The Role of Estimator Variables in Eyewitness Identification

Carolyn Semmler[1], John Dunn[2], Laura Mickes[3] & John T. Wixted[4]

[1]University of Adelaide

[2]University of Western Australia

[3]Royal Holloway, University of London

[4]University of California, San Diego

Author Note

Carolyn Semmler, School of Psychology, University of Adelaide; John Dunn, School of Psychological Science, University of Western Australia; Laura Mickes, Department of Psychology, Royal Holloway, University of London. John T. Wixted, Department of Psychology, University of California, San Diego.

Abstract

Estimator variables are factors that can affect the accuracy of eyewitness identifications but that are outside of the control of the criminal justice system. Examples include (1) the duration of exposure to the perpetrator, (2) the passage of time between the crime and the identification (retention interval), (3) the distance between the witness and the perpetrator at the time of the crime. Suboptimal estimator variables (e.g., long distance) have long been thought to reduce the reliability of eyewitness identifications (IDs), but recent evidence suggests that this is not true of IDs made with high confidence and may or may not be true of IDs made with lower confidence. The evidence suggests that while suboptimal estimator variables decrease discriminability (i.e., the ability to distinguish innocent from guilty suspects), they do not decrease the reliability of IDs made with high confidence. Such findings are inconsistent with the longstanding "optimality hypothesis" and therefore require a new theoretical framework. Here, we propose that a signal-detection-based likelihood ratio account – which has long been a mainstay of basic theories of recognition memory – naturally accounts for these findings.

Keywords: eyewitness identification, confidence and accuracy, estimator variables, system variables

Public Significance Statement: This study challenges the assumption that poor witnessing conditions lead to unreliable eyewitness identification evidence. In particular, it shows that witnesses viewing a perpetrator over a long distance, but providing a confident identification can be accurate. We show how formal recognition memory theories can account for this result.

The Role of Estimator Variables in Eyewitness Identification

According to the Innocence Project (2017), of the 350 wrongful convictions that have been overturned by DNA evidence to date in the United States, approximately 70% involved the misidentification of an innocent suspect by an eyewitness. Research-based efforts to better understand and perhaps reduce that problem have long been guided by the distinction between system variables and estimator variables (Wells, 1978). System variables are factors that are under the control of the legal system, such as the wording of lineup instructions given at the time an eyewitness identification is attempted, whereas estimator variables are factors that are not under the control of the legal system, such as the distance between the perpetrator and the witness at the time of the crime. Here, we focus on estimator variables, a number of which are widely believed to affect the reliability of eyewitness identifications. We argue that estimator variables do not appreciably affect the reliability of identifications made with a particular level of confidence (particularly high confidence), and we offer a signal-detection-based theory of eyewitness identification taken directly from the basic recognition memory literature to account for that surprising result. We offer this theory as an alternative to the "optimality hypothesis" (Deffenbacher, 1980, 2008), which holds that confidence becomes less indicative of accuracy under suboptimal estimator variable conditions. Our signal-detection-based theory consists of a standard likelihood ratio model of recognition memory. This widely used theoretical framework naturally predicts that as the conditions of encoding and retrieval become less favorable, overall accuracy will decline, but the accuracy of a suspect ID made with a particular level of confidence (e.g., the accuracy of a high-confidence suspect ID) will remain unchanged. To illustrate how the model works, we develop it in relation a detailed reanalysis of data from an experiment investigating the effect of distance between a witness and a target individual.

The Prevailing View of Estimator Variables

We begin by taking stock of the prevailing view of the effect of estimator variables on the reliability of eyewitness identification. Consider a recent amicus brief filed by the American Psychological Association (APA) in the case of Commonwealth of Pennsylvania v. Walker (2014), which explains why the APA has standing to weigh in on this matter:

> The American Psychological Association (APA) is the leading association of psychologists in the United States. A nonprofit scientific and professional organization, it has approximately 155,000 members and affiliates, including the vast majority of psychologists holding doctoral degrees from accredited universities in the United States…APA has a rigorous approval process for amicus briefs, the touchstone of which is an assessment of whether the case is one in which there is sufficient scientific research, data, and literature relevant to a question before the court that APA can usefully contribute to the court's understanding and resolution of that question.

On page 9-13, the brief includes a paragraph about each of 6 estimator variables widely believed to affect the reliability of eyewitness identification. The first sentence or two of each paragraph is quoted next in order to succinctly convey the scientific consensus about these variables:

> *a. Passage of Time*. Empirical research establishes that as time passes between an event and an associated identification, the identification becomes increasingly unreliable—put simply, the memory "decays."
>
> *b. Witness Stress*. The level of stress experienced by an eyewitness at the time of exposure to the perpetrator can also affect the reliability of a subsequent identification.
>
> *c. Exposure Duration*. Studies have similarly demonstrated that the reliability of an eyewitness identification diminishes when the witness sees the perpetrator for only a short period of time.
>
> *d. Distance*. As everyday experience tells us, clarity of vision decreases with distance. Experimental research provides specifics about this relationship between distance and the ability to identify faces. The research reveals that—for people with normal vision— this ability begins to diminish at approximately 25 feet, and nearly disappears by approximately 150 feet.
>
> *e. Weapon Focus*. Weapon focus "refers to the visual attention eyewitnesses give to a perpetrator's weapon during the course of a crime"—attention that is "expected … [to] reduce his or her ability to later recall details about the perpetrator or to recognize the perpetrator."
>
> *f. Cross-Race Bias*. Finally, extensive empirical research demonstrates that eyewitnesses are more accurate at identifying perpetrators of their own race than those of a different race.

On page 13, the upshot of the scientific consensus is summarized as follows:

> The point is simply that eyewitness reliability— the linchpin of admissibility under this Court's precedent—is…determined by numerous factors identified by scientific research, many of which (the estimator variables) have nothing to do with the conduct of law enforcement. Eyewitness testimony can be unreliable even where there is no state-created suggestiveness.

Perhaps not surprisingly, these conclusions about the deleterious effect of certain estimator variables on the reliability of eyewitness identification have been increasingly embraced by the legal system. For example, in State v. Almaraz (2013), a ruling from the Idaho State Supreme Court stated the following about the effect of estimator variables:

> In contrast, the research established that the following estimator variables diminish the reliability of a witness's identification: (1) stress; (2) the use of a visible weapon during a crime;[6] (3) the shorter the duration of a criminal event; (4) the greater the distance and the poorer the lighting conditions; (5) increased levels of intoxication; (6) the use of disguises during the crime and changes in facial features between the time of initial observation and a subsequent identification; (7) the greater the period of time between observation and identification to law enforcement;[7] (8) race-bias;[8] and (9) feedback from co-witnesses confirming the identification of a perpetrator (pp. 10-11).

Essentially the same interpretation can be found in jury instructions that are now used in the states of New Jersey and Massachusetts. For example, according to Papailiou, Yokum & Robertson (2015), New Jersey jury instructions admonish jurors that:

> To decide whether the identification testimony is sufficiently reliable evidence to conclude that this defendant is the person who committed these offenses charged, you should evaluate the testimony of the witness in light of the factors for considering credibility that I have already explained to you. In addition, you should consider the following factors that are related to the witness, the alleged perpetrator, and the criminal incident itself (p. 12).

The instructions then consist of one paragraph each about several estimator variables. We reproduce the listed estimator variables and a key sentence or two from each paragraph below:

> a. Stress: Even under the best viewing conditions, high levels of stress can reduce an eyewitness's ability to recall and make an accurate identification. Therefore, you should consider a witness's level of stress and whether that stress, if any, distracted the witness or made it harder for him or her to identify the perpetrator.
>
> b. Duration: The amount of time an eyewitness has to observe an event may affect the reliability of an identification.
>
> c. Weapon Focus: You should consider whether the witness saw a weapon during the incident and the duration of the crime. The presence of a weapon can distract the witness and take the witness's attention away from the perpetrator's face. As a result, the presence of a visible weapon may reduce the reliability of a subsequent identification if the crime is of short duration.
>
> d. Distance: A person is easier to identify when close by. The greater the distance between an eyewitness and a perpetrator, the higher the risk of a mistaken identification.
>
> e. Lighting: Inadequate lighting can reduce the reliability of an identification. You should consider the lighting conditions present at the time of the alleged crime in this case.
>
> f. Disguises/Changed Appearance: The perpetrator's use of a disguise can affect a witness's ability both to remember and identify the perpetrator (pp. 12,13).

This interpretation of how estimator variables affect the reliability of eyewitness identification probably comes as no surprise to the reader because it accords with textbook treatments of the issue. However, in contrast to the prevailing view, our proposal is that none of these estimator variables appreciably affects the reliability of an ID made with a particular level of confidence – for good theoretical reasons (and contrary to the optimality hypothesis). Ironically, that key result – namely, that reliability for a given level of confidence is largely unaffected by estimator variables – may have been overlooked because of the field's once negative view of the information value of confidence. For example, the New Jersey jury instructions presented in Papailiou et al. (2015) provides the following (very common) statement about the information value of eyewitness confidence:

> As I explained earlier, a witness's level of confidence, standing alone, may not be an indication of the reliability of the identification. Although some research has found that highly confident witnesses are more likely to make accurate identifications, eyewitness confidence is generally an unreliable indicator of accuracy (p. 13).

In contrast to the longstanding view that confidence is not predictive of accuracy, a great deal of evidence has now accumulated demonstrating that on an initial eyewitness identification test using a fair lineup, confidence is undeniably predictive of accuracy (e.g., Brewer & Wells, 2006). Moreover, IDs made with high confidence are generally highly accurate under those conditions (Wixted et al., 2015; Wixted & Wells, 2017). This new understanding sets the stage for another surprising claim that we make here: for a given level of confidence, estimator variables have little to no effect on the reliability of eyewitness identifications. The main point of our article is that this non-intuitive result is naturally predicted by a standard likelihood ratio theory of recognition memory (heretofore applied mainly to list-memory paradigms). This longstanding framework is at odds with the optimality hypothesis. Next, we briefly summarize both of these theoretical accounts of the confidence-accuracy relationship.

Likelihood Ratio Models

Recognition memory models need to account for basic empirical phenomena and the most extensively studied is the mirror effect. This is the finding that if a condition gives better recognition performance then it will increase the ability of the observer to respond "old" when the item was presented and "new" when the item was not presented (Glanzer & Adams, 1985). Likelihood ratio models account for the mirror effect by modifying the assumed decision axis so that an observer evaluates the likelihood of the item being old or new on the basis of a computed log likelihood that includes both the familiarity of the item and the background or contextual information about the study conditions (Glanzer, Adams, Iverson, & Kim, 1993). This is in contrast to strength or familiarity based accounts that assume that the observer considers only the familiarity of the item in relation to a set of criteria (in a rating task) or a criterion (in a forced choice task) placed along the strength axis. In sum, these models assume that the observer evaluates an odds ratio associated with the test item, not its level of familiarity. The odds ratio is equal to the likelihood that the item was drawn from the target distribution divided by the likelihood that it was drawn from the lure/filler distribution, or; $LR(x) = (f(x)|S2)/(f(x)|S1)$, where S2 is the height of the target distribution and S1 is the height of the filler distribution. Further, as elaborated below, the likelihood ratio account also predicts that observers will attempt to maintain a constant ratio over weak and strong conditions. They achieve this by adjusting their decision criteria (Stretch & Wixted, 1998). The use of the log-likelihood when unequal variances occur in the target and filler distributions complicates the picture, however, we consider the equal variance case here and show that it provides a good approximation to the data.

The Optimality Hypothesis

The optimality hypothesis is not a statement of the reliability of eyewitness identifications, per se, but is instead a statement about the *correlation* between confidence and accuracy under favorable vs. unfavorable information processing conditions. The proposal is that the confidence-accuracy correlation should vary directly with the optimality of those conditions (Deffenbacher, 1980). In other words, the correlation should be higher when (for example) exposure to the perpetrator is long, distance between the witness and perpetrator is short, and stress is low compared to when exposure to the perpetrator is short, distance between the witness and perpetrator is long, and stress is high. Technically, but improbably, the reduced correlation associated with poorer information processing conditions could arise because eyewitness identifications become perfectly accurate under poor information processing conditions regardless of confidence (100% correct for IDs made with low confidence and 100% correct for IDs made with high confidence). In that case, the confidence-accuracy correlation would drop to 0 when conditions were poor, in accordance with the optimality hypothesis.

Although technically possible, this is certainly not how the optimality hypothesis has been used to help understand the reliability of eyewitness identification. Instead, the argument has been made that under suboptimal conditions of encoding and retrieval, the trustworthiness of eyewitness identifications can be expected to decrease. For example, Deffenbacher (2008) argued that under poor information processing conditions, "…not only will familiar faces be judged to be unfamiliar and unfamiliar faces be judged as familiar more frequently, but the same confidence rating is also more likely to be applied both to a judgment that a face seen before is indeed familiar and to a judgment that another face, never seen before, is also familiar" (p. 819). The optimality hypothesis therefore helps to explain the widespread belief that the usefulness of

eyewitness confidence as an indicator of accuracy – including high confidence – will decrease as information processing conditions get worse.

The Measure of Interest

Prior accounts of the role of estimator variables on the reliability of eyewitness identification have been complicated by the use of the term "accuracy," the meaning of which is not as obvious as intuition might suggest. Accuracy can be measured in an overall sense that takes into account errors of any kind, including failures to identify the perpetrator as well as false identifications of the innocent (measured by overall percent correct, *d'*, area under the ROC, etc.) or in a more specific sense that focuses on the trustworthiness of an identification of a suspect (which does not take into consideration the error of failing to identify the perpetrator). The general measure of accuracy is what we will henceforth refer to as "discriminability," and the more specific measure of accuracy is what we will refer to as "reliability."

Mickes (2015) pointed out that, as a general rule, when the question concerns a system variable, such as simultaneous vs. sequential lineup format, a measure of discriminability (e.g., *d'*, or better yet, area under the ROC curve) usually provides the answer. The same is true in the field of medicine when the question concerns a medically relevant "system variable," such as which diagnostic test for diabetes is the best one to use. By contrast, when the question concerns an estimator variable, such as the effect of distance on the accuracy of a suspect ID, a measure of positive predictive value (PPV) usually provides the answer (Schum, 1981). PPV is the probability that a suspect identification that has been made by an eyewitness is correct. The same considerations also apply to the field of medicine when the question concerns the effect of a medically relevant "estimator variable" on the outcome of a diagnostic test, such as the effect of ethnicity on the likelihood of actually having diabetes given a positive test result.

The key point is that *d′* and PPV measure different things and answer different questions, yet that fact is obscured by the use of the term “accuracy” to refer to both. For example, Cutler (2006) documented the effects of a number of estimator variables on the accuracy of eyewitness identification (e.g., same- vs. cross-race, exposure duration, retention interval, presence vs. absence of a weapon, and eyewitness stress). Research findings were reviewed suggesting that all of these variables can reduce the overall accuracy of eyewitness identifications – a measure that includes the error of failing to identify the perpetrator – and that there is a consensus among eyewitness experts to that effect. Stated differently, researchers agree that these variables reduce discriminability (i.e., the ability of a witness to distinguish between innocent and guilty suspects). Cutler (2006) then made the argument that “Therefore, individuals who must evaluate eyewitness identifications-investigators, attorneys, judges, and jurors-would benefit from education about the effects of estimator variables on identification accuracy” (p. 339). However, as noted by Mickes (2015), judges and jurors are not interested in a measure of discriminability. Instead, they are interested in the *reliability* of a suspect identification that was made by an eyewitness who will end up testifying at trial. In other words, they are interested in PPV (a measure that does not take into account the error of failing to identify the perpetrator).

These considerations reveal why it is essential to be clear about whether the accuracy measure of interest is discriminability or reliability. The basic components of both measures consist of correct suspect IDs (identifying the guilty perpetrator from a lineup) and false suspect IDs (misidentifying the innocent suspect from a lineup). Thus, we describe those two constituent measures next so that we can then illustrate the difference between discriminability and PPV. In police department field studies, these two measures cannot be directly computed, but they can be

directly computed in lab studies because the experimenter knows if the suspect in the lineup is innocent or guilty.

In a typical laboratory task, each observer first watches a perpetrator commit a simulated crime and is later presented with an array of *n* stimuli (*n* = 6 or 8, typically). On target-present trials, the array consists of 1 target (a photo of the perpetrator) and *n* – 1 physically similar fillers (or foils). On target-absent trials, the array consists of *n* fillers. One of those fillers can be designated as the innocent suspect even though, in a fair lineup, the innocent suspect is, from the point of view of the witness, just another filler (i.e., just another person who physically resembles the perpetrator but who is not actually the perpetrator). The observer's job is to indicate whether the target is present in the array and, if so, to specify which person it is. On target-present trials, the observer can correctly identify the target (a correct ID, or a "hit"), incorrectly identify a filler, or incorrectly reject the array. On target-absent trials, the observer can incorrectly identify the filler who serves as an innocent suspect (a false suspect ID, or a "false alarm"), incorrectly identify a distractor, or correctly reject the array.

A measure of discriminability is based on the hit rate and the false alarm rate. The hit rate is the proportion of observers presented with a target-present array who correctly identify the guilty suspect (i.e., number of guilty suspect IDs divided by the number of target-present lineups), and the false alarm rate is the proportion of observers presented with a target-absent array who incorrectly identify the innocent suspect (i.e., number of innocent suspect IDs divided by the number of target-absent lineups). In the common case in which target-absent lineups consist of *n* fillers with no one designated as the innocent suspect, the false alarm rate can be equivalently estimated by counting all filler IDs from target-absent lineups and dividing by *n*.

Participants who do not identify the suspect in target-present and target-absent either identify a filler or reject the lineup. Neither outcome endangers the suspect in the lineup.

The focus on suspect IDs to measure discriminability and reliability (PPV) does not imply that filler IDs or lineup rejections are of no interest to the legal system. Those outcomes are clearly of interest because, for example, they are somewhat indicative of innocence (Wells, Yang & Smalarz, 2015; Wixted & Wells, 2017). Nevertheless, suspect identifications are of most interest because a suspect who is identified by an eyewitness is imperiled, whether the suspect is innocent or guilty. Thus, the discriminability measure of interest here is the ability to discriminate innocent from guilty *suspects*, and the PPV measure of interest here is the probability that a *suspect* who has been identified with a particular level of confidence is guilty.

One additional measurement detail should be briefly addressed before we consider the difference between discriminability and reliability in more detail. As noted by Juslin et al. (1996), judges and jurors are primarily concerned with the reliability of a suspect ID that has been made with the particular level of confidence, not with the *relationship* between confidence and accuracy. For example, if the witness testifying at the trial expressed 100% confidence in an ID of the suspect from a lineup, their question is: “How trustworthy is an ID made with 100% confidence?” Their question is not: “What is the difference in the reliability associated with IDs made with 60% confidence vs. 100% confidence”? The optimality hypothesis and the likelihood ratio account we advance here speak to the relationship between confidence and accuracy. However, the likelihood ratio account also makes it clear that the best way to assess the confidence-accuracy relationship is not by using a correlation coefficient. Instead, the best way to understand the confidence-accuracy relationship is to plot PPV as a function of confidence, which Mickes (2015) referred to as a “confidence-accuracy characteristic” (CAC) plot. Such a

plot not only clearly reveals the relationship between confidence and accuracy, it also provides the information of most interest to judges and jurors, namely, the reliability of an ID made with a particular level of confidence.

Discriminability vs. PPV

The number of hits (guilty suspect IDs) and false alarms (innocent suspect IDs) are used to compute both discriminability and PPV. Figures 1 and 2 illustrate what these two measures capture and how they differ from each other. The left panel of Figure 1 shows 10 individuals in Population 1, 5 falling into Category G and 5 falling into Category I. Those in Category G might be people who have diabetes or who are guilty suspects, whereas those in Category I might be people who do not have diabetes or who are innocent suspects. Because, in this example, the number of Gs equals the number of Is, this is an equal base-rate scenario. Most lineup experiments conducted in the laboratory use equal base rates because half the participants are tested using a target-present lineup and half are tested using a target-absent lineup. Before the diagnostic test is administered (i.e., before the diabetes test is administered or before the lineup is administered), we have no eyewitness-based information about who falls into Category G or Category I. After the test is administered, we have updated eyewitness-based diagnostic information about that.

Discriminability refers to how well the two groups are sorted into their correct categories based on the results of the diagnostic test. High discriminability is characterized by a high hit rate and a low false alarm rate. In this example, the hit rate is .60 (3 of 5 Gs corrected sorted into Category G), and the false alarm rate is .20 (1 of 5 Is incorrectly sorted into Category G). The goal of a police chief is to use the lineup procedure that best sorts innocent and guilty suspects into their correct categories. That is, the police chief is dealing with lineups as a system variable,

and the goal is to find the lineup that simultaneously yields the highest hit rate (so that guilty suspects can be prosecuted) and the lowest false alarm rate (to avoid prosecuting innocent suspects). If one lineup format yields both a higher hit rate and a lower false alarm rate than another, then it would be the objectively superior procedure and would clearly be the one to use. In ways that are described in more detail later, the hit and false alarm rates can be combined to create a single measure of discriminability, the most common ones of which are *d′* (a theory-based measure) or area under the receiver operating characteristic (an atheoretical measure). For the moment, assume that a hit rate of .60 and a false alarm rate of .20 yields a *d′* of approximately 1.0 (we will precisely define *d′* in a later section).

As noted by Mickes (2015), judges and jurors considering an eyewitness who identified a suspect have no control over the lineup procedure that was used, and their question has nothing to do with discriminability. From their perspective, lineup format is an estimator variable, and they are trying to judge the reliability of a suspect ID that was already made. In other words, their question is not "which procedure better sorts innocent and guilty suspects into their correct categories?" but instead is "given that this witness made a suspect ID, what is the probability that the ID is correct?" This is a question about PPV, and it is measured by considering the subset of people who were positively identified (3 Gs and 1 I in the example shown in the left panel of Figure 1). PPV is equal to the number of hits divided by the total number of positive IDs (hits plus false alarms). Thus, for this example, PPV = .75, which means that the probability that the ID is correct is .75 (and the probability is .25 that an innocent suspect was identified instead).

PPV can also be computed from the hit rate (HR) and the false alarm rate (FAR), where HR equals the number of hits divided by the number of target-present lineups (nTP), and FAR equals the number of false alarms divided by the number of target-absent lineups (nTA). As

noted above, the base rate, or prevalence (*p*), of target-present lineups in lab studies is usually .50, where *p* = nTP / (nTP + nTA). For the more general case involving any base rate, PPV = *p*HR / [*p*HR + (1-*p*)*FAR]. For the typical equal-base-rate situation (i.e., *p* = .50), *p* drops out of the equation, so PPV = HR / (HR + FAR). This equal-base-rate version is the PPV value typically analyzed in lineup studies that report CAC curves. The same information could be quantified by converting PPV into an odds ratio, where Odds = PPV / (1 - PPV) = HR / FAR. This is the familiar "diagnosticity ratio." In this case, the diagnosticity ratio would equal .75 / (1 - .75) = = .75 / .25 = 3.0. The diagnosticity ratio is not a useful measure for a system-variable question, but it is a useful measure for an estimator-variable question, especially when it is computed separately for different levels of confidence (e.g., Brewer & Wells, 2006). We will focus mainly on PPV because, for most people, a probability measure is more easily understood than an odds ratio.

The key point is that discriminability and PPV measure completely different things. That fact is most easily illustrated by simply changing base rates, as illustrated in the right panel of Figure 1, while holding the diagnostic performance of the lineup constant. In the right panel, the to-be-diagnosed population (Population 2) consists of 5 Gs (guilty suspects) and 10 Is (innocent suspects), which means that the base rate of guilty suspects in this population is 5 / 15, or .33. The diagnostic performance of Diagnostic Test 2 is identical to that of Diagnostic Test 1 (e.g., they could be the same exact lineup test applied to different populations), so the hit and false alarm rates remain unchanged, and *d'* is still ≈ 1.0 (i.e., discriminability remains unchanged). Because it is insensitive to base rates, discriminability is not a Bayesian measure. However, notice that PPV is now reduced to .60. This means that a suspect ID becomes less trustworthy

even though the identical lineup procedure was used. Because it *is* sensitive to base rates, PPV is a Bayesian measure.

Figure 2 illustrates another concept, which is that PPV can change even when both base rates and discriminability are held constant. The left panel is the same as it was for Figure 1. For the right panel, imagine that everything is the same (same exact test applied to the same exact equal base-rate population) except that now suspect IDs made with low confidence are treated as effective non-IDs. In other words, a higher (more conservative) standard is used such that a low-confidence ID, because of its highly error-prone nature, is counted as a negative test (not as a positive test) for G. In the simplest case, this manipulation would reduce both the hit rate and the false alarm rate (because neither low-confidence hits nor low-confidence false alarms would be counted as positive IDs) but would not change discriminability. Even so, adopting this more conservative standard would increase PPV (and the corresponding diagnosticity ratio), as illustrated in the right panel of Figure 2. Indeed, many studies have now conclusively shown that adopting a more conservative decision rule has the effect of increasing PPV (e.g., Mickes et al., 2012; Mickes et al., 2017).

These illustrations merely underscore the critical point that discriminability and PPV (or the diagnosticity ratio) measure different things that are of interest to different actors in the legal system. Discriminability is the measure of most interest to a policymaker, whose goal is to maximize the hit rate while simultaneously minimizing the false alarm rate. PPV is the measure of most interest to judges and jurors, whose goal is to assess the reliability of a suspect ID that was made with a particular level of confidence (Mickes, 2015).

The Effect of Estimator Variables on the Confidence-Accuracy Relationship

With these considerations in mind, we can return to the main question of interest, which concerns the effect of estimator variables on the reliability of eyewitness identification. Wixted and Wells (2017) recently reviewed the eyewitness identification literature, reanalyzing the data from confidence-accuracy studies in terms of CAC analysis (Mickes, 2015), where the dependent measure – namely, suspect ID accuracy as a function of confidence – is equivalent to plotting PPV computed separately for each level of confidence. It might be better to refer to this relationship as the "confidence-PPV relationship," but we use the standard term "confidence-accuracy relationship" because it is a ubiquitous phrase.

Some of the studies reviewed by Wixted and Wells (2017) manipulated estimator variables. More specifically, those studies manipulated retention interval (short vs. long), exposure duration (short vs. long), attention (full vs. divided), presence vs. absence of a weapon, and match between the race of the witness and the perpetrator (same-race vs. cross-race). The results of those studies are reproduced here in Figure 3 (the ones that included a manipulation of retention interval) and Figure 4 (the ones that manipulated a variety of other estimator variables). It is readily apparent that none of these variables had an appreciable effect on the accuracy (i.e., PPV) of suspect IDs made with high confidence, though they may have had some effect on the accuracy of suspect IDs made with lower levels of confidence. In all cases, regardless of whether the estimator variable was favorable or not, low confidence was associated with relatively low accuracy and high confidence was associated with very high accuracy.

All of the studies considered above were lab studies, but the results they reported are consistent with the results of a recent police department field study that was specifically designed to examine the information value of eyewitness confidence (Wixted, Mickes, Dunn, Clark & W. Wells, 2016). In that study, eyewitness decisions were recorded from six-person photo lineups

administered as part of criminal investigations in the Robbery Division of the Houston Police Department between January 22 and December 5, 2013. This study involved the administration of 348 simultaneous and sequential lineups, the investigators were unaware of the identity of the suspect in each lineup (i.e., double-blind administration was used), and the lineups involved suspects who were unknown to the eyewitnesses prior to the crime. Although measures of estimator variables were not collected as part of this study, a survey of cases from the same division of the Houston Police Department from the previous year reported by W. Wells, Campbell, Li and Swindle (2016) indicated that 61.6% were cross-race cases, 73.5% involved the presence of a weapon, and the average delay between the offense and the identification procedure was over a month (median = 2.5 weeks). Moreover, it seems reasonable to suppose that witness stress was typically high. In other words, the estimator variables were such that one might reasonably assume that the reliability of eyewitness identifications in this study would be very poor. Even so, estimated suspect ID accuracy (estimated PPV) was very similar to what has been observed in lab studies, and estimated high-confidence accuracy was very high. The results are reproduced here in Figure 5.

The key point is that the estimator variables considered above appear to have had virtually no effect on the reliability of IDs made with high confidence (and little to no effect on IDs made with lower levels of confidence). *This is true even though there is no doubt at all that these same variables all had a substantial effect on discriminability*. For example, there is no doubt that forgetting occurs as the retention interval increases. Thus, after a long retention interval, it is more difficult for eyewitnesses to discriminate between innocent and guilty suspects. Nevertheless, the reliability of eyewitness IDs made with high confidence is very high whether the retention interval is short or long. Although counterintuitive, there is no

contradiction between these two observations about the effect of retention interval on the accuracy of eyewitness identifications.

As has been noted before, findings like these are incompatible with the optimality hypothesis, at least to the extent that the hypothesis is assumed to apply to representations of the confidence-accuracy relationship other than the point-biserial correlation coefficient. The point-biserial correlation coefficient statistic has long been known to be inappropriate for assessing the confidence-accuracy relationship because the correlation can be close to zero even when confidence is perfectly predictive of accuracy (as was first shown by Juslin et al., 1996). Thus, a theory that is limited to making predictions about how the size of the correlation coefficient changes as a function of information processing conditions would not be particularly useful. If the optimality hypothesis is assumed to also apply to the confidence-accuracy relationship as depicted in a CAC analysis (i.e., PPV plotted as a function of confidence), then the data shown in Figures 3 and 4 would seem to contradict it.

What theoretical framework makes sense of this unexpected pattern of results – a pattern that, as shown earlier, is completely at odds with the prevailing view of the effect of estimator variables? The main purpose of our article to show how these findings are naturally predicted by a signal-detection-based likelihood ratio theory of recognition memory. Such theories, in one form or another, have long been a cornerstone in the basic recognition memory literature (Glanzer & Adams, 1990; Glanzer, Adams, Iverson & Kim, 1993; McClelland & Chappel, 1998; Osth, Dennis & Heathcote, 2017; Shiffrin & Steyvers, 1997; Stretch & Wixted, 1998; Wixted & Gaitan, 2002). The argument we advance here is that this theoretical framework not only applies to list-memory studies but also to eyewitness identification procedures. We illustrate the likelihood ratio account in relation to empirical data from an estimator variable study reported by

Lindsay, Semmler, Weber, Brewer and Lindsay (2008). They manipulated the distance between a target person and the witness and measured eyewitness identification accuracy using a 6-person simultaneous lineup in which confidence ratings were recorded using a 100-point confidence scale. Next, we briefly describe that study and related work, and then we present our model-based interpretation of the results reported by Lindsay et al. (2008).

Method

Empirical Studies of Effect of Distance on Eyewitness Identification

Very few studies have measured the effect of distance on the accuracy of discriminability and reliability. Using a list-memory paradigm, Lampinen, Erickson, Moore and Hittson (2014) had participants view 8 target individuals from a particular distance and then tested their ability to identify those targets from a list of 16 photographs presented one at a time, with confidence ratings collected using an 8-point scale. Different groups of participants were tested at different distances ranging from 15 ft (~5 m) to 120 ft (~37 m). As might be expected, discriminability (measured by *d'* or ROC analysis) decreased with distance. Although not specifically analyzed, their data appear to indicate that high-confidence accuracy decreased dramatically as discriminability declined with increasing distance (contrary to what has been found for other estimator variables). The generalizability of these findings to memory tested using a lineup with once-tested participants is not clear.

Lindsay et al. (2008)

The only study of distance that tested memory using a lineup, with each participant tested only once, was reported by Lindsay et al. (2008). Original ethics approval was granted by the Social and Behavioural Research Ethics Committee at Flinders University (Approval# 3268). In this study, 11 different research assistants served as the targets (i.e., as the “perpetrators”).

Approximately 1,300 participants (i.e., witnesses) were approached during normal daily activities and asked to observe one of these targets, who was a certain distance away, for about 10 s. The distances varied across participants, but the short distances fell in the range of 4 to 15 m, whereas the long distances fell in the range of 20 to 50 m. After observing the target, the participant was first asked to answer various questions (e.g., questions about how far away the target was and what the target looked like) and was then asked to try to identify the target from a 6-person target-present or target-absent simultaneous photo lineup. The participant-witnesses were randomly assigned to one of three conditions. In the perceptual judgment condition, the participant was still looking at where the target had just been standing while answering questions and taking the simultaneous lineup test. The immediate judgment condition was the same except that the participant turned around and was no longer looking at the spot where the target had just been observed. In the delayed judgment condition, the participants were contacted a day later, at which time they answered the questions and completed the photo lineup test over the internet. After making a lineup decision, all participants provided a confidence rating using a 100-point confidence scale. For our analyses, we collapsed these ratings in low (0-60), medium (70-80) and high (90-100) confidence. Data from studies using a 100-point confidence scale are often collapsed in just this way (e.g., Mickes, 2015) because there would otherwise be too few responses in a given confidence category to meaningfully analyze (especially in the 0-60 range, where relatively few responses tend to be made).

Our main focus will be on the delayed judgment condition because, as noted by Lindsay et al. (2008), witnesses tested using a lineup are not likely to be at the scene of the crime when the test is administered. As they put it, the delayed judgment condition "…arguably most closely approximates the situation for real-life witnesses" (p. 533). In that condition, all responses were

necessarily based on retrieval from long-term memory, as is true of real eyewitnesses who are tested using a photo lineup, and as is also true of most laboratory studies, which typically impose a distractor task between the mock-crime video and the photo lineup test. In the perceptual and immediate judgment conditions, by contrast, participants presumably tried to maintain an active representation of the just-seen target in working memory at all times in order to answer the questions about that person and to then make an identification of the target from the photo lineup.

Results

Figure 6A shows the CAC plots for the short- and long-distance IDs from the delayed judgment condition. As with the other estimator variables considered earlier, the reliability of an eyewitness identification made with a particular level of confidence is apparently unaffected by distance. Moreover, for both distances, high-confidence IDs are highly accurate. Once again, this result is inconsistent with the optimality hypothesis. Figure 6B shows the CAC plots for short- and long-distance IDs from the perceptual and immediate judgment conditions combined. For these conditions, unlike in the delayed judgment condition, and unlike for the other estimator variables shown in Figures 3 and 4, accuracy was reduced under poorer witnessing conditions (long-distance), even when confidence was high. This pattern seems more consistent with the optimality hypothesis in that not only is high-confidence accuracy reduced from 98% correct when distance was short to 90% correct when distance was long, confidence ratings appear to be at least somewhat less predictive of accuracy when distance was long (i.e., the function relating PPV to accuracy is slightly flatter compared to when distance was short).

Our main focus will be on explaining the pattern observed in the delayed judgment condition, which is similar to the pattern that has been observed for a number of other estimator

variables as well (Figures 3 and 4). Does any existing theory make sense of those surprising results? After describing a basic signal detection model and fitting that model to the data from the delayed judgment condition to estimate the effect of distance on *d'* (discriminability), we consider how the PPV results shown in Figure 6A (reliability) correspond to predictions made by a constant likelihood ratio version of that model. The model predicts that PPV will remain essentially unchanged even if an estimator variable like distance has a large effect on *d'* (as would be expected).

A Constant-Likelihood-Ratio Model of Distance

Signal detection theory applied to lineups. Our theoretical account begins with what is arguably the simplest signal detection model of lineup performance, which is illustrated in Figure 7. This basic model is not itself a constant likelihood ratio model but is the foundation of such a model. The signal detection model shown in Figure 7 represents distributions of memory-match signals generated in the minds of once-tested observers presented with target-present or target-absent lineups. Each face in the photo array generates a memory signal of some strength. On average, but not always, the face of the perpetrator will generate a stronger memory signal compared to innocent suspects or fillers. According to this simple model, memory strength values for lures (innocent suspects and fillers) and for targets (guilty suspects) are distributed according to Gaussian distributions with means of μ_{Lure} and μ_{Target}, respectively, and standard deviations of σ_{Lure} and σ_{Target}, respectively. In a fair target-absent lineup, the innocent suspect is, from the witness's point of view, just another filler.[1] A 6-member target-present lineup is conceptualized as 5 random draws from the lure distribution and 1 random draw from the target

[1] Hence, there is only one lure distribution (the use of an unfair lineup in which the innocent suspect stands out from the fillers because of the suspect's resemblance to the perpetrator would require a third distribution with a mean higher than μ_{Lure} and lower than μ_{Target}).

distribution, and a fair 6-member target-absent lineup is conceptualized as 6 random draws from the lure distribution. For simplicity, we consider the equal-variance version of this model (i.e., $\sigma_{Target} = \sigma_{Lure}$), and, as is typically done for equal-variance models, we set $\mu_{Lure} = 0$ and $\sigma_{Target} = \sigma_{Lure} = 1$. Thus, if μ_{Target} is estimated to equal 2.0 when the model is fit to data, it would indicate that the mean of the target distribution is estimated to be 2.0 standard deviations above the mean of the lure distribution. For the equal-variance version of the model, μ_{Target} is the same as *d'*. Thus, for this example, *d'* would be equal to 2.0.

When confidence ratings are supplied by witnesses, they are conceptualized in terms of different decision criteria. Assuming 3 different levels of confidence associated with a positive ID (low, medium or high confidence), there are 3 different confidence criteria. Unlike list-learning studies, confidence ratings are often not taken when the decision is to reject the test item(s), and even when they are, the ratings are not made in relation to a particular face in the lineup (as they are for positive IDs). Thus, we consider confidence for positive IDs here. The parameters *c1* through *c3* in Figure 7 represent the confidence criteria for positive IDs of a suspect or a filler, assuming a 3-point confidence scale. When the model is fit to the data, it not only estimates *d'*, it also estimates the locations of *c1*, *c2*, and *c3*. If, when fit to the data, *c1*, *c2*, and *c3* are estimated to be 1.00, 1.75, and 2.60, it would mean that *c1* is placed 1 standard deviation above the mean of the lure distribution, *c2* is placed 1.75 standard deviations above the mean of the lure distribution, and *c3* is placed 2.60 standard deviations above the mean of the lure distribution. These are the locations of the confidence criteria in Figure 7.

To apply this model to empirical data, a decision rule needs to be specified about when a face should be identified. Using the simplest decision rule, which Clark, Erickson and Breneman (2011) referred to as the Best-Above-Criterion decision rule, an ID is made if the most familiar

person in a lineup (i.e., the "best" person) exceeds *c1*. A different decision rule might be based on the *difference* in the memory strength of the best face and one or more of the other faces in the lineup (referred to as the Best-Next decision rule), but the simplest decision rule makes use of the most familiar face only (i.e., the best face). If the memory strength of the most familiar face in the lineup exceeds *c1* but not *c2*, an ID is made with the lowest level of confidence. If it exceeds *c2* but not *c3*, the ID is made with medium confidence. If the most familiar face in the lineup exceeds *c3*, an ID of that face is made with the highest level of confidence. The model is fit to all of the data from a given condition – that is, it is fit to the frequency counts of guilty suspect IDs and filler IDs made with particular levels of confidence, plus No IDs, from target-present lineups, and to filler IDs made with particular levels of confidence, plus No IDs, from target-absent lineups.

In essence, the likelihood ratio version of the basic model depicted in Figure 7 is a theory about how the confidence criteria shift on the memory-strength axis across conditions that affect *d'*. We fit the basic model to the distance data reported by Lindsay et al. (2008) to estimate *d'* and *c1*, *c2* and *c3* in both the short- and long-distance conditions. After considering the effect of distance on *d'* (which, of course, is expected to be lower in the long-distance condition), we consider whether *c1*, *c2* and *c3* shifted across conditions in the manner predicted by a constant-likelihood ratio account (and, if so, what that outcome predicts about how distance should affect PPV as a function of confidence). How *c1*, *c2* and *c3* shift across conditions is the crux of the issue.

Estimating d' from the empirical distance data. There is no doubt, of course, that as distance increases, discriminability decreases. Indeed, if the distance between the witness and the perpetrator is great enough, discriminability will obviously drop to zero. In this regard, Loftus

and Harley (2005) recount the case of a witness who was 450 ft (~137 m) away from the perpetrator at the time of the crime but identified him at trial nonetheless. Loftus (2010) describes yet another case of a witness who was 271 ft (~83 m) away from the perpetrator at the time of the crime but, again, identified him at trial. Because, as they showed, it is not possible to recognize faces from these distances, it seems likely that these witnesses made their identifications based on something other than a memory match signal (e.g., perhaps due to police pressure). Here, our concern is with identifications that are made based on the strength of a memory-match signal between a face in a photo lineup and the memory of the perpetrator on an initial memory test using a fair lineup.

Table 1 shows the observed frequency counts from the delayed judgment condition of Lindsay et al. (2008), with "don't know" responses included with "No ID" responses. It makes sense to do so because, according to this model, none of the faces in lineups that received a "don't know" response exceeded *c1*. For a given set of parameter values (e.g., μ_{Target} = 2.0 and *c1*, *c2*, and *c3* = 1.0, 1.75 and 2.60, respectively), the model generates a full set of predicted frequency counts similar to the values shown in Table 1. A chi-square goodness-of-fit statistic is then computed comparing the full set of predicted values to the full set of observed values. The 4 parameters are adjusted until the chi-square associated with the observed and predicted values is minimized. The parameter values that minimize chi square are the optimal parameter values.

Table 1

		Short			Long		
		Confidence					
		Low	Med	High	Low	Med	High
TA	Filler IDs	26	17	6	36	11	2
	No IDs		68			58	
TP	Filler IDs	6	4	0	13	1	1
	Suspect IDs	11	9	20	13	14	5
	No IDs		19			38	

When the model in Figure 7 is fit to the data shown in Table 1, it fit the data well, $\chi^2(10) = 12.67$, $p = .242$, and the estimated value of d' was much higher for the short-distance condition (1.73) compared to the long-distance condition (1.20). The difference between these two d' estimates was significant because when the d' values were constrained to be equal, the fit to the data was much worse, $\chi^2(1) = 6.59$, $p = .010$. This expected effect of distance on discriminability is most clearly illustrated by plotting the ROC for each condition. Figure 8 presents the ROC data (correct suspect ID rate vs. false suspect ID rate) computed from the values shown in Table 1. The correct ID rate for each level of confidence is computed by counting the number of correct IDs made from target-present lineups with that level of confidence *or a higher level* of confidence, divided by the total number of target-present lineups. Note that filler IDs from target-present lineups are not included in this calculation, but those responses were taken into account when the model was fit to the data. Because these data were collected from a procedure in which no one was pre-designated as the innocent suspect, the false ID rate for each level of confidence (for plotting the ROC) is computed by counting the number of filler IDs made from target-absent lineups with that level of confidence or higher, divided by the total number of target-absent lineups and then dividing again by the lineup size of 6 (this is the false alarm rate

shown on the x-axis). Dividing by lineup size when there is no designated innocent suspect provides an estimate of the false *suspect* ID rate.

The ROC data clearly show the expected effect of distance on discriminability (higher in the short-distance condition). The smooth curves drawn through the ROC points show the predictions of the best-fitting signal detection model. That is, the upper curve is generated by a signal detection model using the BEST decision rule with d' =1.73, and the lower curve is generated by a signal detection model using the BEST decision rule with d' =1.20.

Estimating c1, c2, and c3 from the empirical distance data. As noted earlier, for judges and jurors, the accuracy measure of interest is reliability (PPV), not d'. In particular, their question concerns PPV associated with the level of confidence expressed by the eyewitness at the time of the initial ID. PPV is a different accuracy measure, and the predictions made by a signal detection model about PPV for each level of confidence is determined not only by d' but also by where *c1, c2,* and *c3* are placed on the memory strength axis.

In the basic memory literature, a great deal of evidence suggests that participants provide confidence ratings in such a way as to maintain constant likelihoods of being correct (e.g., Glanzer, Adams, Iverson & Kim, 1993). The constant likelihood ratio principle is most easily illustrated by considering how *c1* changes as a function of discriminability. Likelihood ratio models assume that a decision criterion is placed on a memory-strength axis based on the odds that a test item with that memory strength was drawn from the target distribution. Graphically, this odds value is given by the height of the target distribution divided by the height of the lure distribution at the point. In the example shown in Figure 7, *c1* is placed at the point on the memory-strength axis where the heights of the target and lure distributions are exactly equal. This means that a test face that generates a memory strength that falls at that point is equally

likely to be a target or a lure (the odds that it is a target or a lure are even). For any memory strength to the right of *c1*, the height of the target distribution is greater than the height of the lure distribution, which means that any face that generates a memory-strength signal that falls above *c1* is more likely to have come from the target distribution than the lure distribution.

A bit to the right of *c1* is a memory strength value where the height of the target distribution is twice the height of the lure distribution. A face with a memory strength that falls at that point is twice as likely to be a target as it is to be a lure. Thus, a witness who is aware of that fact should be more confident that the face is a target (compared to a face that generates a memory strength value that falls where *c1* is placed). Faces that generate even stronger memory-strength values (i.e., even further to the right) are associated with even higher odds ratios that the face is the target. That fact is not visually apparent because it appears that for very strong faces (very far to the right), the target and lure distributions eventually return to being equally high as the heights of both distributions descend towards the *x*-axis. In truth, that is an optical illusion because the ratio continues to increase towards infinity the higher the memory strength of the face (Stretch & Wixted, 1998). Thus, a participant who is aware of that fact should be ever more confident that the face is a target as memory strength increases.

The constant likelihood ratio model makes a clear prediction about how *c1* should change as *d'* decreases under poorer estimator variable conditions (in this case, longer distance). Figure 9 illustrates the prediction. As *d'* decreases, the point on the memory-strength axis where the target and lure distributions intersect (the equal likelihood point) shifts to the left. Thus, to maintain the same 1/1 likelihood ratio associated with the decision criterion, *c1* needs to shift to the left. If *c1* is in fact placed in such a way as to maintain even odds, when the simple signal detection model is fit to the data, the estimate of where *c1* is placed should be lower in the long-distance (low-*d'*)

condition compared to the short-distance (high-*d'*) condition. If that outcome were observed, then, in terms of likelihood ratios, the criterion would not have shifted at all. In both conditions, a positive ID would be made when the odds are greater than even that the face was drawn from the target distribution (i.e., the same decision criterion is used when the criterion is defined as a likelihood ratio). In terms of its placement on the memory-strength axis (which is what a fit of the signal detection model estimates), *c1* will have shifted to the left in the condition with lower discriminability.

The decision criteria for making an ID with higher confidence (*c2* and *c3*) are placed at points on the memory-strength axis associated with likelihood ratios greater than 1/1. For example, *c2* might be placed where the height of the target distribution is 5 times that of the lure distribution, and *c3* might be placed where the height of the target distribution is 10 times that of the lure distribution. How should the estimated locations of *c2* and *c3* on the memory-strength axis change in order to maintain those same likelihood ratios as *d'* decreases under poorer estimator variable conditions? Should they also shift to the left and by the same amount as *c1*? The answer is “no.” As shown by Stretch and Wixted (1998), in order to maintain constant likelihood ratios, the criteria would need to fan out on the memory-strength axis. In fact, the fanning out of the confidence criteria as *d'* decreases is the signature prediction of a constant likelihood ratio model. It is the pattern that would need to be observed for PPV to remain essentially constant for a given level of confidence even though discriminability decreases. In list-memory experiments, the predicted fanning pattern is reliably observed (Stretch & Wixted, 1998). Is the constant-likelihood ratio pattern also observed when the signal detection model is fit to the distance data in Table 1? Indeed it is.

The estimates of *c1*, *c2* and *c3* were 1.40, 1.86, and 2.36, respectively, in the short-distance condition, and 1.39, 2.06, and 2.79, respectively, in the long distance condition. Figure 10 shows a visual representation of the best-fitting signal detection models in the short- and long-distance conditions, and it is apparent that the three confidence criteria fan out as *d'* decreases. If the confidence parameters are constrained to be equal to each other across conditions (i.e., *c1* short = *c1* long, *c2* short = *c2* long, and *c3* short = *c3* long), the model fits worse, though the effect is only marginally significant, $\chi^2(3) = 7.74$, $p = .052$. This result indicates that the criteria likely did shift on the memory-strength axis across conditions in the general manner predicted by a constant likelihood ratio account (i.e., they fanned out as discriminability decreased).

A stronger test of the likelihood ratio account would be to constrain the confidence criteria across the two distance conditions such that they have the same likelihood ratios. In other words, whatever *c1*, *c2* and *c3* are estimated to be in the short-distance condition, *c1*, *c2* and *c3* in the long-distance condition would be constrained to fall at locations on the memory-strength axis that maintain the same corresponding likelihood ratios. Remarkably, imposing this constraint did not worsen the fit to any appreciable degree, $\chi^2(3) = 0.03$, $p = .999$. This result means that the likelihood ratio values associated with *c1*, *c2* and *c3* in the short-distance condition were *virtually identical* to the likelihood ratio values associated with *c1*, *c2* and *c3* in the long-distance condition (i.e., the data are fully consistent with a constant-likelihood ratio model). The likelihood ratios associated with *c1*, *c2* and *c3* for both conditions were estimated to be 2.54, 5.62, and 13.34, respectively. The fact that the data were almost perfectly consistent with a constant likelihood ratio model is presumably coincidental because measurement error

alone would be expected to make an almost perfect outcome unlikely even if the likelihood ratio account is correct. Still, the data clearly support the likelihood ratio account.

The constant likelihood ratios associated with the confidence criteria across conditions predicts that PPV will remain unchanged even as *d'* decreases under poorer estimator variable conditions. Figure 11 shows the predicted CAC plots for the best-fitting signal detection model depicted in Figure 10. Clearly, despite the large and statistically significant effect of distance on discriminability (i.e., despite the significant effect on that measure of accuracy), under a constant-likelihood ratio model, distance should have no effect on PPV (i.e., on this measure of accuracy). Although the best-fitting constant-likelihood ratio model appears to overestimate low-confidence accuracy (estimating it to be about 78% correct, higher than what is seen in the actual data shown in Figure 6), the distance data are generally consistent with this interpretation, and so are the other estimator variable CAC plots shown earlier in Figures 3 and 4. In other words, the observed PPV results follow naturally from a standard constant likelihood ratio account that has long been applied to list-learning studies from the basic recognition memory literature.

Do people really compute likelihood ratios?

An intriguing question – one that lies at the heart of debates about likelihood ratio models in the basic recognition memory literature – concerns how it is that participants manage to maintain constant likelihood ratios for the various confidence criteria across conditions. A computer does it by using the equation for the Gaussian distribution, which is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where *x* represents a particular memory-strength value on the *x*-axis. For the equal-variance model, $\sigma = 1$ for both the target and lure distributions. For the target distribution, $\mu = d'$ and for the lure distribution, $\mu = 0$. Thus, for the target distribution, this equation becomes

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-d')^2/2}$$

and for the lure distribution, it becomes

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-0)^2/2}$$

Consider where on the x-axis *c1* should be placed to achieve a likelihood ratio of *L* (where *L* might equal 5, for example). *c1* should be placed at the value of x that satisfies this equation:

$$L = e^{-(x-d')^2/2} / e^{-(x-0)^2/2}$$

Solving for *x* yields:

$$x = \frac{\ln(L)}{d'} + d'/2$$

This is where on the *x*-axis *c1* should be placed to keep the likelihood ratio at *L*. As *d'* changes, the location on the *x*-axis where *c1* should be placed to maintain the same likelihood ratio also changes. Our signal detection model fits suggest that the confidence criteria do in fact shift on the *x*-axis in accordance with this equation to maintain constant likelihood ratios for *c1*, *c2* and *c3*.

Does that result mean that participants are capable of computing ratios of Gaussian distributions in their heads in order to determine where the confidence criteria should be placed? Skepticism about that possibility has long been a headwind that likelihood ratio models have had to confront. Wixted and Gaitan (2002) argued that what appears to be a remarkable ability to compute Gaussian-based likelihood ratios is perhaps better construed as a reflection of the participant's learning history, something that is often ignored in cognitive models of recognition

memory. According to this view, participants have learned from error-feedback provided by everyday experience that when someone was viewed from a long distance, then a stronger memory-match signal is needed before deciding with high confidence that that a photo matches the previously observed individual (Mickes, Hwe, Wais & Wixted, 2011). Indeed, it is hard to imagine why everyday life experience would not teach this lesson. Other models provide different mechanistic accounts of how constant likelihood ratios are maintained as *d'* changes, and some further assume that the *x*-axis is not best construed in terms of memory strength but is properly construed as a log likelihood ratio variable itself (e.g., Glanzer & Adams, 1990; Osth, Dennis & Heathcote, 2017). Regardless of which view is correct, the fact that recognition memory performance is often accurately predicted by a likelihood ratio model can now be extended to eyewitness identification. Moreover, this account naturally predicts the otherwise surprising finding that estimator variables, while having a large effect on discriminability, have little to no effect on PPV for a given level of confidence.

Although the constant likelihood ratio model provides an adequate account of the CAC data from the delayed judgment condition shown in Figure 6A, it clearly does not provide an adequate account of the CAC data from the perceptual and immediate judgment conditions shown in Figure 6B. For those combined data, the model in Figure 7 does not provide a very good fit in the first place, $\chi^2(10) = 29.9$, $p < .001$ (indicating that the observed data deviate significantly from the data predicted by the best-fitting model). In addition, when a constraint was added such that *c1*, *c2* and *c3* are required to maintain constant likelihood ratios across conditions, the fit was dramatically worse, $\chi^2(3) = 28.9$, $p < .001$. The fit was also dramatically worse when *c1*, *c2* and *c3* were constrained to be equal across conditions or to shift in lockstep across conditions. In short, the basic signal detection model did not characterize these data very

well, and the confidence criteria shifted across conditions in ways that are hard to theoretically conceptualize. The identification decisions in these conditions may differ from what was observed in the delayed judgment condition because participants attempted to actively maintain a representation of the target they had just seen while answering questions and taking the lineup test. Whether or not that explanation is correct, the data are not consistent with the constant likelihood ratio model offered here or with any other simple model of criterion placement that we can identify. Future work will be needed to understand how identification proceeds in these sorts of conditions.

General Discussion

Suboptimal estimator variables have long been thought to compromise the reliability of eyewitness identification, but a considerable body of recent work suggests that this intuitively reasonable assumption is not correct. At best, it is incomplete. As shown in Figures 3 through 6, estimator variables such as retention interval, exposure duration, presence or absence of a weapon, same- vs. cross-race, and full vs. divided attention, have no apparent effect on the accuracy on suspect IDs (defined as PPV) made with high confidence. Often, they have little to no effect on suspect IDs made with lower levels of confidence as well, though an apparent effect on lower levels of confidence shows up fairly often. The data we reanalyze here from Lindsay et al. (2008) indicate that the same may be true for short- vs. long-distance. In the condition of their experiment that most closely approximated the situation for real-life witnesses, confidence was a strong predictor of accuracy whether distance was short or long, and high-confidence accuracy was similarly high in both cases (Figure 6A). However, this was not true in the two less forensically-relevant conditions in which participants were tested after having just observed the

target moments ago. It is not clear why that procedural difference would yield a different pattern of results, but it clearly did (Figure 6B).

As illustrated in Figures 1 and 2, PPV is the information of most interest to judges and jurors (Mickes, 2015). It seems fair to suggest that estimator variables play a less important role than has been assumed for decades. Estimator variables do have an undeniable effect on accuracy measured in terms of discriminability (e.g., *d'*), but that effect is not particularly informative to judges and jurors. Discriminability refers to the ability of eyewitnesses to distinguish between innocent and guilty suspects, and it includes a consideration of failures to correctly identify the guilty suspect. PPV, by contrast, focuses on positive IDs only. In a trial where an eyewitness is testifying, that eyewitness has made a positive ID. The accuracy of that ID is what the court wants to know, and the answer is provided by PPV, not *d'*. In sum, our reanalysis shows that a high confidence identification decision is just as diagnostic of guilt under poor observation conditions as it is under good observation conditions, it is simply the frequency of high confidence identification decisions that change.

On the whole, these results are incompatible with the optimality hypothesis, which holds that IDs made with a particular level of confidence become less reliable when estimator variables are suboptimal. That prediction is clearly wrong for IDs made with high confidence (which are the most important IDs because they often result in witnesses testifying against a suspect), though suspect ID accuracy may be somewhat reduced for IDs made with lower levels of confidence. If so, however, it would mean that confidence becomes *more* predictive of accuracy under worse memory conditions because the slope of the CAC plot would become steeper in the low-*d'* condition (e.g., see the lower right panel of Figure 3). Such a result is the opposite of what the optimality hypothesis predicts.

Here, we advanced the argument that a constant likelihood ratio account predicts that even when an estimator variable has a large effect on discriminability, it will have no effect on PPV measured for different levels of confidence. In the basic memory literature, the results often suggest that participants behave largely in accordance with such a model, though not exactly in accordance with it. For example, in summarizing the results of several list-memory studies. Stretch and Wixted (1998) noted that "In general, although the criteria do fan out in the weak condition as the likelihood ratio model predicts, they do not fan out as much as they should" (p. 1405). Although the distance data analyzed here were almost perfectly consistent with a likelihood ratio account (i.e., the criteria did fan out as much as they should in the weak long-distance condition), it seems likely that participants in general will not be perfectly optimal in this regard. Indeed, as noted above, suboptimal estimator variables do appear to sometimes reduce the accuracy of suspect IDs made with lower levels of confidence (contrary to what an idealized likelihood ratio model would predict). The likelihood ratio account is also consistent with recently published research looking at the impact of age of the witness on identification performance, with Colloff, et al. (2017) showing that older witnesses spread their confidence criteria out to maintain the accuracy of their high confidence identifications. The standard likelihood ratio model from the basic memory literature would seem to offer the only account proposed thus far that can accommodate the surprising fact that estimator variables that have an undeniable effect on discriminability often have a minimal effect on PPV.

References

APA brief for Commonwealth v. Walker (2014). Retrieved from: http://www.apa.org/about/offices/ogc/amicus/walker.aspx

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relation in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12,* 11-30. doi:10.1037/1076-898X.12.1.11

Carlson, C. A. & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC Analysis. *Journal of Applied Research in Memory and Cognition, 3,* 45–53. doi:10.1016/j.jarmac.2014.03.004

Carlson, C. A., Dias, J. L., Weatherford, D. R. & Carlson, M. A. (2016). An investigation of the weapon focus effect and the confidence-accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition,* doi:10.1016/j.jarmac.2016.04.001

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35,* 364-380. doi: 10.1007/s10979-010-9245-1

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging*, *32*(3), 243. http://dx.doi.org/10.1037/pag0000168

Cutler, B. L. (2006). A sample of witness, crime, and perpetrator characteristics affecting eyewitness identification accuracy. *Cardozo Public Law, Policy & Ethics Journal*, *4*, 327-340.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, *4*, 243–260. doi: 0147-7307/80/1200-0243$03.00/0

Deffenbacher, K. A. (2008). Estimating the impact of estimator variables on eyewitness identification: A

fruitful marriage of practical problem solving and psychological theorizing? *Applied Cognitive Psychology, 22,* 815–826.

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30,* 113-125. doi:10.1002/acp.3178

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567.

Innocence Project (2017). Understand the causes: the causes of wrongful conviction. New York: Innocence Project. http://www.innocenceproject.org/causes-wrongful-conviction. Accessed July 14, 2017.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1304-1316.

Lampinen, J.M., Erickson, W.B., Moore, K.N., & Hittson, A. (2014). Effects of distance on face recognition: Implications for eyewitness identification. *Psychonomic Bulletin and Review, 21,* 1489-1494. doi: 10.3758/s13423-014-0641-2

Lindsay, R.C.L., Semmler, C., Weber, N., Brewer, N. & Lindsay, M. (2008). How variations in distance affect eyewitness reports and identification accuracy. *Law and Human Behavior, 32,* 526-535. doi: 10.1007/s10979-008-9128-x

Loftus, G.R. (2010). What can a perception-memory expert tell a jury? *Psychonomic Bulletin & Review, 17*, 143-148. doi:10.3758/PBR.17.2.143

Loftus, G.R. & Harley, E.M. (2005). Why is it easier to recognize someone close than far away? *Psychonomic Bulletin & Review, 12*, 43-65.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724–760.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*, 93-102.

Mickes, L., Seale-Carlisle, T.M., Wetmore, S.A., Gronlund, S.D., Clark, S.E., Carlson, C.A., Goodsell, C.A., Weatherford, D. & Wixted, J.T. (2017). ROCs in eyewitness identification: Instructions versus confidence ratings. *Applied Cognitive Psychology, 31*, 467-477. doi: 10.1002/acp.3344

Mickes, L., Hwe, V., Wais, P. E. & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140*, 239-257.

New Jersey Model Criminal Jury Charges (2012). Retrieved from http://www.judiciary.state.nj.us/pressrel/2012/jury_instruction.pdf

Osth, A.F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology, 92*, 101-126. https://doi.org/10.1016/j.cogpsych.2016.11.007

Palmer, M., Brewer, N., Weber, N. & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55-71.

Papailiou, A. P., Yokum D.V. & Robertson C.T. (2015). The novel New Jersey eyewitness instruction induces skepticism but not sensitivity. *PLoS ONE* 10(12): e0142695. https://doi.org/10.1371/journal.pone.0142695

Read, J. D., Lindsay, D. S., & Nichols, T. (1998). The relation between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? In C. P. Thomson, D. Bruce, J. D. Read, D. Hermann, D. Payne, & M. P. Toglia (Eds.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107-130). Mahwah, NJ: Erlbaum.

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34,* 337–347. doi: 10.1007/s10979-009-9192-x

Schum, D. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance, 27*, 153-196. https://doi.org/10.1016/0030-5073(81)90045-3

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145-166.

State v. Almaraz, 301 P. 3d 242 – 2013.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1397–1410.

Wells, G. L. (1978). Applied eyewitness testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*, 1546-1557. doi:10.1037/0022-3514.36.12.1546

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior, 39*, 99-

122. doi:10.1037/lhb0000125

Wells, W., Campbell, B., Li, Y. & Swindle, S. (2016). The characteristics and results of eyewitness identification procedures conducted during robbery investigations in Houston, TX. *Policing: An International Journal of Police Strategies & Management, 39*, 601 - 619

Wixted, J. T. & Gaitan, S. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior, 30*, 289-305.

Wixted, J. T., Mickes, L., Clark, S., Gronlund, S. & Roediger, H. (2015). Confidence judgments are useful in eyewitness identifications: A new perspective. *American Psychologist, 70*, 515-526. http://dx.doi.org/10.1037/a0039510

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E. & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304-309. doi: 10.1073/pnas.1516814112

Wixted, J. T. & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10-65. https://doi.org/10.1177/1529100616686966

Figure Captions

Figure 1. A schematic illustration of how hit and false alarm rates differ from PPV. Only PPV is affected by the base rate of G (where G = guilty suspects or people with diabetes and I = innocent suspects or people without diabetes).

Figure 2. A schematic illustration of how hit rates, false alarm rates, and PPV are affected by changing the criterion for counting a test result as "positive."

Figure 3. Suspect ID accuracy in terms of percent correct (% correct) as a function of confidence from four studies that manipulated retention interval. The figures are reproduced from a recent review by Wixted and Wells (2017).

Figure 4. Suspect ID accuracy in terms of percent correct (% correct) as a function of confidence from four studies that manipulated a variety of estimator variables. The figures are reproduced from a recent review by Wixted and Wells (2017).

Figure 5. Estimated suspect ID accuracy as a function of confidence from a recent police department field study reported by Wixted et al. (2016).

Figure 6. **A**. Suspect ID accuracy as a function of confidence when the distance between the witness and perpetrator was short vs. when it was long in the delayed condition of Lindsay et al. (2008). **B**. Corresponding data from the perceptual and immediate conditions combined.

Figure 7. Illustration of the basic signal detection model of lineups. According to this model, an ID is made if the face with the strongest memory strength in the lineup exceeds the decision criterion (*c1*). A low-confidence ID is made if that memory strength falls between *c1* and *c2*; a

medium-confidence ID is made if it falls between *c2* and *c3*; and a high-confidence ID is made if it falls above *c3*.

Figure 8. ROC data for short and long distances in the delayed condition of Lindsay et al. (2008). The smooth curves represent the predictions of the best-fitting signal detection model illustrated in Figure 7.

Figure 9. An illustration of how the decision criterion (*c1*) would shift as discriminability declines in order to maintain a constant likelihood ratio of 1/1.

Figure 10. An illustration of the best-fitting signal detection models when distance was short vs. long in the delayed condition of Lindsay et al. (2008).

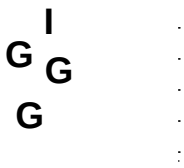
Figure 11. Predicted suspect ID accuracy from the best-fitting models shown in Figure 10.

Figure 1.

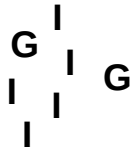
Diagnostic Test 1

Population 1
(5 *G*s and 5 *I*s)

Positive Test for G



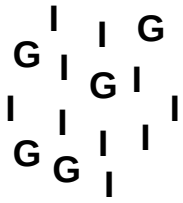
Negative Test for G



Hit rate = 3 *G*s / 5 *G*s = .60
FA rate = 1 *I* / 5 *I*s = .20
PPV = 3 *G*s / 4 positive test results = .75

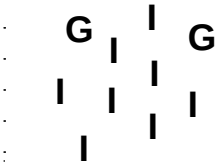
Diagnostic Test 2

Population 2
(5 *G*s and 10 *I*s)



Positive Test for G

Negative Test for G



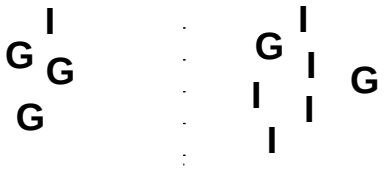
Hit rate = 3 *G*s / 5 Gs = .60
FA rate = 2 *I*s / 10 *I*s = .20
PPV = 3 *G*s / 5 positive test results = .60

Figure 2.

Diagnostic Test 1

Population 1
(5 *G*s and 5 *I*s)

Positive Test for G | Negative Test for G

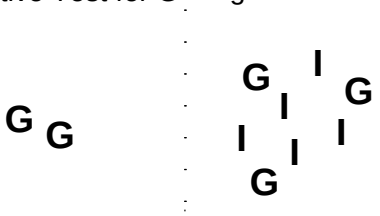


Hit rate = 3 *G*s / 5 *G*s = .60
FA rate = 1 *I* / 5 *I*s = .20
PPV = 3 *G*s / 4 positive test results = .75

Diagnostic Test 2

Population 2
(5 *G*s and 5 *I*s)

Positive Test for G | Negative Test for G



Hit rate = 2 *G*s / 5 *G*s = .40
FA rate = 0 *I*s / 5 Is = 0
PPV = 2 Gs / 2 positive test results = 1.0

Figure 3.

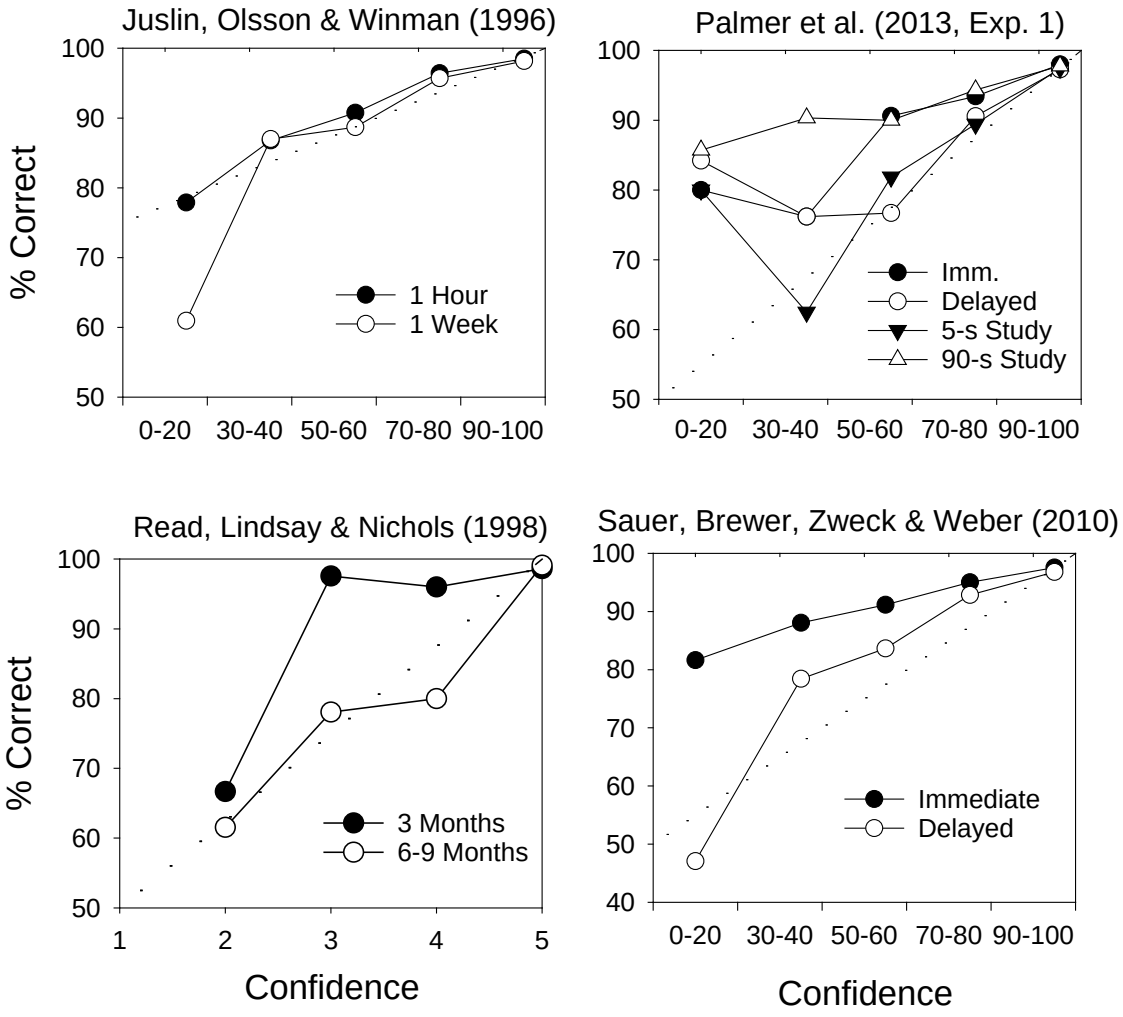

Juslin, Olsson & Winman (1996)
% Correct
100
90
80
70
60
50
0-20
30-40
50-60
70-80
90-100
1 Hour
1 Week
Palmer et al. (2013, Exp. 1)
Imm.
Delayed
5-s Study
90-s Study
Read, Lindsay & Nichols (1998)
3 Months
6-9 Months
1
2
3
4
5
Confidence
Sauer, Brewer, Zweck & Weber (2010)
40
Immediate
Delayed

Figure 4.

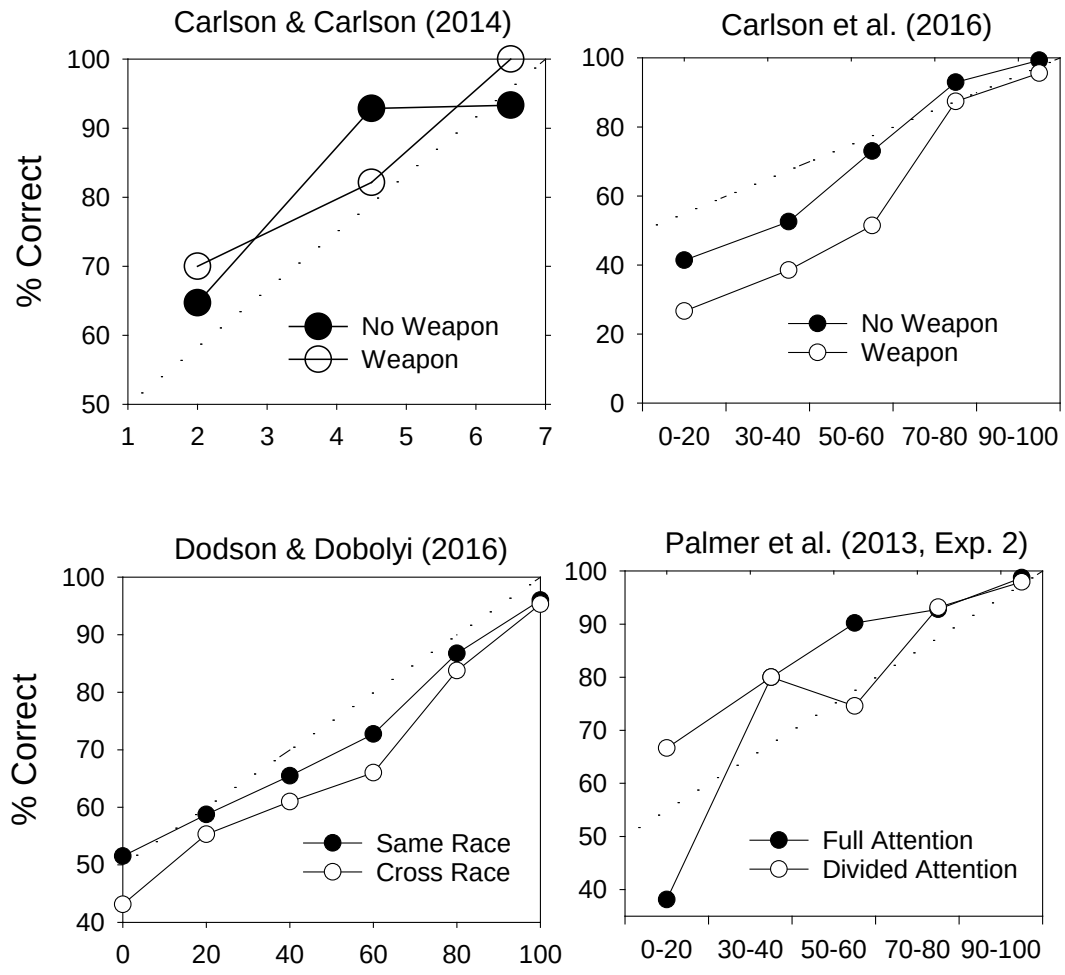

Carlson & Carlson (2014)
% Correct
100
90
80
70
60
50
1
2
3
4
5
6
7
No Weapon
Weapon
Carlson et al. (2016)
100
80
60
40
20
0
0-20
30-40
50-60
70-80
90-100
No Weapon
Weapon
Dodson & Dobolyi (2016)
% Correct
100
90
80
70
60
50
40
0
20
40
60
80
100
Same Race
Cross Race
Palmer et al. (2013, Exp. 2)
100
90
80
70
60
50
40
0-20
30-40
50-60
70-80
90-100
Full Attention
Divided Attention

Figure 5.

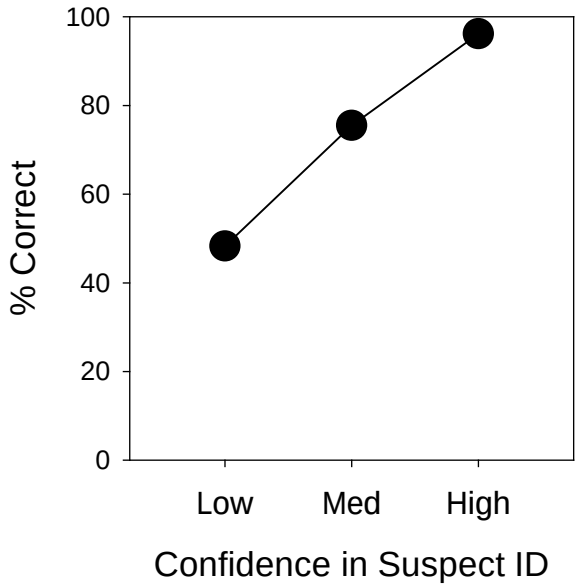

100
80
60
40
20
0
% Correct
Low
Med
High
Confidence in Suspect ID

Figure 6.

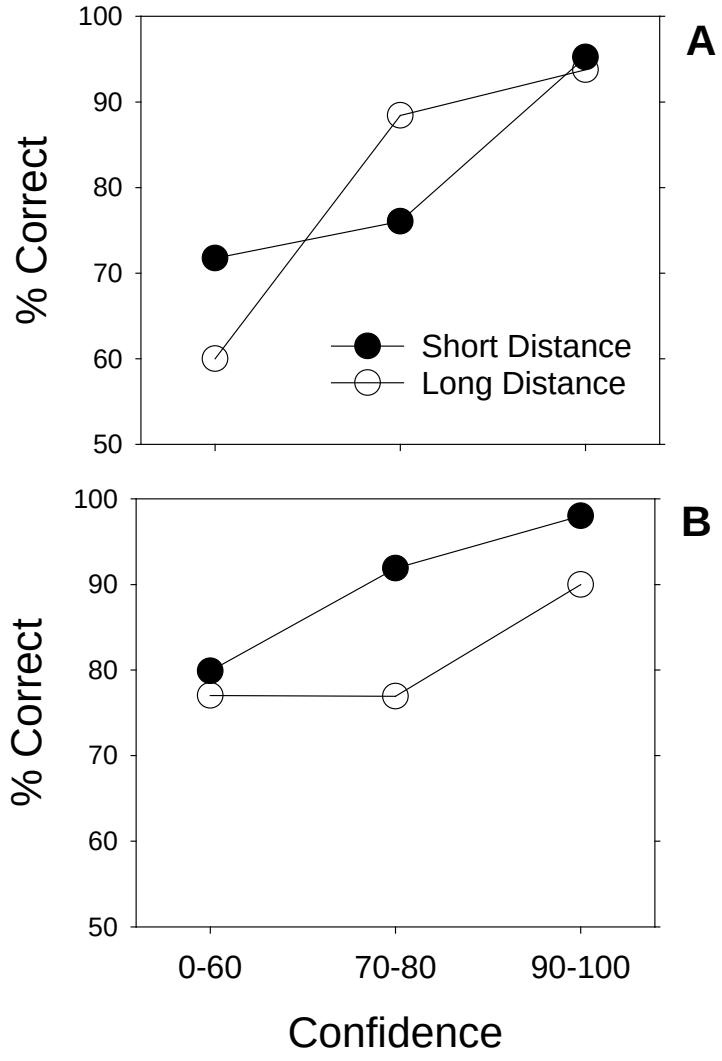

A
B
100
90
80
70
60
50
% Correct
Short Distance
Long Distance
0-60
70-80
90-100
Confidence

Figure 7.

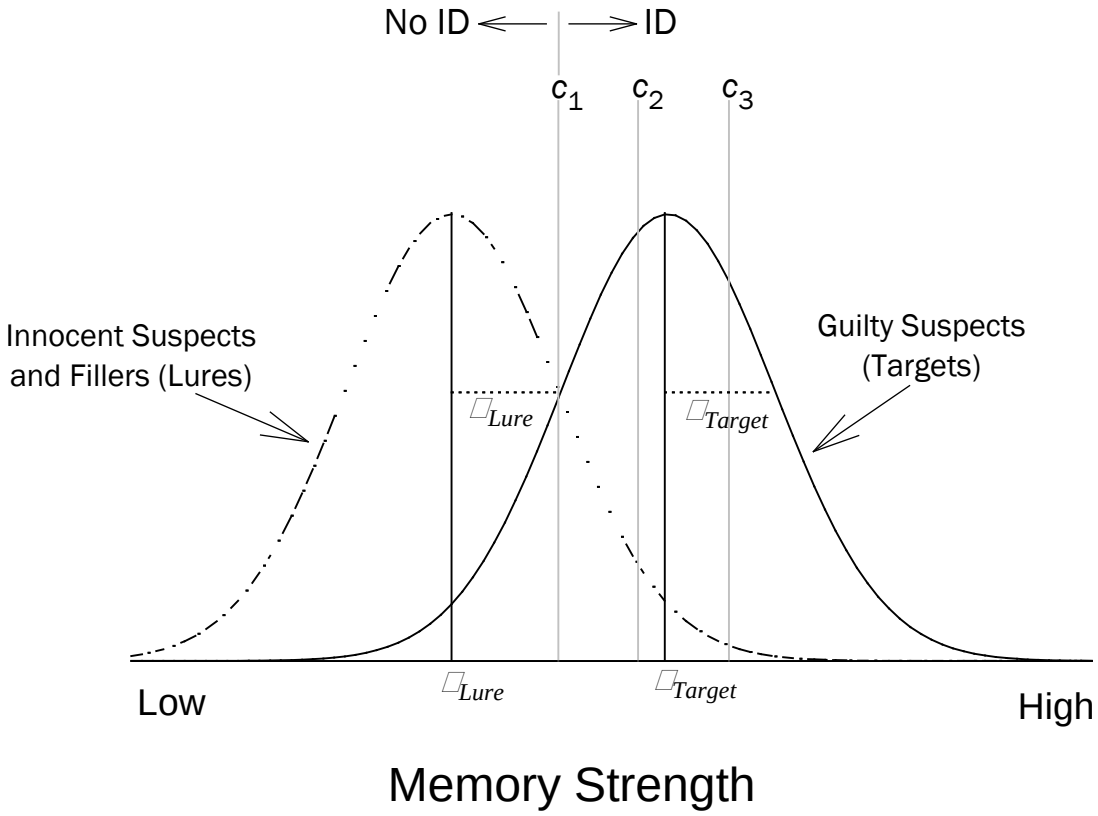

No ID
ID
c_1
c_2
c_3
Innocent Suspects and Fillers (Lures)
Guilty Suspects (Targets)
μ_{Lure}
μ_{Target}
μ_{Lure}
μ_{Target}
Low
High
Memory Strength

Figure 8.

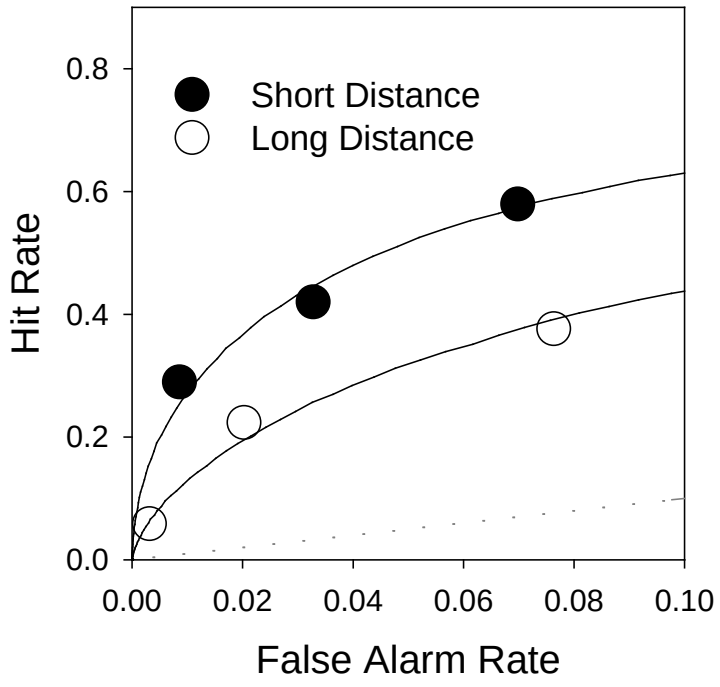

Short Distance
Long Distance
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
0.00
0.02
0.04
0.06
0.08
0.10

Figure 9.

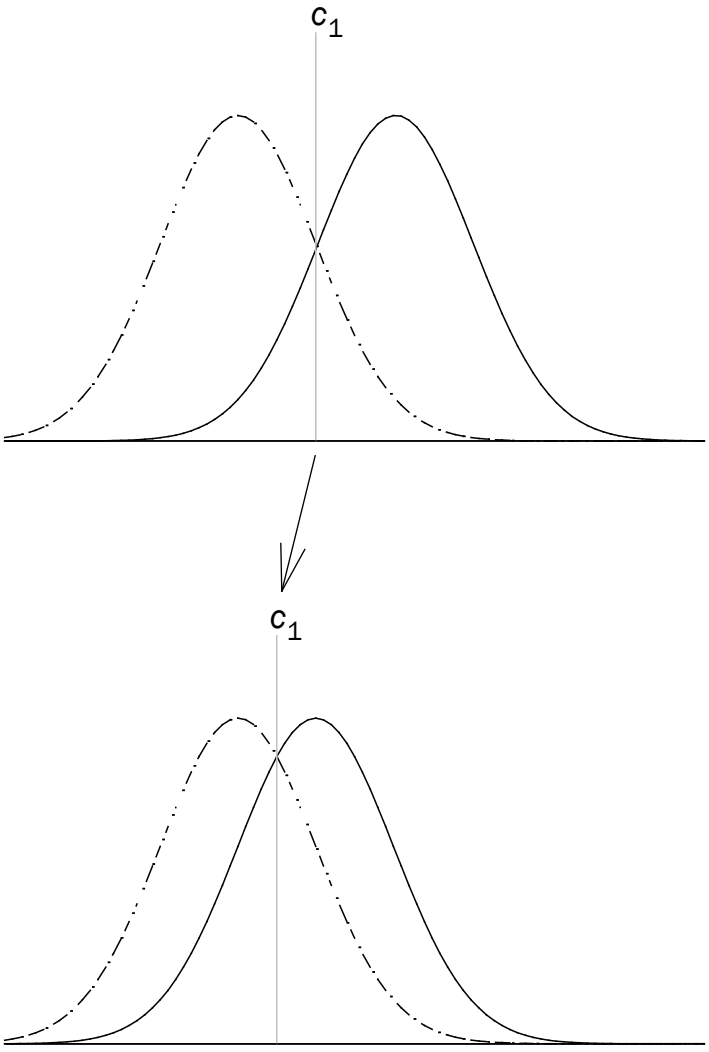
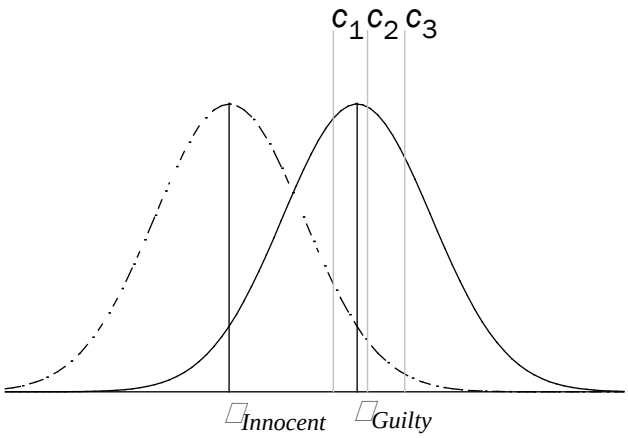
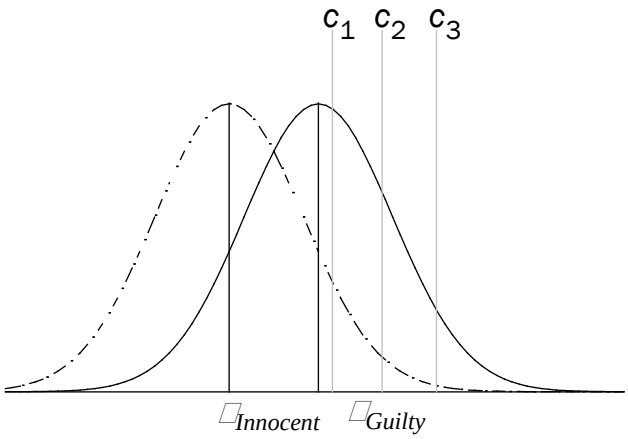


Figure 10.

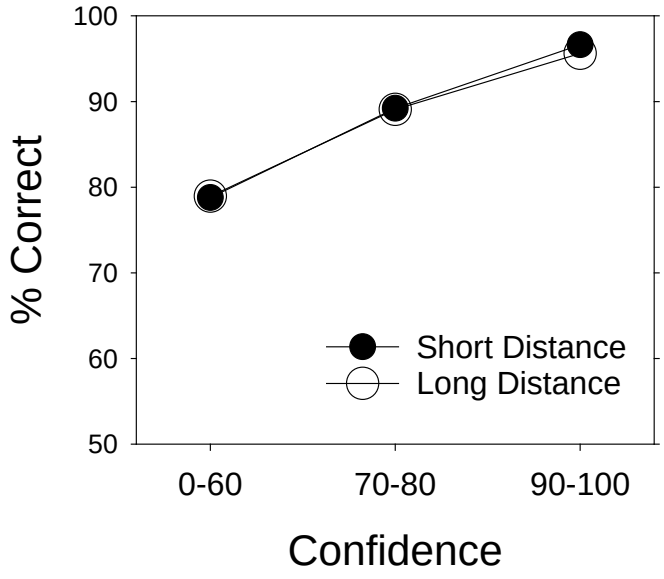


Short Distance
(*d*′ = 1.72)



Long Distance
(*d*′ = 1.20)

Figure 11.


100
90
80
70
60
50
% Correct
Short Distance
Long Distance
0-60
70-80
90-100
Confidence