VOLUME SIXTY THREE

THE PSYCHOLOGY OF LEARNING AND MOTIVATION

Edited by

BRIAN H. ROSS
Beckman Institute and Department of Psychology
University of Illinois, Urbana, Illinois

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

CHAPTER ONE

Conducting an Eyewitness Lineup: How the Research Got It Wrong

Scott D. Gronlund*,[1], Laura Mickes[§], John T. Wixted[¶] and Steven E. Clark[||]

*Department of Psychology, University of Oklahoma, Norman, OK, USA
[§]Department of Psychology, Royal Holloway, University of London, Surrey, England
[¶]Department of Psychology, University of California, San Diego, CA, USA
[||]Department of Psychology, University of California, Riverside, CA, USA
[1]Corresponding author: E-mail: sgronlund@ou.edu

Contents

Psychology of Learning and Motivation, Volume 63
ISSN 0079-7421
http://dx.doi.org/10.1016/bs.plm.2015.03.003

Abstract

A set of reforms proposed in 1999 directed the police how to conduct an eyewitness lineup. The promise of these system variable reforms was that they would enhance eyewitness accuracy. However, the promising initial evidence in support of this claim failed to materialize; at best, these reforms make an eyewitness more conservative. The chapter begins by reviewing the initial evidence supporting the move to description-matched filler selection, unbiased instructions, sequential presentation, and the discounting of confidence judgments. We next describe four reasons why the field reached incorrect conclusions regarding these reforms. These include a failure to appreciate the distinction between discriminability and response bias, a reliance on summary measures of performance that conflate discriminability and response bias or mask the relationship between confidence and accuracy, and the distorting role of relative judgment theory. The reforms are then reevaluated in light of these factors and recent empirical data. We conclude by calling for a theory-driven approach to developing and evaluating the next generation of system variable reforms.

1. INTRODUCTION

In October 1999, the U.S. Department of Justice released a document entitled *Eyewitness Evidence: A Guide for Law Enforcement* (Technical Working Group for Eyewitness Evidence, 1999), which proposed a set of guidelines for collecting and preserving eyewitness evidence (Wells et al., 2000). The guidelines proposed a set of reforms that were expected to enhance the accuracy of eyewitness evidence. The establishment of these guidelines was a noteworthy achievement for psychology, and was heralded as a "successful application of eyewitness research," "from the lab to the police station." Yet, as we shall see, the field got some of these reforms wrong. The goal of this chapter is to examine how that happened.

Intuitively, there would seem to be few kinds of evidence more compelling than an eyewitness confidently identifying the defendant in a court of law. From a strictly legal perspective, eyewitness identification (ID) is direct evidence of the defendant's guilt. Its compelling nature is not surprising if you strongly or mostly agree that memory works like a video recorder, as did 63% of Simons and Chabris' (2011) representative sample of U.S. adults. Of course, the veracity of that claim has been challenged by countless experiments (for reviews see Loftus, 1979, 2003; Roediger, 1996; Roediger & McDermott, 2000; Schacter, 1999) and, in a different way, by the over 1400 exonerations reported by the National Registry of Exonerations

(eyewitness misidentification played a role in 36% of these false convictions) (www.law.umich.edu/special/exoneration/).

There are a number of factors that adversely affect the accuracy of eyewitness ID of strangers and that can help one understand how it is that honest, well-meaning eyewitnesses can make such consequential errors. These include general factors that characterize normal memory functioning, like its constructive nature (Schacter, Norman, & Koutstaal, 1998) and poor source monitoring (Johnson, Hashtroudi, & Lindsay, 1993). But it also includes factors more germane to eyewitness ID, like limitations in the opportunity to observe (Memon, Hope, & Bull, 2003), the adverse effects of stress on attention and memory (Morgan et al., 2004), and the difficulty of cross-racial IDs (Meissner & Brigham, 2001). Wells (1978) referred to factors like these as estimator variables, because researchers can only *estimate* the impact of these variables on the performance of eyewitnesses. There is little the criminal justice system can do to counteract the adverse impact of these factors. Wells contrasted estimator variables with system variables, which are variables that are under the control of the criminal justice system. System variable research can be divided into two categories. One category focuses on the interviewing of potential eyewitnesses (for example, by using the Cognitive Interview, e.g., Fisher & Geiselman, 1992). The other category focuses on ID evidence and how it should be collected. The collection of ID evidence is the focus of this chapter, particularly the role played by the lineup procedure. The aforementioned guidelines pronounced a series of reforms for how to collect ID evidence using lineups that were supposed to enhance the accuracy of that evidence.

The chapter is divided into four main parts. Section 2 reviews the evidence for these reforms at the turn of the twenty-first century—when the recommendations were being made and adopted (Farmer, Attorney General, New Jersey, 2001). We briefly review the empirical evidence supporting the move to description-matched filler selection, unbiased instructions, sequential presentation, discounting confidence judgments, and double-blind lineup administration. Section 3 lays out four reasons why the field reached incorrect conclusions about several of these reforms. These include a failure to appreciate the distinction between discriminability and response bias; a reliance on summary measures of performance that conflate discriminability and response bias; the distorting role of theory; and a resolute (even myopic) focus on preventing the conviction of the innocent. Section 4 reexamines the reforms in light of the factors detailed in Section 3 and recent empirical data. Section 5 lays out the direction forward,

describing a more theory-driven approach to developing and evaluating the next generation of system variable reforms.

2. EYEWITNESS REFORMS

The guidelines focused on many different aspects regarding how a lineup should be conducted, from its construction to the response made by the eyewitness. One reform recommends that a lineup should include only one suspect (Wells & Turtle, 1986). That means that the remaining members of the lineup should consist of known-innocent individuals called fillers. The rationale for the inclusion of fillers is to ensure that the lineup is not biased against a possibly innocent suspect. One factor to consider is how closely the fillers should resemble the perpetrator (Luus & Wells, 1991). To achieve the appropriate level of similarity, another recommendation requires that the fillers should match the description of the perpetrator (as reported by the eyewitness prior to viewing the lineup). Description-matched fillers—that is, fillers chosen based on verbal descriptors—were argued to be superior to fillers chosen based on their visual resemblance to the suspect (Luus & Wells, 1991; Wells, Rydell, & Seelau, 1993). Next, prior to viewing the lineup, an eyewitness should receive unbiased instructions that the perpetrator may or may not be present (Malpass & Devine, 1981). Another suggestion involved how the lineup members should be presented to the eyewitness. The sequential presentation method presented lineup members one at a time, requiring a decision regarding whether #1 is the perpetrator before proceeding to #2, and so on (Lindsay & Wells, 1985; for a review see Gronlund, Andersen, & Perry, 2013). Once an eyewitness rejects a lineup member and moves on to the next option, a previously rejected option cannot be chosen. Also, as originally conceived, the eyewitness would not know how many lineup members were to be presented. Finally, because the confidence that an eyewitness expresses is malleable (Wells & Bradfield, 1998), confidence was not deemed a reliable indicator of accuracy; only a binary ID or rejection decision was forthcoming from a lineup. Another recommendation, not included in the original guidelines, has since become commonplace. This involves conducting double-blind lineups (Wells et al., 1998). If the lineup administrator does not know who the suspect is, the administrator cannot provide any explicit or implicit guidance regarding selecting that suspect. Table 1 summarizes these reforms; the numeric entries refer to the subsections that follow.

Table 1 Eyewitness reforms from Wells et al. (2000)

Proposed reform	Description
One suspect per lineup	Each lineup contains only one suspect and the remainder are known-innocent fillers
2.1 Lineup fillers: filler similarity	Fillers similar enough to the suspect to ensure that the lineup is not biased against a possibly innocent suspect
2.1 Lineup fillers: filler selection	Select fillers based on description of the perpetrator rather than visual resemblance to the suspect
2.2 Unbiased instructions	Instruct eyewitness that the perpetrator may or may not be present
2.3 Sequential presentation	Present lineup members to the eyewitness one at a time as opposed to all at once
2.4 Proper consideration of confidence	Eyewitness confidence can inflate due to confirming feedback
2.5 Double-blind lineup administration	Neither the lineup administrator nor the eyewitness knows who the suspect is

Eyewitness researchers generally rallied behind the merit of these suggested reforms. Kassin Tubb, Hosch, and Memon (2001) surveyed 64 experts regarding the "general acceptance" of some 30 eyewitness phenomena. Several of these phenomena are related to the aforementioned reforms, including unbiased lineup instructions, lineup fairness and the selection of fillers by matching to the description, sequential lineup presentation, and the poor confidence—accuracy relationship. From 70% to 98% of the experts responded that these phenomena were reliable. For example, "The more members of a lineup resemble the suspect, the higher is the likelihood that identification of the suspect is accurate"; "The more that members of a lineup resemble a witness's description of the culprit, the more accurate an identification of the suspect is likely to be"; "Witnesses are more likely to misidentify someone by making a relative judgment when presented with a simultaneous (as opposed to a sequential) lineup"; "An eyewitness's confidence is not a good predictor of his or her identification accuracy" (Kassin et al., 2001, p. 408).

We will briefly review the rationale and the relevant data that supported these reforms (for more details see Clark, 2012; Clark, Moreland, & Gronlund, 2014; Gronlund, Goodsell, & Andersen, 2012). But before doing so, some brief terminology is necessary. In the laboratory, two types of lineup trials are necessary to simulate situations in which the police have

placed a guilty or an innocent suspect into a lineup. A target-present lineup contains the actual perpetrator (a guilty suspect). In the lab, a target-absent lineup is constructed by replacing the guilty suspect with a designated innocent suspect. If an eyewitness selects the guilty suspect from a target-present lineup, it is a correct ID. An eyewitness makes a false ID when he or she selects the innocent suspect from a target-absent lineup. An eyewitness also can reject the lineup, indicating that the guilty suspect is not present. Of course, this is the correct decision if the lineup is target-absent. Finally, an eyewitness can select a filler. In contrast to false IDs of innocent suspects, filler IDs are not dangerous errors because the police know these individuals to be innocent.

2.1 Proper Choice of Lineup Fillers

There are two factors to consider regarding choosing fillers for a lineup. Filler similarity encompasses how similar the fillers should be to the suspect. Once the appropriate degree of similarity is determined, filler selection comprises how to choose those fillers. Regarding filler similarity, Lindsay and Wells (1980) varied whether or not the fillers matched a perpetrator's description. They found that the false ID rate was much lower when the fillers matched the description. The correct ID rate also dropped, but not significantly. Therefore, according to this reform, fair lineups (fillers match the description) are better than biased lineups (the fillers do not match the description).

If fair lineups are better, how does one go about selecting those fillers? Two methods were compared. The suspect-matched approach involves selecting fillers who visually resemble a suspect; the description-matched approach requires selecting fillers who match the perpetrator's verbal description. Wells et al. (1993) compared these two methods of filler selection and found no significant difference in false ID rates, but description-matched selection resulted in a greater correct ID rate. Lindsay, Martin, and Webber (1994) found similar results.

Navon (1992) and Tunnicliff and Clark (2000) also noted that suspect-matched filler selection could result in an innocent suspect being more similar to the perpetrator than any of the fillers. Navon called this the backfire effect, which Tunnicliff and Clark describe as follows: An innocent person becomes a suspect because the police make a judgment that he matches the description of the perpetrator, but the fillers are chosen because they are judged to match the innocent suspect, not because they are judged to match the perpetrator's description. Consequently, the innocent suspect is more

likely to be identified because he or she is once removed from the perpetrator (matches the description), but the suspect-matched fillers are twice removed (they match the person who matches the description). Based on the aforementioned data, and this potential problem, the guidelines declared description-matched filler selection superior.

2.2 Unbiased Instructions

Malpass and Devine (1981) compared two sets of instructions. Biased instructions led participants to believe that the perpetrator was in the lineup, and the accompanying response sheet did not include a perpetrator-not-present option. In contrast, participants receiving unbiased instructions were told that the perpetrator may or may not be present, and their response sheets included an explicit perpetrator-not-present option. Malpass and Devine found that biased instructions resulted in more choosing from the target-absent lineups. Other research followed that showed that biased instructions resulted in increased choosing of the innocent suspect from target-absent lineups without reducing the rate at which correct IDs were made from target-present lineups (e.g., Cutler, Penrod, & Martens, 1987). A meta-analysis by Steblay (1997) concluded in favor of unbiased instructions.

2.3 Sequential Presentation

Lindsay and Wells (1985) were the first to compare simultaneous to sequential lineup presentation. They found that sequential lineups resulted in a small, nonsignificant decrease to the correct ID rate (from 0.58 to 0.50), but a large decrease in the false ID rate (from 0.43 to 0.17). Two experiments by Lindsay and colleagues (Lindsay, Lea, & Fulford, 1991; Lindsay, Lea, Nosworthy, et al., 1991) also found large advantages for sequential lineup presentation. A meta-analysis by Steblay, Dysart, Fulero, and Lindsay (2001) appeared to confirm the existence of the sequential superiority effect.

2.4 Proper Consideration of Confidence

Wells and Bradfield (1998) showed that confirming a participant's choice from a lineup led to an inflation of confidence in that decision, and an enhancement of various other aspects of memory for the perpetrator (e.g., estimating a longer and better view of the perpetrator, more attention was paid to the perpetrator). Therefore, it was important for law enforcement to get a confidence estimate before eyewitnesses received any

feedback regarding their choice. But that confidence estimate, even if uncontaminated by feedback, played a limited role in the reforms. This limited role stood in contrast to the important role played by confidence as deemed by the U.S. Supreme Court (Biggers, 1972). Confidence is one of the five factors used by the courts to establish the reliability of an eyewitness.

2.5 Double-Blind Lineup Administration

A strong research tradition from psychology and medicine supports the importance of double-blind testing to control biases and expectations (e.g., Rosenthal, 1976). Regarding lineups, the rationale for double-blind lineup administration is to ensure that a lineup administrator can provide no explicit or implicit guidance regarding who the suspect is. Phillips McAuliff, Kovera, and Cutler (1999) compared blind and nonblind lineup administration. They relied on only target-absent lineups, and found that blind administration reduced false IDs when the lineups were conducted sequentially, but not simultaneously. The lack of empirical evidence at the time the reforms were proposed likely explains why double-blind administration was not among the original reforms. There has been some research since. Greathouse and Kovera (2009) found that the ratio of guilty to innocent suspects identified was greater for blind lineup administrators. However, Clark, Marshall, and Rosenthal (2009) showed that blind testing would not solve all the problems of administrator influence. In sum, there remains relatively little evidence evaluating the merits of double-blind lineup administration. Consequently, its status as a reform has more to do with the historical importance of blind testing in other fields than the existence of a definitive empirical base involving lineup testing.

The story of the eyewitness reforms appeared to be complete at the dawn of the twenty-first century. Yes, honest well-meaning eyewitnesses could make mistakes, but the adoption of these reforms would reduce the number of those mistakes and thereby enhance the accuracy of eyewitness evidence. And nearly everyone believed this, from experts in the field (e.g., Kassin et al., 2001), to the criminal justice system (e.g., The Justice Project, 2007; the Innocence Project), textbook writers (e.g., Goldstein, 2008; Robinson-Riegler & Robinson-Riegler, 2004), lay people (see Schmechel, O'Toole, Easterly, & Loftus, 2006; Simons & Chabris, 2011), and the media (e.g., Ludlum's (2005) novel, *The Ambler Warning*; *Law and Order: SVU* (McCreary, Wolf, & Forney, 2009)). An important standard of proof, a meta-analysis, had been completed for several of the reforms, confirming the conclusions. However, the narrative surrounding

these eyewitness reforms, and indeed eyewitness memory in general, has shifted in important ways in the last decade.

3. IMPACT OF THE REFORMS MISCONSTRUED

Why did support coalesce around the aforementioned set of reforms? Clark et al. (2014) addressed this question at some length, and the analysis presented here, built around four fundamental ideas, is similar to that articulated by Clark et al. The first idea is that the field focused almost exclusively on protecting the innocent (the benefit of the reforms), and not the accompanying costs (reduced correct IDs of guilty suspects). The second involves the distinction between response bias (the willingness to make a selection from a lineup) and discriminability (the ability to discriminate guilty from innocent suspects). The third idea highlights the role played by the reliance on performance measures that (1) conflated response bias and discriminability, or (2) masked the relationship between confidence and accuracy. The final idea implicates the role played by theory in the development of a research area, in this case relative judgment theory (Wells, 1984): The rationale for the enhanced accuracy of many of the reforms was that the reforms reduced the likelihood that an eyewitness relied on relative judgments.

3.1 Focus on Benefits, Discount Costs

Eyewitness researchers generally have focused on the benefits of the reforms, and disregarded the costs. That is, they have emphasized the reduction in the false IDs of innocent suspects, while downplaying the reduction in correct IDs of guilty suspects (see Clark, 2012). Due to the failure to appreciate the difference between discriminability and response bias, and a reliance on measures that conflated these factors (see next two subsections), more *conservative* (protecting the innocent) became synonymous with *better*. This focus on protecting the innocent, coupled with the fact that the reforms generally induce fewer false IDs, fed the momentum of these reforms across the United States "like a runaway train," (G. Wells, quoted by Hansen, 2012).

Of course, reducing the rate of false IDs is a noble goal, and an understandable initial reaction to the tragic false convictions of people like Ronald Cotton (Thompson-Cannino, Cotton, & Torneo, 2009), Kirk Bloodsworth (Junkin, 2004), and too many others (e.g., Garrett, 2011). False convictions take a terrible toll on the falsely convicted and his or her family. False convictions also take a financial toll. An investigation by the Better Government

Association and the Center on Wrongful Convictions at Northwestern University School of Law showed that false convictions for violent crimes cost Illinois taxpayers $214 million (*Chicago Sun Times*, October 5, 2011). A recent update estimates that the costs will top $300 million (http://www.bettergov.org/wrongful_conviction_costs_keep_climbing, April, 2013).

But the narrative surrounding these reforms was distorted by this understandable focus on the innocent. For example, Wells et al. (2000, p. 585) wrote: "Surrounding an innocent suspect in a lineup with dissimilar fillers increases the risk that the innocent suspect will be identified (Lindsay & Wells, 1980)." That is undoubtedly true, but surrounding a *guilty* suspect in a lineup with dissimilar fillers also increases the chances that a guilty suspect will be chosen. Both innocent suspect and guilty suspect choosing rates must be considered. A full understanding of the contribution of factors like lineup fairness to eyewitness decision making requires consideration of both sides of the story.

The other side of the story is that if an innocent person is convicted of a crime, the actual perpetrator remains free and capable of committing more crimes. The aforementioned *Sun Times* article also reported on the new victims that arose from the 14 murders, 11 sexual assaults, 10 kidnappings, and at least 62 other felonies committed by the actual Illinois perpetrators, free while innocent men and women served time for these crimes. Similar occurrences are conceivable if a reform merely induces more conservative responding, which decreases the rate of false IDs (the benefit) but also decreases the rate of correct IDs (the cost). The ideal reform would seek to minimize costs *and* maximize benefits.

3.2 Discriminability versus Response Bias

An eyewitness ID from a lineup involves a recognition decision. That is, the options are provided to the eyewitness, who has the choice to select someone deemed to be the perpetrator, or to reject the lineup if the perpetrator is deemed not to be present. But because there are a limited number of options available, it is possible that an eyewitness can be "correct" (choose the suspect) by chance. For example, if there are five fillers and one suspect in the lineup, even someone with no memory for the perpetrator but who nevertheless makes an ID from the lineup has a one in six chance of picking the suspect. Consequently, it is important to take into account this "success by chance" when dealing with recognition memory data, especially because "success by chance" varies across individuals (and testing situations) due to differences in the willingness to make a response. An example will make this clear.

Imagine that students are randomly assigned into one of two groups: a neutral group or a conservative group. All students take an identical multiple-choice exam, but one in which the students can choose not to respond to every question. The neutral group is awarded +1 point for each correct answer and deducted −1 point for each incorrect answer. The conservative group receives +1 point for each correct answer but −10 points for each incorrect answer. Because the cost of making an error is much greater in the conservative group, the students in this group will be less likely to answer a question. Instead, these students will make a response only if they are highly likely to be correct (i.e., highly confident). They have set a "conservative" criterion for making a response. As a result of their conservative criterion, Table 2 reveals that these students have only responded correctly to 48% of the questions (in this hypothetical example). In contrast, the students in the neutral group will be more likely to answer the questions because they are penalized less for an incorrect answer. As a result of their "liberal" criterion, they have responded correctly to 82% of the questions.

Would it be fair to assign grades (which reflect course knowledge) based on percent correct? No, because the conservative group will be more careful when responding because the cost of an error is high. This results in fewer correct answers. But the differential cost of an error affects only the students' willingness to respond (affecting response bias), not their course knowledge (not affecting discriminability, which is the ability to distinguish correct answers from fillers). Note also the corresponding role that confidence plays in the answers that are offered. The conservative students will only answer those questions for which they are highly confident whereas the neutral students will be highly confident in some answers but will answer other questions (some correctly) despite being less than certain.

In recognition memory, the need to disentangle discriminability from response bias has long been known (e.g., Banks, 1970; Egan, 1958). The principal solution to this problem in the recognition memory literature involves the application of signal-detection theory (SDT) (e.g., Macmillan & Creelman, 2005). SDT provides a means of separately estimating, from a hit (correct ID) and false alarm (akin to a false ID) rate, an index of

Table 2 Hypothetical data from the neutral and conservative groups

	% Correct	Hit rate	False alarm rate	d'	β
Neutral group	82%	0.82	0.14	2.00	0.165
Conservative group	48%	0.48	0.02	2.00	2.108

discriminability (d') and a separate index of response bias (i.e., a willingness to make a response, e.g., β).

The hypothetical data from the neutral and conservative groups are shown in Table 2. The neutral group has a higher percent correct, hit rate, and false alarm rate than the conservative group, but d' is identical. That means the groups have the same ability to distinguish correct answers from fillers, but the response bias differs, as reflected by the β values (which is higher for the conservative group). Despite the fact that the need to separate discriminability and response bias has been known since the 1950s, eyewitness researchers often relied on measures that conflated the two, as we shall see next.

3.3 Measurement Issues

The neutral versus conservative students' example illustrates that one cannot simply rely on a direct comparison of correct ID rates (or hit rates) across, for example, simultaneous versus sequential presentation methods, to determine which one is superior. Eyewitness researchers recognized this fact, and therefore jointly considered correct and false IDs to compute an index of the probative value of an eyewitness ID. One common probative value measure, the diagnosticity ratio (Wells & Lindsay, 1980), took the ratio of the correct ID rate to the false ID rate. If the diagnosticity ratio equals 1.0, it indicates that the eyewitness evidence has no probative value; a chosen suspect is just as likely to be innocent as guilty. But as that ratio grows, it signals that the suspect is increasingly likely to be guilty rather than innocent. It is assumed that the best lineup presentation method is the one that maximizes the diagnosticity ratio, and the reforms were evaluated relying on this (or a related ratio-based) measure.

3.3.1 Diagnosticity Ratio

As revealed by Wixted and Mickes (2012), the problem with comparing one diagnosticity ratio from (for example) simultaneous presentation to one diagnosticity ratio from sequential presentation is that the diagnosticity ratio changes as response bias changes. In particular, the diagnosticity ratio increases as the response bias becomes more conservative. Gronlund, Carlson, et al. (2012) and Mickes, Flowe, and Wixted (2012) demonstrated this empirically. Wixted and Mickes (2014) showed how this prediction follows from SDT; Clark, Erickson, and Breneman (2011) used the WITNESS model to show the same result. The problem is obvious: If a range of diagnosticity ratios can arise from a simultaneous lineup test, which value should

be used to compare to a sequential lineup test? (Rotello, Heit, and Dubé (in press) illustrate how similar problems with dependent variables in other domains have led to erroneous conclusions.) The solution proposed by Wixted and Mickes (2012) was to conduct a receiver operating characteristic (ROC) analysis of eyewitness IDs. ROC analysis traces out discriminability across *all* levels of response bias. It is a method widely used in a variety of diagnostic domains including weather forecasting, materials testing, and medical imaging (for reviews see Swets, 1988; Swets, Dawes, & Monahan, 2000), and is an analytic (and nonparametric) technique closely tied to SDT.

In the basic cognitive psychology literature, SDT has long been used to conceptualize the level of confidence associated with a recognition memory decision. SDT is useful for conceptualizing an eyewitness task because a lineup is a special type of recognition test, one in which an eyewitness views a variety of alternatives and then makes a decision to either identify one person or to reject the lineup. The specific version of SDT that has most often been applied to recognition memory is the unequal-variance signal-detection (UVSD) model (Egan, 1958).

In the context of eyewitness memory, the UVSD model specifies how the subjective experience of the memory strength of the individuals in the lineup is distributed across the population of guilty suspects (targets) and innocent suspects (lures). Assuming the use of fair lineups in which the innocent suspect does not resemble the perpetrator any more than the fillers do, the lure distribution also represents the fillers in a lineup. The model represents a large population of possible suspects and fillers (hence the distributions), although in any individual case there is only one suspect and (typically) five fillers in a lineup. According to this model (illustrated in Figure 1), the mean and standard deviation of the target distribution (the

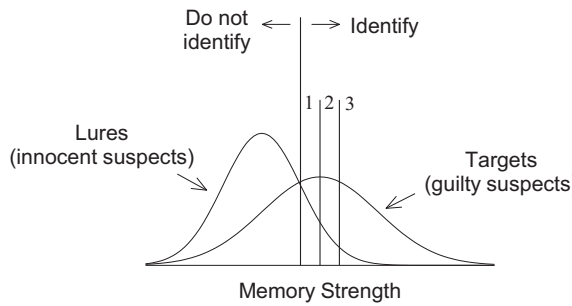

Figure 1 A depiction of the standard unequal-variance signal-detection model for three different levels of confidence, low (1), medium (2), and high (3).

actual perpetrators) are both greater than the corresponding values for the lure distribution.

A key assumption of SDT is that a decision criterion is placed somewhere on the memory strength axis, such that an ID is made if the memory strength of a face (target or lure) exceeds it. The correct ID rate is represented by the proportion of the target distribution that falls to the right of the decision criterion, and the false ID rate is represented by the proportion of the lure distribution that falls to the right of the decision criterion. These theoretical considerations apply directly to eyewitness' decisions made using a showup (i.e., where a single suspect is presented to the eyewitness, for a review see Neuschatz et al., in press), but they also apply to decisions made from a lineup once an appropriate decision rule is specified (Clark et al., 2011; Fife, Perry, & Gronlund, 2014; Wixted & Mickes, 2014). One simple rule holds that eyewitnesses first determine the individual in the simultaneous lineup who most closely resembles their memory for the perpetrator and then identify that lineup member if the subjective memory strength for that individual exceeds the decision criterion.

Figure 1 also shows how SDT conceptualizes confidence ratings associated with IDs made with different degrees of confidence (1 = low confidence, 2 = medium confidence, and 3 = high confidence). Theoretically, the decision to identify a target or a lure with low confidence is made when memory strength is high enough to support a confidence rating of 1, but is not high enough to support a confidence rating of 2 (i.e., when memory strength falls between the first and second decision criteria). Similarly, a decision to identify a target or a lure with the next highest level of confidence is made when memory strength is sufficient to support a confidence rating of at least 2 (but not 3). A high-confidence rating of 3 is made when memory strength is strong enough to exceed the rightmost criterion.

An ROC curve is constructed by plotting correct IDs as a function of false IDs. Figure 2 (left-hand panel) depicts an ROC curve based on the signal-detection model in Figure 1. For the left-hand-most point on the ROC, the correct ID rate is based on the proportion of the target distribution that exceeds the high-confidence criterion (3), and the false ID rate is based on the proportion of the lure distribution that exceeds that same criterion. For the next point on the ROC, the correct ID rate reflects the proportion of the target distribution that exceeds the medium-confidence criterion (2), and the false ID rate is based on the proportion of the lure distribution that exceeds that same criterion. The correct and false ID rates continue to accumulate across all the decision criteria, sweeping out a curve

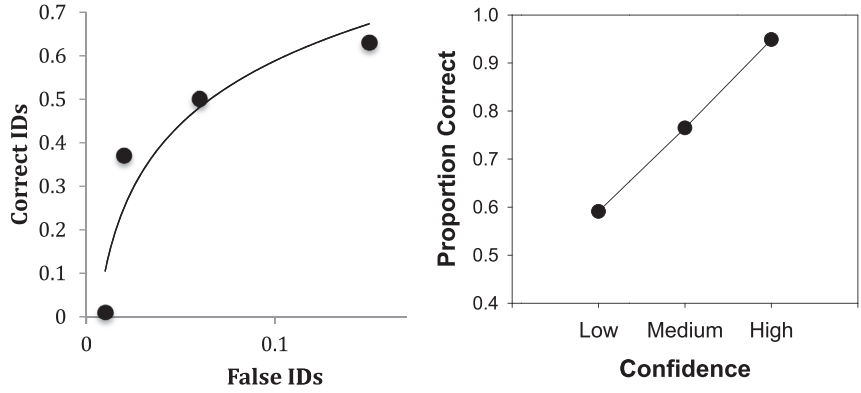

Figure 2 The left-hand panel depicts a receiver operating characteristic curve based on the signal-detection model in Figure 1. The high-confidence criterion results in a correct ID rate of 0.37 and a false ID rate of 0.02; the medium-confidence criterion results in a correct ID rate of the 0.50 and a false ID rate of 0.06; the low-confidence criterion results in a correct ID rate of 0.63 and a false ID rate of 0.15. The right-hand panel depicts the calibration curve for the same model using these same response proportions. For a calibration curve, the proportion correct in each confidence category (0.37/(0.37 + 0.02); 0.13/(0.13 + 0.04); 0.13/(0.13 + 0.09)) is plotted as a function of subjective confidence.

that displays the discriminability for a given reform as a function of different response biases. The best performing reform is indicated by the ROC curve closest to the upper left-hand corner of the space. See Gronlund, Wixted, and Mickes (2014) for more details about conducting ROC analyses in lineup studies.

The reliance on measures like the diagnosticity ratio that conflate discriminability and response bias led researchers to conclude that some of the recommended reforms were more accurate than the procedure they were replacing (Clark et al., 2014). However, as we shall see, several of the recommended reforms were merely more conservative in terms of response bias, not more accurate. Moreover, the reliance on measures that conflated discriminability and bias was not the only measurement issue that led eyewitness researchers astray. The widespread use of an unsuitable correlation measure also allowed an incorrect conclusion to be reached regarding the relationship between confidence and accuracy.

3.3.2 Point-Biserial Correlation

The relationship between eyewitness confidence in an ID decision and the accuracy of that decision was evaluated by computing the point-biserial correlation. The point-biserial correlation assesses the degree of relationship

between accuracy, coded as either correct or incorrect, and subjective confidence. Research at the time the reforms were proposed showed a weak to moderate relationship between confidence and accuracy. Wells and Murray (1984) found a correlation of only 0.07, although a higher correlation (0.41) was reported when the focus was on only those individuals who made a choice from the lineup (Sporer, Penrod, Read, & Cutler, 1995). This seemingly unimpressive relationship[1] between confidence and accuracy dovetailed nicely with the malleability of confidence demonstrated by Wells and Bradfield (1998). This is why an eyewitness' assessment of confidence played little role in the reforms. But that began to change with a report by Juslin, Olsson, and Winman (1996).

Juslin et al. (1996) argued that eyewitness researchers needed to examine the relationship between confidence and accuracy using calibration curves. Calibration curves plot the relative frequency of correct IDs as a function of the different confidence categories (i.e., the subjective probability that the person chosen is the perpetrator). Figure 2 (right-hand panel) depicts a calibration curve based on the signal-detection model in Figure 1. In contrast to the construction of ROC curves, where we compute the area in the target and lure distributions that fall above a confidence criterion, here we take the areas in the target and lure distributions that fall between adjacent confidence criteria. For example, 13% of the target distribution falls above criterion 1 but below criterion 2, with 9% of the lure distribution falling in that same range. That means that the accuracy of these low-confidence suspect IDs is 13/(13 + 9) or 59%. The accuracy is higher for those suspect IDs that fall between criteria 2 and 3, 13% of the target distribution and 4% of the lure distribution, making the accuracy 77% (13/(13 + 4)). Finally, the accuracy is higher still for the highest confidence suspect IDs, those that fall above criterion 3 (95% = 37/(37 + 2)).

Juslin et al. (their Figure 1) showed that the point-biserial correlation masked the relationship between confidence and accuracy. To illustrate the point, they simulated data that exhibited perfect calibration; perfect calibration implies that (for example) participants that are 70% certain of a correct ID have 70% correct IDs. But by varying the distribution of responses across the confidence categories, Juslin et al. showed that the point-biserial

[1] Although r is not the best statistic for evaluating the relationship between confidence and accuracy, $r = 0.41$ actually signals a strong relationship. The first clinical trial for a successful AIDS drug was so successful that the research was halted so that the control group could also get the drug: $r = 0.28$ was the effect size (Barnes, 1986).

correlation could vary from 0 to 1 despite perfect calibration. More recent efforts (e.g., Brewer & Wells, 2006) using calibration show a much stronger relationship between confidence and accuracy than was understood at the time the reforms were proposed. We shall return to the implications of this finding.

The reliance on measures that conflated discriminability and response bias, or masked the relationship between confidence and accuracy, was major contributor to how the impact of the eyewitness reforms came to be misconstrued. Another major contributor was the role of a theory developed in response to the initial empirical tests of the reforms.

3.4 Role of Theory

> *Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve*
>
> **Popper (1972).**

Theory is vital to the evolution of a science. Theories are testable; they organize data, help one to conceptualize why the data exhibit the patterns they do, and point to new predictions that can be tested. However, theory also can distort data through confirmation biases, publication biases, and selective reporting (see Clark et al., 2014; Ioannidis, 2008; Simmons, Simonsohn, & Nelson, 2011). We believe that this distorting effect of theory is especially likely when two conditions are met. First, a theory has the potential to distort when it is not formally specified. It is difficult to extract definitive predictions from verbally specified theories (Bjork, 1973; Lewandowsky, 1993) because the lack of formalism makes the workings of the model vague and too flexible. A formally specified theory, on the other hand, forces a theoretician to be explicit (and complete) about the assumptions that are made, which make transparent the reasons for its predictions, and provides a check on the biases of reasoning (Hintzman, 1991). Second, a theory has the potential to distort when it has no competitors (Jewett, 2005; Platt, 1964). Such was the state of the field of eyewitness memory at the time of the reforms.

Relative judgment theory has been the organizing theory for eyewitness memory for 30 years (Wells, 1984, 1993). Wells proposed that faulty eyewitness decisions largely arose from a reliance on relative judgments. Relative judgments involve choosing the individual from the lineup who looks most like (is the best match to) the memory of the perpetrator relative

to the other individuals in the lineup. An extreme version of relative judgment theory would have an eyewitness choosing someone from every lineup, but that is not what happens. Instead, a decision criterion is needed to determine if the best-matching individual from a lineup should be chosen or whether the lineup should be rejected. Wells contrasted relative judgments with absolute judgments. Absolute judgments involve determining how well each individual in the lineup matches memory for the perpetrator, and results in choosing the best-matching individual if its match strength exceeds a decision criterion. Absolute judgments are assumed to entail no contribution from the other lineup members.

In addition to the absolute-relative dichotomy, comparable dichotomies posited other "reliable versus unreliable" contributors to eyewitness decisions (see also Clark & Gronlund, 2015). One dichotomy was automatic versus deliberative processes (Charman & Wells, 2007; Dunning & Stern, 1994); a deliberative strategy (e.g., a process of elimination) was deemed inferior to automatic detection ("his face popped out at me"). A second dichotomy involved true recognition versus guessing (Steblay, Dysart, & Wells, 2011). The additional correct IDs that arose from use of the nonreform procedure were deemed "lucky guesses" and therefore should be discounted because they were accompanied by additional false IDs. Irrespective of the dichotomy, the reforms were thought to be beneficial because they reduced reliance on these unreliable contributions. In what follows, we focus on the relative versus absolute dichotomy, although the arguments we make apply equally to the other dichotomies.

The initial version of relative judgment theory led people to believe that a reliance on absolute judgments reduced false IDs but not correct IDs. The first studies conducted comparing the reforms to the existing procedures reported data consistent with this outcome. The four reforms reviewed by Clark et al. (2014)—lineup instructions, lineup presentation, filler similarity, and filler selection[2]—showed an average *gain* in correct IDs for the reforms of 8%, and an average decrease in false IDs for the reforms of 19%. There apparently was no cost to the reforms in terms of reduced correct IDs, and a clear benefit in terms of reduced false IDs. Clark (2012) called this the no-cost view; Clark and Gronlund (2015) referred to it as the strong version of relative judgment theory's accuracy claim. In other words, the

[2] Granted, description-matched filler selection was designed to increase the correct ID rate relative to suspect-matched filler selection, so the increase in the correct ID rate should not be viewed as surprising for that reform.

shift from relative to absolute judgments reduces false ID rates but has little effect on correct ID rates, thereby producing a "no-cost" accuracy increase. This was the version of relative judgment theory in place at the time the reforms were enacted. An SDT alternative would intuitively predict a trade-off between costs and benefits arising from these reforms. But because the reforms appeared to increase accuracy rather than engender a criterion shift, a signal-detection-based alternative explanation failed to materialize as a competitor theory.

Most scientific theories evolve as challenging data begin to accumulate, but principled modifications need to be clearly stated and the resulting predictions transparent. However, this may not be the case when a verbally specified theory is guiding research. As conflicting evidence began to accumulate contrary to the strong version (see summary by Clark, 2012), a weak version arose that claimed that the proportional decrease in false IDs is greater than the proportional decrease in correct IDs. But without a clear operationalization of how the model worked, it was not clear whether this was really what relative judgment theory had predicted all along (Clark et al., 2011). We suspect that if this trade-off was acknowledged sooner, an SDT alternative might have challenged the widespread acceptance of relative judgment theory. The following example makes clear the role a competitor theory can play in interpreting data.

One of the major sources of empirical support for relative judgment theory came from an experiment by Wells (1993). Participants viewed a staged crime, and then were randomly assigned to view either a 6-person target-present lineup or a 5-person target-removed lineup. The target-present lineup contained the guilty suspect and five fillers; the target-removed lineup included only the five fillers. In the target-present lineup, 54% of the participants chose the guilty suspect and 21% rejected the lineup. According to the logic of relative judgment theory, if participants are relying on absolute judgments when they make eyewitness decisions, approximately 75% of the participants should have rejected the target-removed lineup: the 54% that could have identified the guilty suspect if he had been present, plus the 21% that would even reject the lineup that included the guilty suspect. But instead, in apparent support for the contention that eyewitnesses rely on relative judgments, most target-removed participants selected a filler (the next-best option). The target-removed rejection rate was only 32%, not 75%. This finding is considered by many (Greene & Heilbrun, 2011; Steblay & Loftus, 2013; Wells et al., 1998) to offer strong support for the fact that eyewitnesses rely on relative judgments.

Although this result is intuitively compelling, it is difficult to definitively evaluate the predictions because the predictions arose from a verbally specified model. There are many examples of this in the wider literature. To take one example from the categorization literature: Do we summarize our knowledge about a category (e.g., birds) by storing in memory a summary prototype that captures most of the characteristics shared by most of the category members, or do we instead store all the category examples we experience? Posner and Keele (1970) showed that participants responded to a category prototype more strongly than to a specific exemplar from the category, even if the prototype had never before been experienced. This was thought to demonstrate strong evidence for the psychological reality of prototypes as underlying categorization decisions. But Hintzman (1986) took a formally specified memory model that stored only exemplars and reproduced the same performance advantage for the test of a prototype. The model accomplished this because it made decisions by matching a test item to everything in memory. Although a prototype matches nothing exactly, as the "average" stimulus, it closely matches everything resulting in a strong response from memory.

Clark and Gronlund (2015) applied a version of the WITNESS model (Clark, 2003) to Wells' (1993) target-removed data. The WITNESS model is a formally specified model of eyewitness decision making, and one that has an SDT foundation. Consequently, the model can provide definitive predictions, as well as serve as a competitor to relative judgment theory. Clark and Gronlund implemented a version of WITNESS that makes absolute judgments (compares a lineup member to criterion and chooses that lineup member if the criterion is exceeded). They showed that the model could closely approximate the Wells' data. This is unexpected given that these data are regarded as providing definitive evidence of the reliance on relative judgments. Moreover, a formal model reveals an explanation for the data that a verbally specified theory often cannot. Assume that there are two lineup alternatives above criterion in the target-present lineup. One of those typically is the target, and the other we refer to as the next-best. Because the target, on average, will match memory for the perpetrator better than the next-best, the target is frequently chosen. But it is clear that by moving that same lineup into the target-removed condition (sans the target), the same decision criterion results in the choosing of the next-best option. That is, the "target-to-filler-shift" thought indicative of a reliance of relative judgments may signal nothing of the sort. This raises questions about the empirical support favoring relative judgment theory.

Clark et al. (2011) undertook an extensive exploration of relative and absolute judgments in the WITNESS model to seek theoretical support for the superiority of absolute judgments. They explored the parameter space widely for both description-matched (same fillers in target-present and target-absent lineups) and suspect-matched (different fillers in target-present and target-absent lineups). They found that relative versus absolute judgments made little difference for description-matched lineups in many circumstances (see also Goodsell, Gronlund, & Carlson, 2010); some circumstances exhibited a slight relative judgment advantage. In contrast, the suspect-matched lineups showed a more robust absolute judgment advantage. Here was the theoretical support for the predictions of relative judgment theory; a reliance on absolute judgments did enhance performance for the types of lineups that the police typically construct.

But Fife et al. (2014) limited the scope of this finding. They showed that the WITNESS model parameters that govern the proportional contributions of relative versus absolute judgments covary with the decision criterion. That means that the model typically is unable to uniquely identify the proportion of relative versus absolute judgment contribution given only ID data. Figure 3 shows three ROC curves generated by the WITNESS model for the largest absolute judgment advantage reported by Clark et al. (2011). Although there is a detectable difference between a 100% relative and 0% relative judgment rule, there is little difference between a 0% relative rule and a 75% relative rule. This is not strong evidence for the superiority of absolute judgments if a model that is predominantly relative (75%) is very similar to one that is absolute (0% relative). At the present time, both the empirical and the theoretical support for the predictions of relative judgment theory are unsettled. Indeed, Wixted and Mickes (2014) suggested that comparisons among lineup members (a form of relative judgment) actually facilitate the ability of eyewitnesses to discriminate innocent versus guilty suspects.

Fully understanding the theoretical contributions of relative versus absolute judgments to eyewitness ID decision making will require more work. The aforementioned parameter trade-off may not arise if relative-absolute judgments are instantiated differently in the WITNESS model, or if additional data like confidence or reaction times are considered. Moreover, as Clark et al. (2011) noted, the empirical evaluation of these predictions also is complicated by a number of factors. For example, it is unlikely that any experimental manipulation would be so strong that all of the participants in one condition would use a pure absolute judgment strategy and all of

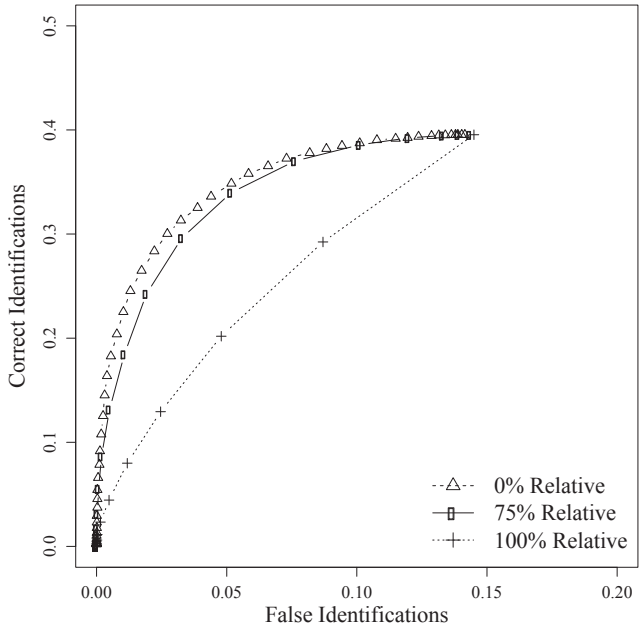

Figure 3 Three receiver operating characteristic curves generated by the WITNESS model for the largest absolute judgment advantage reported by Clark et al. (2011). Although there is a difference between a 100% relative and 0% relative judgment rule, there is little difference between a 0% relative rule (i.e., an absolute rule) and a 75% relative rule. *Figure modified with kind permission from Springer Science and Business Media, Psychonomic Bulletin & Review, (2014), 21, 479–487, Revisiting absolute and relative judgments in the WITNESS model., Fife, D., Perry, C., & Gronlund, S. D., Figure 4.*

the participants in the other condition would use a pure relative judgment strategy. To the extent that the manipulation is not 100% successful, or that participants use a mixed strategy, the differences might be difficult to detect empirically.

A theory can abet confusion within a research area in several ways. It can engender confirmation biases. For instance, in a meta-analysis comparing simultaneous and sequential lineups, Steblay et al. (2011) reported that the sequential lineup produced a 22% decrease in the false IDs compared to the simultaneous lineup, compared to only an 8% decrease in correct IDs arising from sequential lineups. (Clark (2012) reported other problems with this meta-analysis.) This result ostensibly signals clear support for the sequential lineup reform. But the 22% value was misleading because it arose from a failure to distinguish between filler IDs and false IDs. For studies that do not designate an innocent suspect, researchers typically estimate a false

ID rate by dividing the choosing rate by the number of fillers. Once the correction is made, the estimated decrease in the false ID rate resulting from sequential lineups is only 4% (Clark et al., 2014). Steblay (1997) made a similar error regarding the effectiveness of unbiased lineup instructions.

A theory also can induce publication biases. Clark et al. (2014) reported evidence of this in the Steblay et al. (2011) simultaneous-sequential meta-analysis. The unpublished studies reported by Steblay et al. showed a trade-off between costs (reduced correct IDs in sequential) and benefits (reduced false IDs in sequential). However, the studies that were published during this same period indicated that the benefits of sequential lineups outweighed the costs. In other words, the unpublished data supported a conservative criterion shift arising from sequential lineups, not a discriminability advantage.

4. REEVALUATION OF THE REFORMS

The narrative surrounding the reforms has changed in the last decade. The data have changed, shifting from showing the benefits of the reforms to showing that the reforms often produce a conservative criterion shift, not an improvement in discriminability. It took a while for researchers to sort this out for the reasons discussed above: an almost exclusive focus on protecting the potentially innocent suspect, reliance on measures that conflated discriminability and response bias, and the distorting role of relative judgment theory. In this next section, we assess the current state of the reforms, examining the recent data, the implications of the development of competing theories, and the broader implications of more clearly assessing the relationship between confidence and accuracy. We begin with a current view of the empirical data.

4.1 Decline Effects

Clark et al. (2014) examined the evolution of the empirical findings regarding four of the reforms from the time of the initial studies through to those studies published by 2011. The reforms were: filler similarity, filler selection, lineup instructions, and lineup presentation. Recall that the comparison for filler similarity involves less versus more similar fillers; the comparison for filler selection involves suspect- versus description-matched fillers; the comparison for lineup instructions is between biased and unbiased instructions; and the comparison for lineup presentation is between simultaneous and sequential.

As Clark et al. (2014) reported, and we noted above, the first studies that made the comparisons between the initial procedures and the recommended reforms (filler similarity—Lindsay & Wells, 1980; filler selection—Wells et al., 1993; lineup instructions—Malpass & Devine, 1981; lineup presentation—Lindsay & Wells, 1985) resulted in data that exhibited no costs and large benefits. But Clark et al. showed that, when viewed in the context of the data that followed, those results were outliers. For example, they compared the d' difference between the recommended and the nonrecommended procedures. The average d' advantage favoring the reforms for these initial studies was 0.81. But the average d' difference for an aggregate of all studies was −0.02. Clark et al. also completed another assessment of the representativeness of the initial studies, determining what proportion of studies had results less extreme than the results of the initial studies. For the d' comparisons, those proportions were 0.91, 0.97, 0.89, and 0.87, for filler similarity, filler selection, lineup instructions, and lineup presentation, respectively. These initial studies were not poorly conducted, but in hindsight it is clear that their results were unrepresentative, and too influential. Table 3 provides a summary of the current view of these eyewitness reforms.

The discriminability benefit of the reforms reported in the initial studies did not withstand the test of replication. Ioannidis (2008; Schooler, 2011) calls these *decline effects*. Decline effects are not unique to psychology, and there are many factors that contribute including publication bias and the file-drawer problem, a bias toward publishing positive results (not null effects), the biasing effect of initial studies, and the distorting role of theory. The data as they stand today provide no support for these four reforms if the criterion for success is increased accuracy (i.e., discriminability). A report

Table 3 Current understanding of the impact of these eyewitness reforms

Reform	Current view
Fair fillers	Induces more conservative responding but no change to discriminability
Description-matched fillers	Induces more conservative responding but no change to discriminability
Unbiased instructions	Induces more conservative responding but no change to discriminability
Sequential presentation	Induces more conservative responding but reduces discriminability
Role for initial confidence	Initial eyewitness confidence is meaningfully related to eyewitness accuracy

released by the National Academy of Sciences in October, 2014 (*Identifying the Culprit: Assessing Eyewitness Identification*), stated "The committee concludes that there should be no debate about the value of greater discriminability — to promote a lineup procedure that yields less discriminability would be akin to advocating that the lineup be performed in dim instead of bright light," p. 80.

4.2 Alternative Theoretical Formulations

Relative judgment theory dominated the eyewitness literature for 30 years, and the time has come to consider alternative theoretical formulations. Here we consider three: a signal-detection-based theory, the question of whether eyewitness memory is mediated by discrete processes or a continuous underlying substrate, and consideration of the role recollection might play in eyewitness decision making.

4.2.1 Signal-Detection Alternative

Relative judgment theory purported to explain how the recommended reforms reduced reliance on relative judgments and encouraged reliance on absolute judgments. It claimed to describe how it was that correct IDs were little affected by these reforms, but false IDs would decrease. As mentioned above, no theoretical alternative arose to challenge relative judgment theory. But in light of the results reported by Clark et al. (2014), an alternative theoretical approach is needed. Moreover, calls have been made for that theory to be formally specified (e.g., Clark, 2008). Clark's (2003; Clark et al., 2011) WITNESS model was the first signal-detection-based theory to answer that call (Clark et al., 2011). Recently, Wixted and Mickes (2014) proposed an alternative theoretical implementation. We consider the Wixted and Mickes theory here because it explicitly addresses ideas that have been raised in this chapter, including the need for ROC analysis of lineup data, and, due to its grounding in SDT, the strong positive relationship between eyewitness confidence and accuracy.

The theory, embedded in an UVSD framework, is depicted in Figure 1. One of the things that makes the theory beneficial is the way in which it can enhance our understanding of relative judgment theory. For example, Wixted and Mickes (2014) illustrated that the diagnosticity ratio increases as response bias becomes more conservative. We can illustrate the same thing using the criteria depicted in Figure 1. For the distributions depicted, the correct and false ID rates for the most liberal criterion (1) are 0.63 and 0.15, making the diagnosticity ratio equal to 4.2 (0.63/0.15). Recall that

the correct ID rate is based on the proportion of the target distribution that lies above criterion 1; the false ID rate is based on the proportion of the lure distribution that lies above criterion 1. For the more conservative criterion 2, the correct and false ID rates are 0.50 and 0.06, and the diagnosticity ratio increases to 8.3. For the even more conservative criterion 3, the correct and false ID rates are 0.37 and 0.02 and the diagnosticity ratio even greater at 18.5. We can bookend these values by selecting the most liberal criterion setting at the far left tails of the distributions, which would result in correct and false ID rates of essentially 1.0 and 1.0 and a diagnosticity ratio of approximately 1.0. At the other extreme, we can set the criterion far out in the right-hand tail of the target distribution, where the false ID rate becomes vanishingly small (e.g., 0.001), greatly increasing the diagnosticity ratio (>50). Note that the diagnosticity ratio varies over this wide range despite the discriminability, by definition, not changing. For a more detailed treatment of why the diagnosticity ratio and response bias are related in this manner, see the Appendix in Wixted and Mickes. For empirical confirmation, see Gronlund, Carlson, et al. (2012) and Mickes et al. (2012). In sum, viewed through an SDT framework, it is clear why the diagnosticity ratio is an inappropriate measure for evaluating reforms that induce changes in response biases. Moreover, it underscores the necessity for ROC analysis to assess these reforms.

4.2.2 Continuous or Discrete Mediation

The UVSD model assumes that continuous latent memory strengths mediate recognition judgments. The memory strengths could arise from a familiarity process (e.g., Gillund & Shiffrin, 1984), or as the sum of familiarity and a graded recollection signal (Wixted & Mickes, 2010), or as a match value representing the similarity of the test face to the memory of the perpetrator (Clark, 2003). A face in the lineup is matched to memory and the resulting memory signal is compared to a decision criterion. A positive response is made if the memory signal exceeds criterion, otherwise a negative response is made. If the test face had been studied previously, the response would be classified a hit (a correct ID), but if the test face had not been studied previously, the response would be classified as a false alarm (a false ID). The continuous evidence that mediates recognition judgments in the UVSD model can be contrasted with the discrete mediation posited by Wells and colleagues.

Wells, Steblay, and Dysart (2012; Steblay et al., 2011) proposed that, in addition to the reforms purportedly increasing the likelihood that

eyewitnesses rely on absolute judgments, they also implicitly posited that discrete processes mediated recognition memory in eyewitness ID. They called the two processes (among other labels) "true recognition" and "guessing." Wells and colleagues assumed that if a face in a lineup is the perpetrator, there are two paths by which that face could be identified. One path relies on a detection process (many would equate detection with recollection, e.g., see Yonelinas, 1994). If the perpetrator is detected, he is positively identified with high confidence. Wells et al. referred to this as a *legitimate* hit. However, if the detect process fails, an eyewitness can still make a guess and select the perpetrator with a $1/n$ chance (where n is the size of the lineup). (If the lineup is biased, the likelihood of guessing the perpetrator could be greater than $1/n$.) Wells et al. referred to this as an *illegitimate* hit. The idea of the reforms was that it would reduce eyewitnesses' reliance on guessing (reduce illegitimate hits) and move them toward judgments based on true recognition (legitimate hits). Wells and colleagues' idea revisits the debate between discrete-state and continuous signal-detection-based models from the basic recognition memory literature (for a review see Egan, 1975; Macmillan & Creelman, 2005; Pazzaglia, Dubé, & Rotello, 2013).

The operation of recognition memory as described by Wells and colleagues is reminiscent of a single high-threshold recognition memory theory (Swets, 1961). For example, take the perpetrator from the target-present lineup. The assumption is that participants respond from one of two cognitive states, detect or nondetect. One probability governs the likelihood of detecting the perpetrator, and with the complementary probability participants enter the nondetect state, a state from which they make a guess. If the lineup is fair, the probability of guessing the perpetrator is $1/n$.

The standard testing grounds for these two classes of models in the recognition memory literature has been the shape of ROC curves (Green & Swets, 1966). Discrete-state models predict linear ROC functions; continuous evidence models generally predict curvilinear ROC functions. The data generally are consistent with continuous evidence models (Pazzaglia et al., 2013). But recently, discrete-state models have been proposed that relax assumptions regarding how detect states are mapped onto response categories (Province & Rouder, 2012), allowing discrete-state models to produce curvilinear ROC functions. Alternative means of testing between these classes of models are being developed (e.g., Rouder, Province, Swagman, & Thiele, under review). Kellen and Klauer (2014) developed one such alternative. They had participants study lists of words, and varied the strength of these words by having some studied once (weak) and some studied three

times (strong). At test, sets of four words were presented, each set containing one previously studied word and three previously unstudied words. The participants ranked each word in the set from most-to-least likely to have been studied before. The key statistic to be computed is the conditional probability that a previously studied word would be ranked second given that it had not been ranked first. According to SDT, this conditional probability should increase as memory strength increases. In contrast, the discrete-state model predicts that the conditional probability should remain constant as memory strength increases. Kellen and Klauer showed that the conditional probability was greater for the strong memory tests, consistent with SDT and supporting the claim that continuous evidence mediates recognition memory. Work is underway utilizing this new paradigm in an eyewitness context to pit the UVSD and the true recognition accounts against one another.

4.2.3 Role for Recollection

The role that recollection might play in eyewitness ID needs to be explored further. Gronlund (2005) proposed a dual-process account for why sequential presentation could result in superior performance in some circumstances. Gronlund (2004) had participants study the heights of pairs of men and women depicted in photographs. Height information was presented either as the actual height (the man is 5′8″) or in a comparative manner (the man is taller than the woman). Recognition testing involved either the sequential or simultaneous presentation of different height options. Performance was especially good in the comparative height condition when the height of the man and woman was equal (man = woman). Specifically, when participants studied a man = woman pair, but the sequential presentation of the test options did not include that option, participants correctly rejected the test at very high rates. Gronlund (2005) proposed a dual-process explanation for these data, positing special encoding for the man = woman stimulus due to its distinctive status (Healy, Fendrich, Cunningham, & Till, 1987). Furthermore, because research has shown a tight coupling of distinctiveness and recollection (e.g., Dobbins, Kroll, Yonelinas, & Yui, 1998; Hunt, 2003; Mäntylä, 1997), Gronlund (2005) proposed that recollection was responsible for retrieving this distinctive information, and that recollection was more likely given sequential presentation. The consideration of multiple options in a simultaneous test could stretch limited cognitive resources that otherwise could be used to support recollection (e.g., Craik, Govoni, Naveh-Benjamin, & Anderson, 1996).

Carlson and Gronlund (2011) found support for a contribution of recollection using a face recognition paradigm. They varied perpetrator distinctiveness and sequential or simultaneous testing, and had participants make ID decisions and remember-know-guess (RKG) judgments (Gardiner & Richardson-Klavehn, 2000). They found evidence for the greater use of recollection (a recall-to-reject strategy, Rotello, 2001) in target-absent sequential lineups. But Meissner, Tredoux, Parker, and MacLin (2005) used a multiple-lineup paradigm and found no evidence of a greater reliance on recollection arising from sequential lineups. Finally, Palmer, Brewer, McKinnon, and Weber (2010) had participants view a mock crime and make ID decisions accompanied by RKG judgments and recollection ratings (which assessed graded recollection, e.g., Wixted, 2007). They found that correct IDs accompanied by a "remember" report were more accurate than those accompanied by a "know" report, but that benefit was redundant with the contribution of response confidence (an effect recently replicated by Mickes, in press). However, they found that they could better diagnose eyewitness accuracy by taking graded recollection *ratings* into account, even after ID confidence was considered.

Now that the influence of relative judgment theory is waning, there is much to be done theoretically to enrich our understanding of eyewitness decision making. It is vital to have a competitor theory, and that now exists (Clark, 2003; Wixted & Mickes, 2014). Moreover, these new theories are specified formally, which facilitates empirical and theoretical development. Next, the correspondence between true recognition/guessing and the single high-threshold model, allows Wells and colleagues' (Steblay et al., 2011; Wells et al., 2012) conjecture to be pitted against SDTs and subjected to empirical tests. Finally, dual-process conceptions of recognition involving either all-or-none or graded recollection contributions need to be explored. The next step in the evolution of the eyewitness reforms must be driven by theory, an idea upon which we will expand in Section 5.

4.3 Role for Confidence

The consensus at the time of the reforms, a view still widely held today (see Lacy & Stark, 2013), is that eyewitness confidence is not reliably related to ID accuracy. Krug (2007) reported that the confidence—accuracy relationship is "relatively weak or nonexistent." Moreover, confidence can be inflated by confirming feedback (e.g., Wells & Bradfield, 1998). In light of these conclusions, the New Jersey Supreme Court ruled (Henderson, 2011) that if a defendant can show that suggestive police procedures may have influenced

an eyewitness, but the judge nevertheless allows the eyewitness to testify, jurors will be instructed that eyewitness confidence is generally an unreliable indicator of accuracy (p. 5, http://www.judiciary.state.nj.us/pressrel/2012/jury_instruction.pdf). Nevertheless, jurors find high-confidence eyewitnesses to be very compelling (Cutler, Penrod, & Stuve, 1988), and the U.S. Supreme Court (Biggers, 1972) points to eyewitness confidence as one of the factors a judge should weigh to determine if an ID is reliable.

A signal-detection framework predicts a meaningful relationship between confidence and accuracy (Mickes, Hwe, Wais, & Wixted, 2011), and presenting the data as a calibration curve, as illustrated in the right-hand panel of Figure 2, best reveals this relationship. Recent data (e.g., Palmer, Brewer, Weber, & Nagesh, 2013) have supported the existence of this meaningful relationship. However, it is important to note that a meaningful relationship only holds for the confidence reported by an eyewitness at his or her *initial* ID attempt, before any confirming feedback is delivered and before any additional time has passed.

The existence of a meaningful confidence—accuracy relationship for an eyewitness' initial choice from a lineup changes the narrative surrounding eyewitness memory. It suggests that there is more to learn from an eyewitness report than has often been acknowledged. In light of these developments, Wixted, Mickes, Clark, Gronlund, and Roediger (in press) argued that jurors should weigh the confidence reported by an eyewitness during the initial ID. In other words, an ID accompanied by a confidence report of 95% is more likely to be correct than an ID accompanied by a confidence report of 60%. Of course, this does not imply that an eyewitness who is 100% confident is 100% accurate, but it does imply that an eyewitness who is 100% confident is (on average) much more likely to be accurate than one that is 60% confident. But more work remains to be done on a variety of issues involving confidence judgments, including how different eyewitnesses use the same scale, should eyewitnesses state their degree of confidence using their own words or a numeric scale, what scale is best to use, and how do the police decipher and interpret these confidence judgments (see Dodson & Dobolyi, in press).

Perhaps the most compelling evidence for the potential of a reliance on initial confidence comes from Garrett's (2011) analysis of 161 of the DNA exoneration cases in which faulty eyewitness evidence played a role. In 57% of these cases (92 out of 161), the eyewitnesses reported they had *not* been certain at the time of the initial ID. If this low confidence (or zero confidence for those eyewitnesses that initially selected a filler or rejected the

lineup) was taken seriously, these innocent individuals might never have been indicted and, consequently, never falsely convicted. However, if the criminal justice system is going to rely on eyewitness confidence, it provides important motivation for conducting double-blind lineup testing to eliminate feedback that could taint the initial confidence report.

The development of new theory has cast relative judgment theory and the reforms in a new light. A signal-detection-based theory is consistent with the empirical results as they currently stand. This includes the meaningful relationship between initial confidence and accuracy. Also, three of the reforms (filler similarity, filler selection, unbiased instructions) can be understood as inducing a conservative criterion shift. In contrast, sequential presentation actually reduces discriminability (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Mickes et al., 2012). How does new theory address that result?

Wixted and Mickes (2014; see also Goodsell et al., 2010) proposed a diagnostic-feature-detection hypothesis to explain the reduced discriminability of sequential lineup presentation. Discriminability from simultaneous lineups is superior because, by seeing all the options at once, eyewitnesses can determine what features to pay attention to and what features are redundant and therefore not diagnostic. For example, if all the individuals in the lineup are young Hispanic males with shaved heads, putting attention on any of those cues will not help discriminate the lineup members. Generally speaking, focusing on shared (i.e., nondiagnostic) features will not help eyewitness to distinguish between innocent and guilty suspects. Rather, eyewitnesses must attend to the diagnostic cues that will differentiate the perpetrator from the fillers—and from innocent suspects. Eyewitnesses viewing a sequential lineup can engage in the same type of sorting of nondiagnostic from diagnostic cues as the lineup unfolds. After viewing the second young bald Hispanic male, eyewitness can shift attention to other cues. Consequently, discrimination is predicted to be superior when the suspect (guilty or innocent) is placed later in the sequential lineup. This is what Carlson, Gronlund, and Clark (2008) and Gronlund, Carlson, Dailey, and Goodsell (2009) have found. Clearly, new theory can point to new avenues for exploration, the proposed reliance on initial eyewitness confidence being the first such avenue.

5. FOUNDATION FOR NEXT-GENERATION REFORMS

The next generation of reforms must be grounded in theory (see also McQuiston-Surrett, Tredoux, & Malpass, 2006). An explanation for how

and why a reform does what it claims provides a foundation for making inferences about how the reform will perform in other situations. One criticism of the application of psychological research to real criminal cases is that the conclusions reached in the lab do not exactly match, or are not entirely germane, to real world cases (Konecni & Ebbesen, 1979). How does one determine if the circumstances surrounding the particular crime under discussion, given this particular eyewitness, and this particular lineup, sufficiently resemble the circumstances surrounding the experiment being discussed? Of course, that goal can never be attained, because all possible experiments can never be conducted. However, the answer that can be provided is to develop theory that seeks to understand how various empirical circumstances affect a reform.

5.1 Theory-Driven Research

Hugo Münsterberg (1908) typically gets the credit for conducting the first experimental research directed at integrating psychology and the law. Munsterberg wrote about a number of factors that can change a trial's outcome, including faulty eyewitness ID and untrue confessions. But Munsterberg also is relevant to the argument we have made regarding how the field reached the wrong conclusions regarding some of the reforms. For that purpose, it is helpful to contrast Munsterberg with one of his contemporaries, Arnold (1906; cited in Bornstein & Penrod, 2008). Munsterberg and Arnold took different approaches to the examination of eyewitness memory. Munsterberg took an applied approach to the problem, and made frequent use of examples and anecdotes, but Arnold saw value in theory. Arnold was concerned about processes and general principles of memory. Munsterberg's approach carried the day in psychology and law research, and a focus on phenomena, cases, and applications, was to the detriment of research progress in the field. We are not the first to make this appraisal (Lane & Meissner, 2008).

Eyewitness research needs to be conducted in concert with the development and evaluation of theory. However, theory testing will require conducting different kinds of experiments than have been the norm. Theory testing will require a shift from maximizing the external validity and realism of the studies, to a focus on internal validity and the underlying psychological processes that operate to produce various phenomena. This will necessitate experiments that generate more than one observation per participant. For example, Meissner et al. (2005) used a multiple-lineup paradigm to evaluate the contributions of recollection and familiarity in simultaneous and

sequential lineups. Participants studied eight faces in succession, and then were tested using 16 lineups (a target-present and a target-absent lineup for each studied face). To test theory, we often need to analyze performance at the level of the individual rather than at the level of the group. Of course, highly controlled laboratory experiments are not going to be sufficient. Once hypotheses are developed and evaluated in these controlled settings, it will be important to verify that the conclusions scale-up to more realistic situations. But eyewitness researchers must add highly controlled experiments that seek to test theory as a complement to the more realistic experiments that have dominated the literature to date.

Theory development and testing in eyewitness memory will also require consideration of additional dependent variables. Right now, data from eyewitness experiments are sparse, often consisting of only response proportions for suspect IDs, filler IDs, and rejections. Reaction time data play a large role in theory development and testing in the broader cognitive literature (e.g., Ratcliff & Rouder, 1998; Ratcliff & Starns, 2009). There has been some consideration of reaction time data in the eyewitness literature (e.g., Brewer, Caon, Todd, & Weber, 2006), but as a postdictor of eyewitness accuracy and not in the service of theory development. Future theorizing also must account for metacognitive judgments like prospective and retrospective confidence judgments. The need for a better understanding of confidence is clear given in Wixted et al.'s (in press) call for jurors to rely on initial eyewitness confidence. Prospective confidence judgments (do you think you can ID the perpetrator?) might influence which eyewitnesses are, or are not, shown a lineup. In real crimes, eyewitnesses sometimes report to the police that they will not be able to make an ID; perhaps because they did not think they got a good view of the perpetrator, or were a bystander rather than the victim. How accurate are those judgments? Do eyewitnesses that believe that they *cannot* make an ID, but nevertheless are shown a lineup, perform more poorly than those eyewitnesses that believe they can make an ID (and would that be reflected in their level of confidence in that ID)? Finally, the availability of sophisticated neuroscience tools can provide an unparalleled window into cognitive function. There have been efforts to apply these tools to try to separate accurate from inaccurate memories (Rosenfeld, Ben Shakhar, & Ganis, 2012; Schacter, Chamberlain, Gaessar, & Gerlach, 2012). These tools hold great promise for advancing theory, if the data are interpreted in the context of appropriate theoretical frameworks (Wixted & Mickes, 2013).

At the conclusion of Wells, Memon, and Penrod's (2006) overview of eyewitness evidence, they propose that eyewitness researchers have been unadventurous by focusing all their reform efforts on the lineup. Instead, they ask us to consider what researchers might dream up if the lineup never existed.

> *Operating from scratch, it seems likely that modern psychology would have developed radically different ideas. For instance, brain-activity measures, eye movements, rapid displays of faces, reaction times, and other methods for studying memory might have been developed instead of the traditional lineup*
> *Wells et al., p. 69.*

Although we agree that new ideas and new procedures should be tried, it is important that these "radically different ideas" are embedded in appropriate theoretical frameworks.

5.2 Cost and Benefits

New reforms must consider both benefits and costs. But eyewitness researchers must rely on policy makers to decide if it is more important to protect the innocent, implicate the guilty, or whether each is equally important. For example, the recent National Academy of Sciences report (*Identifying the Culprit: Assessing Eyewitness Identification*, October, 2014) recommended adopting unbiased lineup instructions. Given that the data show no discriminability difference between biased and unbiased instructions (see Clark et al., 2014), this recommendation must be based on the fact that the National Academy attaches greater social good to protecting the innocent, which the more conservative responding induced by unbiased instructions accomplishes. We agree with this recommendation, but point out that this is a different justification for the adoption of this reform than what was offered by Wells et al. (2000), and that the recommendation only makes sense if a greater social good is attached to protecting innocent suspects than protecting innocent victims who may suffer at the hands of guilty suspects who are incorrectly freed from suspicion.

Once a determination is made of the relative weight to give to benefits versus costs, SDT can guide researchers in their choice of what reforms are best at achieving the desired goal. In particular, SDT specifies two factors that are vital for evaluating diagnostic domains, and for governing where eyewitnesses place their response criteria (see Clark, 2012, for a review of this issue). One factor is the relative frequency of target-present versus target-absent lineups in the criminal justice system. In other words, how

often do the police put guilty versus innocent suspects into lineups. These base rates are difficult to estimate. We cannot simply assume that if someone selected from a lineup is eventually convicted that they were guilty. The many Innocence Project DNA exonerations disprove that. The base rates also are influenced by when different jurisdictions conduct lineups. Some may choose to conduct a lineup early in an investigation, especially if there is little other evidence to consider. These lineups might contain a relatively high number of innocent suspects. Another jurisdiction may conduct a lineup only after other evidence has created probable cause implicating the suspect (Wells & Olson, 2003). These lineups might have relatively few innocent suspects.

As mentioned above in the context of recommending unbiased instructions, the other factor that influences where an eyewitness places his or her response criterion is the utilities of the various responses that result. For example, if we follow Blackstone's maxim that it is "… better that ten guilty persons escape than that one innocent suffer" (Blackstone, 1769, p. 352), the cost of a false ID is 10x greater than that of a miss, and eyewitnesses should set a conservative criterion (although not as conservative as if the cost of a false ID is 100x greater than a miss, as Benjamin Franklin wrote in 1785). Of course, other policy makers may feel differently (see Volokh, 1997 for a historical and legal review of the many perspectives on the proper ratio of false acquittals to false convictions), as might the general public if the crime is a serious one (de Keijser, de Lange, & van Wilsem, 2014). The important point, however, is that the choice of these utilities is a matter for members of society and their policy makers, not eyewitness researchers. Given that SDT provides the machinery for converting the chosen utilities, given the base rates, into optimal criteria placement, instructions and procedures can be tailored to induce eyewitnesses, and the criminal justice system more broadly, to adopt the optimal criteria placements. That is how new reforms need to be evaluated.

6. CONCLUSIONS

The U.S. Department of Justice document entitled *Eyewitness Evidence: A Guide for Law Enforcement* (Technical Working Group for Eyewitness Evidence, 1999) proposed a set of guidelines for collecting and preserving eyewitness evidence (Wells et al., 2000). The proposed reforms were expected to enhance the accuracy of eyewitness evidence by stipulating

how to conduct an eyewitness lineup. However, the reforms do not enhance the accuracy of eyewitness evidence, at best, they increase eyewitness conservatism. Given the number of innocent people who have been falsely convicted, and the unknown number of innocent people still behind bars due to faulty eyewitness evidence, increased conservatism is important. But that was not the promise of these reforms. The goal of this chapter was to describe how it was that the field originally reached the wrong conclusions regarding many of these reforms.

The chapter began by reviewing the empirical evidence supporting the move to description-matched filler selection, unbiased instructions, sequential lineup presentation, and the discounting of confidence judgments. We discussed four reasons why the field reached incorrect conclusions regarding these reforms. The reasons included a failure to appreciate the distinction between discriminability and response bias, a reliance on summary measures of performance that conflated discriminability and response bias or masked the relationship between confidence and accuracy, the distorting role of relative judgment theory, and a strong focus on preventing the conviction of the innocent. We next reexamined the reforms in light of recent empirical data (exhibiting decline effects) and illustrated the importance of alternative theoretical formulations that can compete with relative judgment theory. A possible new system variable reform was discussed whereby a jury takes the validity of initial eyewitness confidence seriously. However, this, and future system variable reforms, must be motivated and rigorously evaluated in the context of theory.

In hindsight, for all the aforementioned reasons, advocacy on behalf of the sequential lineup and several of the other reforms got ahead of the science. In an article titled "Applying applied research: Selling the sequential line-up," Lindsay (1999, p. 220) wrote: "Obviously the first step in any application of research is to obtain potentially useful data. This is the area in which psychologists excel. We identify potential problems and test possible solutions to those problems." But eyewitness researchers must be careful once they step beyond this point. Lindsay goes on to say, "Once a solution (or at least a superior procedure) has been found and replicated, we feel justified in suggesting that practitioners would benefit from altering their behavior to take advantage of the knowledge generated by our research." At some point, everyone who engages in research on an important topic like eyewitness ID wants his or her research to have an impact. However, requiring that any future reforms are understood theoretically is one way to ensure that advocacy does not get ahead of the science.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation grant SES-1060902 to Scott Gronlund, NSF grant SES-1155248 to John Wixted and Laura Mickes, and NSF grant SES-061183 to Steve Clark. The content is solely the responsibility of the authors and does not necessarily reflect the views of the National Science Foundation.

REFERENCES

Arnold, G. F. (1906). *Psychology applied to legal evidence and other constructions of law*. Calcutta: Thacker, Spink & Co.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*, 81–98.

Barnes, D. M. (1986). Promising results halt trial of anti-AIDS drug. *Science, 234*, 15–16.

Biggers, Neil v. (1972). 409 U.S. 188.

Bjork, R. A. (1973). Why mathematical models? *American Psychologist, 28*, 426–433.

Blackstone, W. (1769). *Commentaries on the Laws of England* (Vol. II) (Book IV).

Bornstein, B. H., & Penrod, S. D. (2008). Hugo Who? G. F. Arnold's alternative early approach to psychology and law. *Applied Cognitive Psychology, 22*, 759–768.

Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. *Law and Human Behavior, 30*, 31–50.

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.

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.

Carlson, C. A., & Gronlund, S. D. (2011). Searching for the sequential line-up advantage: a distinctiveness explanation. *Memory, 19*, 916–929.

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 14*, 118–128.

Charman, S. D., & Wells, G. L. (2007). Eyewitness lineups: is the appearance-change instruction a good idea? *Law and Human Behavior, 31*, 3–22.

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17*, 629–654.

Clark, S. E. (2008). The importance (necessity) of computational modelling for eyewitness identification research. *Applied Cognitive Psychology, 22*, 803–813.

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: psychological science and public policy. *Perspectives on Psychological Science, 7*, 238–259.

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.

Clark, S. E., & Gronlund, S. D. (2015). Mathematical modeling shows that compelling stories do not make for accurate descriptions of data. In J. G. W. Raaijmakers, R. Goldstone, M. Steyvers, A. Criss, & R. M. Nosofsky (Eds.), *Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin*. Psychology Press.

Clark, S. E., Marshall, T. E., & Rosenthal, R. (2009). Lineup administrator influences on eyewitness identification decisions. *Journal of Experimental Psychology: Applied, 15*, 63–75.

Clark, S. E., Moreland, M. B., & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin & Review, 21*, 251–267.

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125*, 159–180.

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). Improving the reliability of eyewitness identification: putting context into context. *Journal of Applied Psychology, 72*, 629–637.

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12*, 41–55.

Dobbins, I. G., Kroll, N. E. A., Yonelinas, A. P., & Liu, Q. (1998). Distinctiveness in recognition and free recall: the role of recollection in the rejection of the familiar. *Journal of Memory and Language, 38*, 381–400.

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: a criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*, 345–357.

Dodson, C.S. & Dobolyi, D.G. Misinterpreting eyewitness expressions of confidence: the featural justification effect. *Law and Human Behavior*, in press.

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology, 67*, 818–835.

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

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.

Farmer, J. J., Jr. (2001). *Attorney general guidelines for preparing and conducting photo and live lineup identification procedures*.

Fife, D., Perry, C., & Gronlund, S. D. (2014). Revisiting absolute and relative judgments in the WITNESS model. *Psychonomic Bulletin & Review, 21*, 479–487.

Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Charles C. Thomas.

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. E. Tulving, & F. I. M. Craik (Eds.), *The oxford handbook of memory* (pp. 229–244). New York, NY: Oxford University Press.

Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.

Goldstein, E. B. (2008). *Cognitive psychology: Connecting mind, research, and everyday experience*. Cengage Learning.

Goodsell, C. A., Gronlund, S. D., & Carlson, C. A. (2010). Exploring the sequential lineup advantage using WITNESS. *Law and Human Behavior, 34*, 445.

Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior, 33*, 70–82.

Greene, E., & Heilbrun, K. (2011). *Wrightsman's psychology and the legal system* (7th ed.). Belmont, CA: Wadsworth.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: Wiley.

Gronlund, S. D. (2004). Sequential lineups: shift in criterion or decision strategy? *Journal of Applied Psychology, 89*, 362–368.

Gronlund, S. D. (2005). Sequential lineup advantage: contributions of distinctiveness and recollection. *Applied Cognitive Psychology, 19*, 23–37.

Gronlund, S. D., Andersen, S. M., & Perry, C. (2013). Presentation methods. In B. Cutler (Ed.), *Reform of eyewitness identification procedures*. APA Publications.

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 15*, 140–152.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S., Wooten, A., et al. (2012). Showups versus lineups: an evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 1*, 221–228.

Gronlund, S. D., Goodsell, C. A., & Andersen, S. M. (2012). Lineup procedures in eyewitness identification. In L. Nadel, & W. Sinnott-Armstrong (Eds.), *Memory and law*. New York: Oxford University Press.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23*, 3–10.

Hansen, M. (2012). Show me your ID: cops and courts update their thinking on using eyewitnesses. *American Bar Association Journal*, 18–19.

Healy, A. F., Fendrich, D. W., Cunningham, T. F., & Till, R. E. (1987). Effects of cuing on short-term retention of order information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 413–425.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley, & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56) Hillsdale, NJ: Erlbaum.

Hunt, R. R. (2003). Two contributions of distinctive processing to accurate memory. *Journal of Memory and Language, 48*, 811–825.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648.

Jewett, D. L. (2005). What's wrong with single hypotheses? Why is it time for strong-inference-PLUS. *Scientist, 19*, 10–11.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28.

Junkin, T. (2004). *Bloodsworth: The true story of the first death row inmate exonerated by DNA*. Chapel Hill, NC: Algonquin Books.

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.

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research: a new survey of the experts. *American Psychologist, 56*, 405–416.

de Keijser, J. W., de Lange, E. G. M., & van Wilsem, J. A. (2014). Wrongful convictions and the blackstone ratio: an empirical analysis of public attitudes. *Punishment & Society, 16*, 32–49.

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1795–1804.

Konecni, V. J., & Ebbesen, E. B. (1979). External validity of research in legal psychology. *Law and Human Behavior, 3*, 39–70.

Krug, K. (2007). The relationship between confidence and accuracy: current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice, 3*, 7–41.

Lacy, J. W., & Stark, C. E. L. (2013). The neuroscience of memory: Implications for the courtroom. *Nature Reviews Neuroscience, 14*, 649–658.

Lane, S. M., & Meissner, C. A. (2008). A "middle road" approach to bridging the basic-applied divide in eyewitness identification research. *Applied Cognitive Psychology, 22*(6), 779–787.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science, 4*, 236–243.
Lindsay, R. C. L. (1999). Applying applied research: selling the sequential line-up. *Applied Cognitive Psychology, 13*, 219–225.
Lindsay, R. C. L., Lea, J. A., & Fulford, J. A. (1991). Sequential lineup presentation: technique matters. *Journal of Applied Psychology, 76*, 741–745.
Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., et al. (1991). Biased lineups: sequential presentation reduces the problem. *Journal of Applied Psychology, 76*, 796–802.
Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: a problem for the match-to-description lineup foil selection strategy. *Law and Human Behavior, 18*, 527–541.
Lindsay, R. C. L., & Wells, G. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4*, 303–313.
Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*, 556–564.
Loftus, E. F. (1979). Malleability of human memory. *American Scientist, 67*, 312–320.
Loftus, E. (2003). Our changeable memories: legal and practical implications. *Nature Reviews Neuroscience, 4*, 231–234.
Ludlum, R. (2005). *The Ambler warning*. New York: St. Martins.
Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15*, 43–57.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66*, 482–489.
Mäntylä, T. (1997). Recollections of faces: remembering differences and knowing similarities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1203–1216.
McCreary, J., Wolf, D., & Forney, A. W. (2009). Unstable. In *Wolf Films, Law & Order: Special victims unit*. Universal City, California: NBC.
McQuiston-Surrett, D., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. simultaneous lineups: a review of methods, data, and theory. *Psychology, Public Policy, and Law, 12*, 137–169.
Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*, 603–616.
Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: a dual-process signal detection theory analysis. *Memory & Cognition, 33*, 783–792.
Memon, A., Hope, L., & Bull, R. H. C. (2003). Exposure duration: effects on eyewitness accuracy and confidence. *British Journal of Psychology, 94*, 339–354.
Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361–376.
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.
Mickes, L. 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*, in press.
Morgan, C. A., Hazlett, G., Doran, A., Garrett, S., Hoyt, G., Thomas, P., et al. (2004). Accuracy of eyewitness memory for persons encountered during exposure to highly intense stress. *International Journal of Law and Psychiatry, 27*, 265–279.

Münsterberg, H. (1908). *On the witness stand*. New York: Doubleday.
Navon, D. (1992). Selection of lineup foils by similarity to suspect is likely to misfire. *Law and Human Behavior, 15*, 43–57.
Neuschatz, J.S., Wetmore, S.A., Key, K., Cash, D., Gronlund, S.D., & Goodsell, C.A. Comprehensive evaluation of showups. In M. Miller & B. Bornstein (Eds.), *Advances in psychology and law*. New York: Springer, in press.
Palmer, M. A., Brewer, N., McKinnon, A. C., & Weber, N. (2010). Phenomenological reports diagnose accuracy of eyewitness identification decisions. *Acta Psychologica, 133*, 137–145.
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.
Pazzaglia, A. M., Dubé, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139*, 1173–1203.
Phillips, M. R., McAuliff, B. D., Kovera, M. B., & Cutler, B. L. (1999). Double-blind photoarray administration as a safeguard against investigator bias. *Journal of Applied Psychology, 84*, 940–951.
Platt, J. R. (1964). Strong inference. *Science, 146*, 347–353.
Popper, K. (1972). *Objective knowledge: An evolutionary approach*. Oxford: Clarendon Press.
Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304–308.
Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences, 109*, 14357–14362.
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*, 347–356.
Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*, 59–83.
Robinson-Riegler, G., & Robinson-Riegler, B. (2004). *Cognitive psychology: Applying the science of the mind*. Allyn and Bacon.
Roediger, H. L. (1996). Memory illusions. *Journal of Memory and Language, 35*, 76–100.
Roediger, H. L., & McDermott, K. B. (2000). Distortions of memory. In F. I. M. Craik, & E. Tulving (Eds.), *The oxford handbook of memory* (pp. 149–164). Oxford, England: Oxford University Press.
Rosenfeld, J. P., Ben Shakhar, G., & Ganis, G. (2012). Physiologically based methods of concealed memory detection. In W. Sinnott-Armstrong, F. Schauer, & L. Nadel (Eds.), *Neuroscience, philosophy and law*. Oxford University Press.
Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Irvington Publishers.
Rotello, C. M. (2001). Recall processes in recognition memory. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 183–221). San Diego, CA: Academic Press.
Rotello, C.M., Heit, E., & Dubé, C. When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, in press.
Rouder, J.N., Province J.M., Swagman A.R., & Thiele J.E. From ROC curves to psychological theory, under review.
Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist, 54*, 182–203.
Schacter, D. L., Chamberlain, J., Gaessar, B., & Gerlach, K. D. (2012). Neuroimaging of true, false, and imaginary memories: findings and implications. In L. Nadel, & W. Sinnott-Armstrong (Eds.), *Memory and law*. New York: Oxford University Press.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49*, 289–318.
Schmechel, R. S., O'Toole, T. P., Easterly, C., & Loftus, E. F. (2006). Beyond the ken: testing juror's understanding of eyewitness reliability evidence. *Jurimetrics Journal, 46*, 177–214.
Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature, 470*, 437.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.
Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: a representative survey of the U.S. population. *PLoS One, 6*(8), e22757.
Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: a meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*, 315–327.
Henderson, State v. Larry R., A-8–08, 062218. (New Jersey Supreme Court, 2011).
Steblay, N. (1997). Social influence in eyewitness recall: a meta-analytic review of lineup instruction effects. *Law and Human Behavior, 21*, 283–297.
Steblay, N. K., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: a meta-analytic comparison. *Law and Human Behavior, 25*, 459–473.
Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: a meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17*, 99–139.
Steblay, N., & Loftus, E. (2013). Eyewitness identification and the legal system. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 145–162). Princeton, NJ: Princeton University Press.
Swets, J. A. (1961). Is there a sensory threshold? *Science, 134*, 168–177.
Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26.
Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: United States Department of Justice, Office of Justice Programs.
Thompson-Cannino, J., Cotton, R., & Torneo, E. (2009). *Picking cotton: Our memoir of injustice and redemption*. New York, NY: St. Martin's Press.
Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: matching suspects or descriptions. *Law and Human Behavior, 24*, 231–258.
Volokh, A. (1997). *N guilty men* (pp. 173–216). University of Pennsylvania Law Review.
Wells, G. L. (1978). Applied eyewitness-testimony research: system variables and estimator variables. *Journal of Personality and Social Psychology, 12*, 1546–1557.
Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14*, 89–103.
Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*, 553–571.
Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360–376.
Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*(3), 776–784.
Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: a successful application of eyewitness research. *American Psychologist, 6*, 581–598.

Wells, G. L., Memon, A., & Penrod, S. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*, 45–75.
Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells, & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). New York, NY: Cambridge University Press.
Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*, 277–295.
Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*, 835–844.
Wells, G. L., Small, M., Penrod, S. J., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: recommendations for lineups and photospreads. *Law and Human Behavior, 22*, 603–647.
Wells, G. L., Steblay, N. K., & Dysart, J. E. (2012). Eyewitness identification reforms: are suggestiveness-induced hits and guesses true hits? *Perspectives on Psychological Science, 7*, 264–271.
Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: the importance of lineup models. *Psychological Bulletin, 99*, 320–329.
Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.
Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*, 1025–1054.
Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7*, 275–278.
Wixted, J. T., & Mickes, L. (2013). On the relationship between fMRI and theories of cognition: the arrow points in both directions. *Perspectives on Psychological Science, 8*, 104–107.
Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*, 262–276.
Wixted, J.T., Mickes, L., Clark, S., Gronlund, S. & Roediger, H. Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, in press.
Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354.