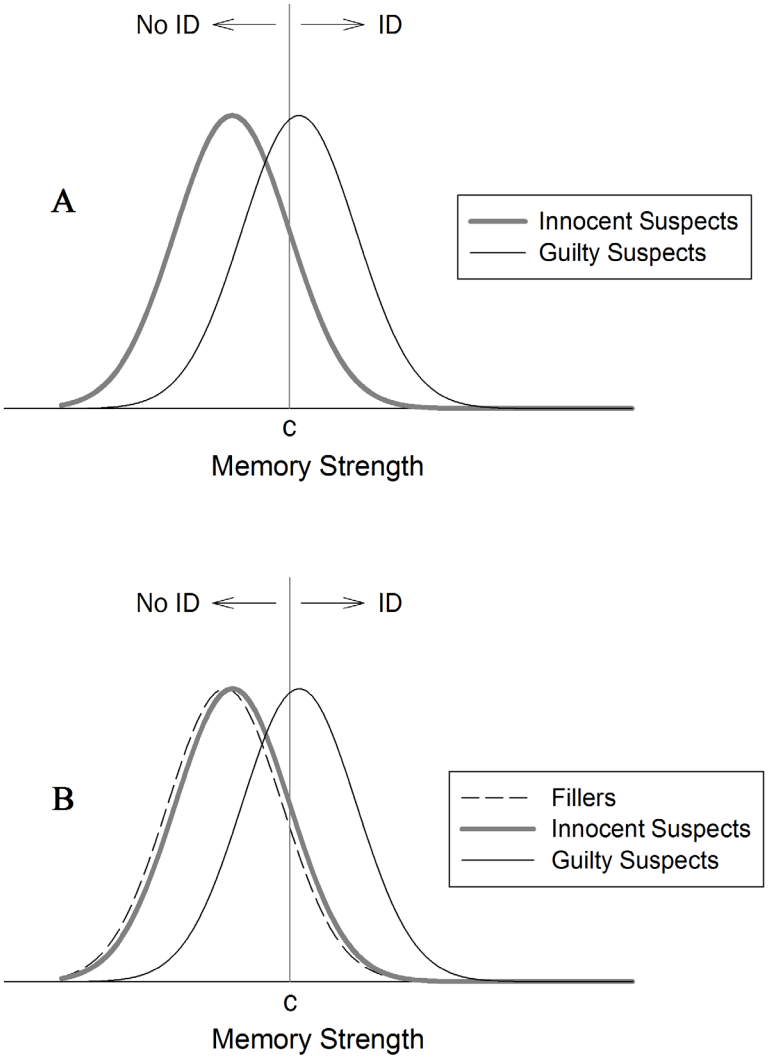

Figure 4. Predicted lineup and showup ROCs when discriminability between innocent and guilty suspects is equated for both procedures (as represented by the corresponding lineup and showup models shown in Figure 2).

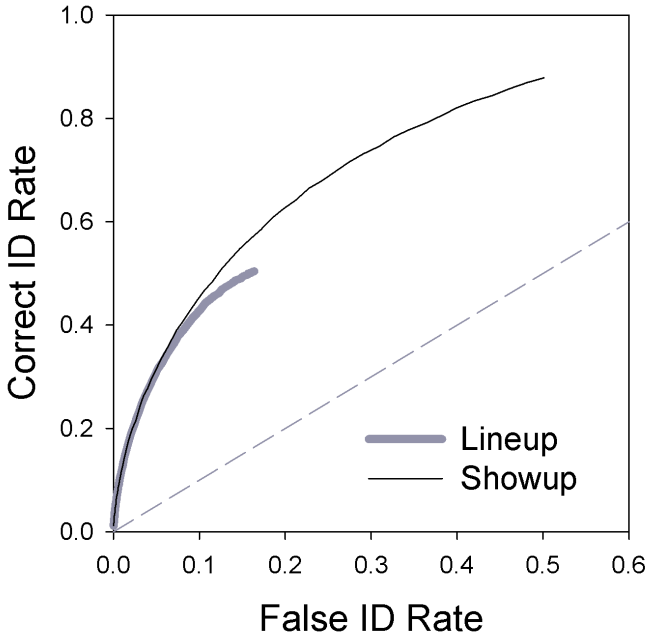

Figure 5. (A) Signal-detection lineup model that corresponds to hypothetical lineup data in Table 4A. (B) Signal-detection lineup model that corresponds to hypothetical lineup data in Table 4B. To keep the correct and false ID rates the same, the criterion in Model B is shifted to the left (to compensate for filler siphoning, which would otherwise yield lower correct and false ID rates for Model B).

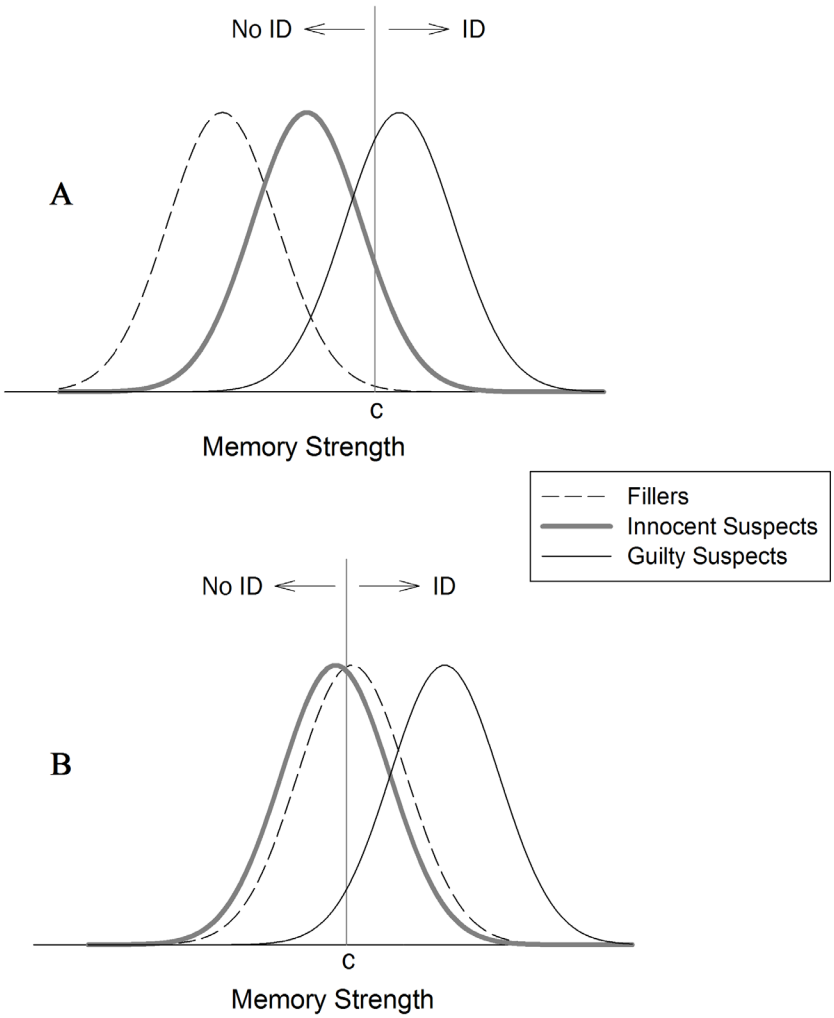

Figure 6. Predicted ROC data for Models A and B in Figure 5. (**A**) The data show a plot of the correct ID rate from target-present lineups vs. the filler ID rate from target-absent lineups. (**B**) The data show a plot of the correct ID rate from target-present lineups vs. the false ID rate (i.e., innocent suspect ID rate) from target-absent lineups.