Psychon Bull Rev (2014) 21:251–267
DOI 10.3758/s13423-013-0516-y

THEORETICAL REVIEW

Evolution of the empirical and theoretical foundations of eyewitness identification reform

Steven E. Clark · Molly B. Moreland · Scott D. Gronlund

Published online: 21 November 2013

Abstract Scientists in many disciplines have begun to raise questions about the evolution of research findings over time (Ioannidis in *Epidemiology*, *19*, 640–648, 2008; Jennions & Møller in *Proceedings of the Royal Society, Biological Sciences*, *269*, 43–48, 2002; Mullen, Muellerleile, & Bryan in *Personality and Social Psychology Bulletin*, *27*, 1450–1462, 2001; Schooler in *Nature*, *470*, 437, 2011), since many phenomena exhibit decline effects—reductions in the magnitudes of effect sizes as empirical evidence accumulates. The present article examines empirical and theoretical evolution in eyewitness identification research. For decades, the field has held that there are identification procedures that, if implemented by law enforcement, would increase eyewitness accuracy, either by reducing false identifications, with little or no change in correct identifications, or by increasing correct identifications, with little or no change in false identifications. Despite the durability of this *no-cost* view, it is unambiguously contradicted by data (Clark in *Perspectives on Psychological Science*, *7*, 238–259, 2012a; Clark & Godfrey in *Psychonomic Bulletin & Review*, *16*, 22–42, 2009; Clark, Moreland, & Rush, 2013; Palmer & Brewer in *Law and Human Behavior, 36, 247–255*, 2012), raising questions as to how the no-cost view became well-accepted and endured for so long. Our analyses suggest that (1) seminal studies produced, or were interpreted as having produced, the no-cost pattern of results; (2) a compelling theory was developed that appeared to account for the no-cost pattern; (3) empirical results changed over the years, and subsequent studies did not reliably replicate the no-cost pattern; and (4) the no-cost view survived despite the accumulation of contradictory empirical evidence. Theories of memory that were ruled out by early data now appear to be supported by data, and the theory developed to account for early data now appears to be incorrect.

Keywords Eyewitness memory · Memory · Replicability · Decline effects

"Many hypotheses proposed by scientists as well as non-scientists turn out to be wrong. But science is a self-correcting enterprise. To be accepted, all new ideas must survive rigorous standards of evidence" (Sagan, 1980, p. 73). This article is about the evolution of data and theory in psychological science, how ideas that are later shown to be false become widely held in the first place, and how they are maintained despite the accumulation of disconfirming evidence.

This article focuses specifically on data and theory related to eyewitness identification. There is a consensus among social scientists and legal scholars that mistaken eyewitness identification is a primary cause of wrongful convictions in the U.S. (Garrett, 2011; Gross, Jacoby, Matheson, Montgomery, & Patil, 2005; Gross & Shaffer, 2012). The link between false convictions and false identifications, combined with over 100 years of psychological research on eyewitness memory (Loftus, 1979; Munsterberg, 1908), has led researchers to recommend procedural changes to increase the accuracy of eyewitness identification evidence and reduce the risk of false identification errors that send innocent people to prison. Examples of these recommendations include the following: constructing lineups in such a way that the suspect does not stand out; instructing the witness that the perpetrator of the crime may not be in the lineup; presenting the lineup members sequentially, rather than all at once; and refraining from behaviors that could influence the witness's decision.

Many state and local law enforcement jurisdictions have adopted the new procedures, including the states of Connecticut, New Jersey, North Carolina, Ohio, Texas, and

Electronic supplementary material The online version of this article (doi:10.3758/s13423-013-0516-y) contains supplementary material, which is available to authorized users.

S. E. Clark · M. B. Moreland
University of California, Riverside, Riverside, CA, USA

S. D. Gronlund
University of Oklahoma, Norman, OK, USA

S. E. Clark (✉)
Department of Psychology, University of California, Riverside, Riverside, CA 92521, USA
e-mail: clark@ucr.edu

Wisconsin, as well as Denver, CO, Santa Clara, CA, Suffolk County, MA, and Ramsey and Hennepin Counties, MN. This reform movement has gained considerable momentum in the last few years, "like a runaway train" (Wells, quoted by Hansen, 2012).

This momentum has been driven in part by the claim that the recommended procedures increase identification accuracy—either by reducing the risk of false identifications of the innocent, with little or no loss of correct identifications of the guilty, or by increasing correct identifications of the guilty, with little or no increase in false identifications of the innocent. There are two parts to this claim: first, that the recommended procedures increase overall accuracy, and second, that the increase in accuracy does not involve "mere trade-offs" (Wells, Rydell, & Seelau, 1993, p. 835) of costs and benefits. The pursuit and attainment of this pattern—increased accuracy *with no cost*[1]—has dominated eyewitness identification research for 30 years. The empirical claim that the no-cost pattern is routinely obtained has been widely held, not only for the recommendations discussed in this article, but for others as well (Charman & Wells, 2007; Haw & Fisher, 2004; Wells, 1984), and not only among eyewitness identification researchers (Lindsay, 1999; Wells et al., 1998; Wells, Steblay, & Dysart, 2011), but also among legal scholars (Findley, 2008; Garrett, 2008), policy-makers (Wisconsin Attorney General, 2006), and the popular media (Fenster, 2012; Gawande, 2001; Hart, 2012).

These claims are contradicted by data (Clark, 2005, 2012a; Clark & Godfrey, 2009; Clark, Rush, & Moreland, 2013; McQuiston-Surrett, Malpass, & Tredoux, 2006; Palmer & Brewer, 2012). Procedures that decrease the false identification rate also decrease the correct identification rate, and procedures that increase the correct identification rate also increase the false identification rate. Many of the recommended procedures appear to be no more accurate, or even less accurate, than nonrecommended comparisons.

The relationship between correct and false identification rates is important not only for criminal justice policy, but also for theories of memory and decision making. Specifically, the no-cost pattern showing asymmetrical, independent variation in correct and false identifications is not easily explained by signal detection (Egan, 1958; Wixted, 2007) or matching models of recognition memory (e.g., Clark, 2003; Clark & Gronlund, 1996; Shiffrin & Steyvers, 1997). Such theories, which describe the decision component of recognition memory in terms of a comparison with an adjustable criterion, were dismissed from the eyewitness literature nearly 30 years ago in favor of a theory centered on a distinction between absolute and relative judgments (Wells, 1984). This distinction provided a framework that gathered a wide variety of different procedural reforms under a unified theoretical umbrella. This compelling and intuitive theory has had a profound impact on eyewitness identification research. However, as we will argue, it also may have led researchers to misinterpret empirical results and maintain the no-cost view even as empirical evidence accumulated against it.

This article explores several questions about the evolution of theory and data in eyewitness identification research. How did the no-cost view come to be so widely accepted? How did the no-cost view affect theory development, and how did theory development affect the no-cost view? Did the empirical results change over time? If the results did change, why did the field seem not to notice? There has recently emerged a great deal of interest in—and concern with—how research findings change over time (Ioannidis, 2005, 2008; Jennions & Møller, 2002; Lehrer, 2010; Leimu & Koricheva, 2004; Mullen, Muellerleile, & Bryant, 2001; Schooler, 2011), with some researchers suggesting that many empirical associations may be false or inflated (Ioannidis, 2005, 2008), due to a variety of methodological and data-analytic factors (Simmons, Nelson, & Simonsohn, 2011).

The present article is organized as follows. In the next section, we describe the experimental paradigm for eyewitness identification research and how it maps on to police procedures in actual criminal investigations. The second section describes the recommended procedures and relevant data and how the basic results have evolved over the last 30 years. The third section considers various factors that may have maintained the no-cost view in the face of disconfirming data.

Eyewitness identification in criminal investigations and in experimental research

In a typical lineup identification procedure, the police place the suspect among some number of other individuals, sometimes referred to as *fillers*, who are known to be innocent. The suspect may be guilty or innocent, and the witness may identify the suspect, identify a filler, or identify no one from the lineup. Of these possible responses, suspect identifications have a unique role in the criminal justice system, because they constitute direct evidence of the suspect's guilt. Thus, it is important to distinguish between a false identification of a suspect who is innocent versus an identification of a lineup filler. Both responses are incorrect identifications of a person who is innocent of the crime, but only the false identification of an innocent suspect has the potential to lead to prosecution

[1] There is some variation in how this claim is expressed in the research literature. In some cases, the claim is that there is *no* cost associated with the recommended procedure (Lindsay, 1999), whereas in other cases, the claim is that there is little or no cost. To the extent that costs are sometimes acknowledged, they are often described as uncertain and so small as to be functionally nonexistent (i.e., the correct identification rate "might be slightly harmed"; Wells, Memon, & Penrod, 2006, p. 62). Thus, in the research literature, no cost and little or no cost are not differentiated in any meaningful way; hence, we use the phrase "no cost" here.

and a false conviction. Because the fillers are known to be innocent, an identification of a filler is a known error, and only in the most unusual cases is a filler ever prosecuted after having been identified. Because these distinctions are sometimes confused, we will use the term *correct identification* to refer *only* to the identification of a suspect who is guilty and the term *false identification* to refer *only* to the identification of a suspect who is innocent. As will be shown later, the failure to distinguish between identifications of fillers and false identifications of innocent suspects plays a large role in the misinterpretation of data.

In actual criminal investigations, it can be difficult to determine whether an identification of a suspect is a correct identification of the guilty or a false identification of the innocent. Consequently, eyewitness identification research relies heavily (although not entirely) on an experimental paradigm in which participants become witnesses to a *staged* crime that is either performed live or recorded and presented on video. Such staged crimes cannot fully capture the chaos, emotion, and high stakes of a real criminal investigation. However, they have one critical advantage: The identity of the perpetrator, typically an actor or confederate of the experimenter, is known to a certainty. This allows researchers to clearly examine both possibilities regarding the guilt of the suspect. After witnessing the staged crime, some witnesses are shown a guilty-suspect lineup, whereas others are shown an innocent-suspect lineup.[2]

The responses of experimental witnesses, like those of real witnesses, can be categorized as identifications of the suspect, identifications of a filler, or nonidentifications, and the various response rates can be calculated. Some experiments, however, report only the total identification rate for innocent-suspect lineups, without distinguishing between a filler identification and the false identification of an innocent suspect. For those experiments, the false identification rate can be estimated by dividing the total identification rate by the number of people in the lineup. With this background, we turn our attention to the recommended procedures and the relevant data.

Recommended procedures and relevant data

Eyewitness researchers have made many recommendations for reform. However, because we are interested in how empirical results change over time, we focus on reforms for which there are sufficient data to analyze changes over time. Thus, we consider three[3] recommendations: (1) the use of unbiased instructions, (2) the sequential presentation of lineups, and (3) the selection of more similar fillers who match the witness's description of the perpetrator. The comparison conditions for the first two recommendations are biased instructions and simultaneous presentation of the lineup, respectively. The third recommendation involves two comparison conditions: (1) between fillers who do or do not match the description of the perpetrator and (2) between fillers who match the witness's verbal description of the perpetrator versus fillers who match the visual appearance of the suspect. The reforms are described briefly below.

Instructions to the witness

Researchers have recommended that the lineup administrator instruct the witness that the perpetrator of the crime may or may not be present in the lineup and that the witness is not required to identify anyone. Such instructions are referred to as "unbiased" and are contrasted with what are referred to as "biased" instructions that state or imply that the perpetrator *is* in the lineup and that the witness should identify that person (implying that none-of-the-above is not an appropriate response). Until recently, the results of studies comparing biased and unbiased lineup instructions have been summarized as showing "strong effects" for the reduction of false identifications, with "little if any" effect on correct identifications (Wells & Seelau, 1995, p. 773). More recent summaries of the literature acknowledge that the correct identification rate "might be slightly harmed" by the use of unbiased instructions (Wells, Memon, & Penrod, 2006, p. 62) but that unbiased instructions have "little effect" on correct identifications.

Sequential lineup presentation

The sequential lineup presents lineup members one at a time, and the witness is required to respond "yes" (that's the perpetrator) or "no" (that's not the perpetrator) as each lineup member is presented. The sequential presentation is contrasted with what has been for many years the standard procedure, which is to present all lineup members to the witness simultaneously. Wells et al. (2011) recently summarized the research literature as "decades of laboratory research showing that the sequential procedure reduces mistaken identifications with little or no reduction in accurate identifications" (p. x). This pattern of results has been called the "sequential superiority effect" (Steblay, Dysart, & Wells, 2011).

[2] In the eyewitness identification research literature, the guilty-suspect lineup is often referred to as a perpetrator-present, culprit-present, or target-present lineup, and the innocent-suspect lineup is often referred to as a perpetrator-absent, culprit-absent, or target-absent lineup. These terms do not capture an important aspect of lineups as they are conducted in real criminal investigations. Police lineups always include a suspect; the question is whether that suspect is guilty or innocent.

[3] One reform that we do not consider here is the recommendation to conduct lineups with a blind lineup administrator who does not know the position of the suspect in the lineup. With only one published study, there are insufficient data for analysis.

Description-matched filler selection

The selection of lineup fillers is a balancing act: If the fillers are too dissimilar, even a nonwitness can pick out the suspect; if the fillers are too similar, a witness with an accurate memory of the perpetrator may have difficulty distinguishing the perpetrator from the fillers. Luus and Wells (1991) argued that this balance could best be achieved by selecting fillers who match the witness's description of the perpetrator. The recommendation is based on two lines of research. One line compares lineups with more similar fillers who match a description of the perpetrator with lineups with less similar fillers who mismatch a description of the perpetrator. If the innocent suspect matches the witness's description of the perpetrator but the fillers mismatch the witness's description of the perpetrator, the fillers (but not the suspect) can be ruled out as implausible, leading to high rates of false identification. The other line of research compares lineups with fillers who match a description of the perpetrator with lineups with fillers who match the visual appearance of the suspect. According to Luus and Wells, by selecting fillers on the basis of their similarity to the visual appearance of the suspect, the fillers may be too similar to the guilty suspect, leading to an unnecessary reduction of correct identification rates.

According to the no-cost claims, lineups with fillers who match the description of the perpetrator produce a lower false identification rate with little or no loss of correct identifications, relative to lineups with fillers who mismatch the description of the perpetrator (Lindsay & Wells, 1980), and they produce a higher correct identification rate with no increase in false identifications, relative to lineups composed of fillers who are similar to the visual appearance of the suspect (Luus & Wells, 1991; Wells et al., 1993). In this article, the comparison between matching and not matching to a description of the perpetrator is discussed under the heading of *filler similarity*, and the comparison between matching to description versus matching to suspect is discussed under the heading of *filler selection method*.

Relevant data

The no-cost claims associated with each of these recommendations were assessed through meta-analytic reviews conducted by Clark (2012a) and Clark et al. (2013). The correct and false identification rates for each of the four comparisons, averaged across studies, are shown in Table 1, and the results of the individual studies are given in the Appendix. Each analysis shows a trade-off relationship between correct and false identifications, contrary to the no-cost claims. Unbiased instructions and sequential lineups reduce both the correct and false identification rates, relative to biased instructions and simultaneous lineups. Lineups composed of fillers who match the witness's description of the perpetrator produce decreases in both correct and false identification rates, relative to lineups composed of fillers who mismatch the witness's description of the perpetrator, and produce increases in both correct and false identification rates, relative to lineups composed of fillers selected to match the visual appearance of the suspect.

Table 1 Correct and false identification rates in recommended (R) and nonrecommended (N) eyewitness identification procedures

	Correct	False
Biased instructions (N)	.59	.15
Unbiased instructions (R)	.50	.09
Simultaneous (N)	.54	.15
Sequential (R)	.43	.09
Less similar fillers (N)	.67	.31
More similar fillers (R)	.59	.17
Suspect-matched fillers (N)	.46	.07
Description-matched fillers (R)	.53	.15

In the next section of the article, we discuss the origin of the no-cost view and the evolution of the relevant data over the last 30 years. We then turn to the central question of the article: Why did the no-cost view persist for nearly 30 years, when it is unequivocally contradicted by data?

Origin, evolution, and maintenance of the no-cost claim

Origin: early data and theory

Two studies published in the early 1980s examined the effects of biased instructions and biased lineup composition on eyewitness identification. Malpass and Devine (1981) showed that unbiased instructions resulted in a lower false identification rate and a slightly (but not significantly) higher correct identification rate. Lindsay and Wells (1980) showed that the false identification rate was much lower when lineups were composed with fillers who matched the description of the perpetrator on salient characteristics (ethnicity, facial hair, hair color) than when they were composed with fillers who mismatched on those salient characteristics. The correct identification rate did decrease, but the decrease was not statistically significant and was smaller than the decrease in the false identification rate.

Both papers emphasized the asymmetry in results. Malpass and Devine (1981) noted that "errors were relatively low with the (perpetrator) present, irrespective of the instructions, whereas biased instructions led to a very high error rate with the (perpetrator) absent" (p. 486). Lindsay and Wells (1980) acknowledged the decrease in the correct identification rate with better matching fillers but also noted that the loss of correct identifications was much smaller than the reduction

in false identifications: "Our findings provide the criminal justice system with a reasonable means of improving the reliability of eyewitness testimony at apparently little cost" (p. 310).

Perhaps the key factor in establishing the no-cost view was the development of a decision theory (Wells, 1984) that appeared to account for the asymmetrical results of Lindsay and Wells (1980) and Malpass and Devine (1981) and established the no-cost pattern as a result that not only had been obtained, but also *should* be obtained. Although not explicitly stated, the theory was developed as a necessary alternative to theories of recognition based on signal detection theory and criterion adjustment (see Wells, 1984, p. 93, noting that unbiased instructions "did not simply make witnesses more cautious since accurate identifications were...unaffected"). According to such theories, people make recognition decisions by computing an index of strength or familiarity of the test item. If the index is above some adjustable criterion, they respond "yes," and if that index is below that criterion, they respond "no." A core prediction of such models is that true positives and false positives must covary with adjustments in the decision criterion. This theoretical framework is foundational to all theories of recognition memory (Clark & Gronlund, 1996; Wixted, 2007). The asymmetrical variation in correct and false identifications appears to contradict any account based on a criterion shift.

According to Wells (1984), the results from Malpass and Devine (1981) and Lindsay and Wells (1980) were not due to a shift in the decision criterion but, rather, were explained in terms of a distinction between absolute and relative judgments. According to Wells's theory, an identification based on a relative judgment is one in which the "witness seems to be choosing the lineup member who most resembles the witness's memory *relative* to the other lineup members" (p. 92), whereas an identification based on an absolute judgment requires that the match "must exceed some cut-off or threshold" (p. 95). According to the theory, the relative judgment decision rule is a "useful and unflawed strategy" for guilty-suspect lineups (Wells et al., 1993) but is "fallacious," "dysfunctional," and "dangerous" for innocent-suspect lineups, because "choosing the person who most resembles the criminal is necessarily going to produce an error whenever the true offender is absent from the lineup" (p. 93). The shift from relative to absolute judgments should reduce false identification rates but have little or no effect on correct identification rates.

The sequential lineup was devised as a direct test of this theory. Lindsay and Wells (1985) reasoned that if lineup members were presented sequentially and witnesses were required to make a "yes" (that's him) or "no" (that's not him) decision for each lineup member, the tendency toward making relative judgments would decrease, on the assumption that it is difficult to make comparisons among lineup members presented one at a time. The results showed a substantial decrease in the false identification rate, from .43 in the simultaneous lineup condition to .17 in the sequential lineup condition, and a much smaller decrease in the correct identification rate, from .58 in the simultaneous lineup condition to .50 in the sequential lineup condition. Lindsay and Wells (1985) emphasized this asymmetry in the results, noting a "reduction of inaccurate identifications without loss of accurate identifications" (p. 562).

It is not hard to see how the no-cost view caught on. In 1985, there were several empirical results that showed the no-cost pattern and an intuitive and compelling theory that not only seemed to account for it,[4] but also mandated it as the pattern that *should* be observed. As was noted by Wells et al. (1993), "the goal of finding lineup-identification methods that can reduce false-identification rates without damaging accurate-identification rates has dominated the conceptual and operational approach of eyewitness theorists" (p. 835).

Evolution: changes in empirical results over time

The average correct and false identification rates shown in Table 1 suggest that the results first observed by Malpass and Devine (1981), Lindsay and Wells (1985), Lindsay and Wells (1980), and Wells et al. (1993) have not held up over time. In this section of the article, we examine how and to what extent the results have evolved over time.

We examine the correct and false identification rates, as well as several measures of overall accuracy, based on both correct and false identification rates. We examine the difference in correct identification rates and the difference in false identification rates, the log of the ratio of correct/false ratios, d', and Pearson's r. Some of the performance measures are quite straightforward and require little explanation, whereas others require more explanation.

Differences in correct identification rates, false identification rates, and d'

The differences in correct (C) and false (F) identification rates for recommended (R) and nonrecommended (N) procedures are calculated as $C_R - C_N$ and $F_R - F_N$, respectively. Thus, both of these differences reflect the correct identifications that are lost and the false identifications that are avoided as negative numbers. The d' statistics are calculated from the correct and false

[4] In addition to explaining these three empirical results, Wells (1984) also suggested that the distinction between absolute and relative judgments could account for the relationship between eyewitness confidence and eyewitness accuracy, thought at that time to be very weak, and also account for the results from an experiment in which a "blank lineup" was presented prior to the guilty-suspect lineup.

identification rates, and the difference in d' is given simply as $d'_R - d'_N$, which gives a positive number when the recommended procedure results in superior performance relative to the nonrecommended procedure. The two remaining performance measures, the log of the ratio of C/F ratios and Pearson's r, require lengthier explanations.

Log of the ratio of C/F ratios

The C/F ratio is the accuracy measure most frequently used in the eyewitness identification literature (see Wells & Lindsay, 1980), and it is calculated by simply dividing the correct identification rate by the false identification rate. The interpretation of the ratio is straightforward. If the correct and false identification rates are .5 and .1, the C/F ratio is 5.0, implying that the suspect is 5 times more likely to be identified when guilty than when innocent. Despite this intuitive interpretation of the C/F ratio, it has a number of problems that have been articulated elsewhere (Clark, 2012a; Mickes, Flowe, & Wixted, 2012; Wixted & Mickes, 2012). We use it here, with some adjustments,[5] because of its widespread use in the eyewitness identification literature. (We will have more to say about the C/F ratio later.)

Proportional changes in correct and false identification rates

Pearson's r is used here as a measure of proportional changes in correct and false identification rates.[6] The value of r will be zero if correct, and false identification rates change proportionally across procedures. The value of r will be positive if the false identification rate decreases disproportionally more than the correct identification rates. The value of r will be negative if the decrease in correct identification rates is disproportionally larger than the decrease in false identification rates.

Analyses

To assess the change in the patterns of results, we computed *cumulative effect size* functions (Lau et al., 1992; Leimu & Koricheva, 2004; Mullen et al., 2001). The cumulative effect size averages over all of the studies published up to and including a given year. Because the cumulative effect size is calculated as data accumulate over time, the averages are based on ever increasing numbers of comparisons.

The analyses presented here focus primarily on published studies, for two reasons. First, published studies reflect what would have been known at any given point in time. Second, the U.S. Supreme Court in *Daubert v. Dow Pharmaceuticals* (1993) established peer-reviewed publication as an important consideration for the admissibility of expert testimony. Although all of the data analyzed here are from published studies, some of the critical comparisons were not reported in the original publications. For example, Malpass and Devine (1980) did not analyze the instructions manipulation in their published study. The critical results, comparing biased and unbiased instructions, appeared as an unpublished study by Malpass, Devine, and Bergen (1980) in Steblay's (1997) meta-analysis. However, the data were published by Malpass and Devine (1981b), but without the biased–unbiased comparison. Also, a study by Smith, Lindsay, Pryke, and Dysart (2001) examined performance in simultaneous and sequential lineups; however, the published paper made no reference to the sequential lineups and published only the simultaneous lineup data.[7] In such studies, where the study *was* published but the critical data were analyzed years later, the results are incorporated into the cumulative effect size function for the year the critical analyses were made public.

Cumulative effect functions are shown in Fig. 1 for lineup instructions, lineup presentation, filler similarity, and filler selection method, respectively. Each column consists of five panels, showing cumulative effect size functions for (1) the difference in correct identification rates, $Correct_R - Correct_N$; (2) the difference in false identification rates, $False_R - False_N$; (3) the log of the ratio of C/F ratios, $\log[(C/F)_R/(C/F)_N]$; (4) the difference in d' statistics, $d'_R - d'_N$; and (5) Pearson's r. Each panel shows a scatterplot where each point represents a single comparison, plus two cumulative effect functions, one based on the mean and one based on the median, and the slope of the regression line. Table 2 provides the slope (b) and correlation (r) for each dependent measure, along with the standard

[5] Clark (2012a) compared recommended and nonrecommended procedures as the difference between the two ratios—that is, $C_R/F_R - C_N/F_N$. However, if performance in a given study is low, this difference may underestimate the changes in performance. The point can be illustrated with two hypothetical cases, one in which $C_R/F_R = 3$ and $C_N/F_N = 2$, and one in which $C_R/F_R = 12$ and $C_N/F_N = 11$. The difference in the ratios is 1.0 in both cases, but the ratios of the ratios, 3/2 and 12/11, are very different. The disadvantage of ratio measures is that they are asymmetric and very sensitive to small numbers in the denominator (see Clark, 2012a; Clark, Erickson, & Breneman, 2011). Thus, the log of the ratio is used here.

[6] Pearson's r is calculated from the chi-square ($r = \sqrt{\chi^2/N}$) on the basis of the 2 × 2 table defined by correct and false identification rates for recommended and nonrecommended procedures.

[7] The simultaneous and sequential lineups were indeed part of the same study, since data in both conditions were in the same data file. We thank Steve Smith for sharing the original data file.

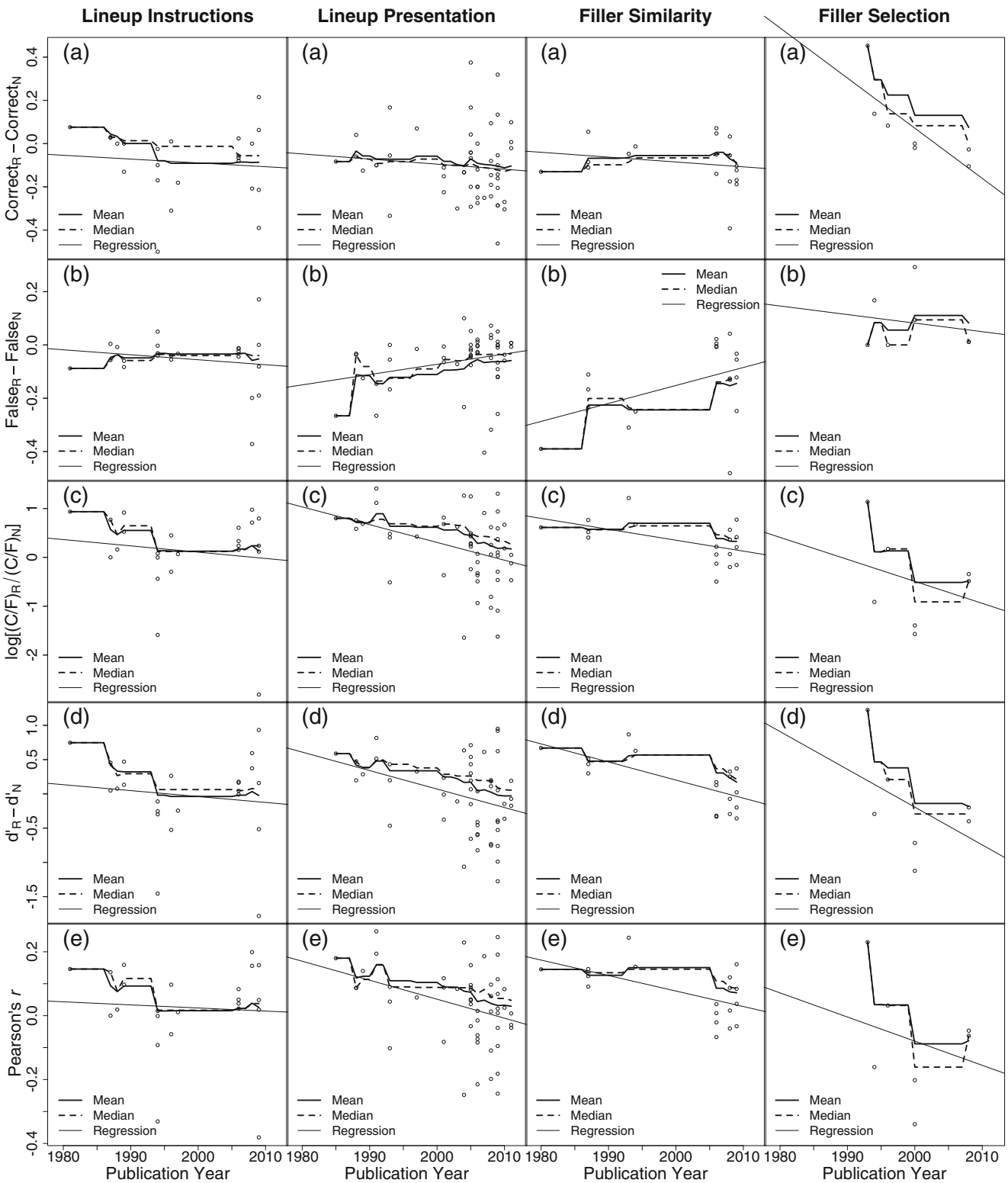

Fig. 1 Cumulative effect size functions for lineup instructions, lineup presentation, filler similarity, and filler selection. **a** Correct identification rate difference. **b** False identification rate difference. **c** Log of the ratio of C/F ratios. **d** d' difference. **e** Pearson's r

Table 2 Regression analyses for lineup instructions, lineup presentation, filler similarity, and filler selection

		Lineup instructions	Lineup presentation	Filler similarity	Filler selection
$Correct_R - Correct_N$	*r*	−.099	−.100	−.201	−.777
	b	−.002	−.002	−.002	−.023
	b(*SE*)	.004	.003	.003	.008
	p	.654	.484	.424	.040
$False_R - False_N$	*r*	−.169	.266	.452	−.178
	b	−.002	.004	.007	−.003
	b(*SE*)	.002	.002	.003	.008
	p	.442	.060	.060	.703
$\log [(C/F)_R/(C/F)_N]$	*r*	−.137	−.359	−.527	−.297
	b	−.013	−.036	−.022	−.045
	b(*SE*)	.020	.014	.009	.065
	p	.532	.010	.025	.518
$d'_R - d'_N$	*r*	−.125	−.346	−.698	−.455
	b	−.009	−.027	−.026	−.055
	b(*SE*)	.015	.010	.007	.048
	p	.569	.013	.001	.305
Pearson's *r*	*r*	−.062	−.341	−.572	−.256
	b	−.001	−.006	−.005	−.008
	b(*SE*)	.003	.002	.002	.013
	p	.779	.014	.013	.579

error of the slope and the *p* value for significance testing.[8]

For the top two rows of the figure, negative values represent decreases in correct and false identification rates. The interpretation of the slope depends on whether the function is moving away from zero, which indicates that the differences have increased over time, or toward zero, which indicates that the differences have decreased over time. Positive values in the bottom three rows of the figure represent the accuracy advantage for the recommended procedure relative to the nonrecommended procedure. Negative slopes represent a decline effect.

Lineup instructions The cumulative effect size functions comparing biased and unbiased instructions are shown in the far left column of Fig. 1. Panels a and b show that the reductions in correct and false identification rates have changed only slightly over time. Panels c, d, and e show that the performance advantage for unbiased instructions has also decreased only slightly over the past 32 years. However, none of the correlations approached statistical significance. A comparison of the mean and median cumulative effect size functions for correct identification rates in panel a may explain the nonsignificance of what appear to be clear trends. The function is much lower for the mean than for the median, suggesting that the mean is pulled downward by a few cases in which the the loss of correct identifications in the unbiased instructions condition was very large, which indeed was the case. Thus, there is little evidence that the effects of unbiased instructions have changed significantly since 1981.

Lineup presentation The cumulative effect size functions comparing sequential and simultaneous lineups are shown in the second column of Fig. 1. The loss of correct identifications due to sequential presentation has increased slightly, but not significantly, over time (panel a, $r = -.10$, $b = -.002$, $p = .48$). The reduction in the false identification rate has decreased over time (panel b, $r = .27$, $b = .003$ $p = .06$). These two effects combine such that the accuracy advantage due to sequential presentation also has decreased significantly over time (for $\log[(C/F)_R/(C/F)_N]$, $r = -.36$, $b = -.036$, $p = .01$; for $d'_R - d'_N$, $r = -.35$, $b = .03$, $p = .01$; for Pearson's *r*, $r = -.34$, $b = -.006$, $p = .01$).[9]

Filler similarity The cumulative effect size functions comparing more similar versus less similar fillers is shown in the third column of Fig. 1. The figure shows (panel a) that the loss of correct identifications due to increased filler similarity has increased only very slightly since Lindsay and Wells's (1980) first study. However, the reduction in false identification rates (panel b) has decreased considerably, $r = .452$, $b = .007$, $p = .06$. These two results combine to produce a

[8] We note two points of clarification. (1) Pearson's *r* is calculated for two purposes. It represents the association between the publication year and effect size, and it also is a measure of the proportional change in correct and false identification rates in the bottom row of Fig. 1. (2) Our calculation of the slope differs from the calculation of the cumulative slope by Mullen et al. (2001; see pp. 1455–1456 for details).

[9] A recent meta-analytical review of simultaneous and sequential lineups, by Steblay et al. (2011), used a different set of studies than those used in the present analysis, which leaves open the question as to whether the pattern reported here is due to the difference in sets of studies. To investigate that possibility, we repeated the analyses based on the 27 "gold standard" simultaneous–sequential comparisons analyzed by Steblay et al. The results, using data from Steblay et al., show the same pattern of results as described above. The details are as follows: correct identification rate, $r = -.15$, $b = -.002$, $p = .45$; false identification rate, $r = .37$, $b = .004$, $p = .055$; log of the C/F ratios, $r = -.58$, $b = -0.05$, $p = .002$; d', $r = -.54$, $b = -.03$, $p = .004$; Pearson's *r*, $r = -.60$, $b = -.008$, $p = .001$.

decrease in the accuracy advantage due to the use of more similar fillers (for $\log[(C/F)_R/(C/F)_N]$, $r = -.53$, $b = -.02$, $p = .03$; for $d'_R - d'_N$, $r = -.70$, $b = -.03$, $p = .001$; and for Pearson's r, $r = -.57$, $b = -.005$, $p = .01$).

Filler selection method The predicted pattern of results, comparing lineups with description-matched fillers with lineups with suspect-matched fillers, differs from the other three comparisons. Luus and Wells (1991)predicted a no cost pattern, but in an opposite direction—specifically, that description-matched filler selection would increase the correct identification rate, with no increase in false identifications. The far-right column of Fig. 1 shows that the increase in correct identification rates has decreased, $r = -.78$, $b = -.023$, $p = .04$, and the difference in false identifications has changed slightly, $r = -.178$, $b = -.003$, $p = .70$. The combination of these two results is a decrease in the accuracy advantage for description-matched lineups (for $\log[(C/F)_R/(C/F)_N]$, $r = -.30$, $b = -.05$, $p = .52$; for $d'_R - d'_N$, $r = -.46$, $b = -.06$, $p = .31$; and for Pearson's r, $r = -.26$, $b = -.008$, $p = .58$). The correlations are not close to statistical significance, although the effect sizes are of moderate size and comparable to the other analyses. The combination of moderate effect sizes and statistical nonsignificance is due to the small number of comparisons ($n = 7$). The separation of the mean and median suggest the presence of an outlier in the set of studies, and indeed there is; the initial study by Wells et al. (1993) showed a very large advantage for description-matched filler selection over suspect-matched filler selection, which was not shown in any of the subsequent studies. The large accuracy advantage reported by Wells et al. (1993) has evolved into a small accuracy disadvantage.

Summary Although there is some statistical variation, the analyses show a common pattern. In each case, the first published study is unrepresentative of the corpus of studies, as is illustrated in Tables 3 and 4. Table 3 shows the effect sizes for the first published study and for the meta-analytic aggregate (means). In each case, the first published study showed an impressive accuracy advantage for the recommended procedure that declined, disappeared, or reversed in the aggregate. Table 4 shows the z-scores for each of the four seminal studies, for each dependent measure, along with the proportion (*pr*) of cases within each normalized distribution that are less extreme than the results of the first published study. If the seminal studies are representative of the full set of studies, these *pr* values should fall near .5. The observed *pr* values, however, range from .55 to .98 across all dependent measures and from .75 to .98 for the accuracy measures.

Importantly, for sequential lineups and filler similarity, the declines in the accuracy advantages did not come about because subsequent studies showed greater costs but, rather, because subsequent studies showed smaller benefits and a less favorable cost/benefit ratio, relative to the initial studies. Thus, for these comparisons, it is not the case that the no-cost pattern was shown in the initial studies and lost in subsequent studies. Rather, the no-cost pattern, although often described in the initial studies, was not shown.

Maintenance of the increased accuracy with no-cost view

The claim that the recommended identification procedures increase identification accuracy—by reducing false identifications with little or no loss of correct identifications or by increasing correct identifications with little or no increase in false identifications—is unambiguously contradicted by data. So why has this view been widely held for the last 30 years? The answer to that question involves some speculation; one cannot establish causal relationships. With that caveat, we make the following observations and arguments.

The same combination of data and theory that gave rise to the no-cost view may have played an important role in maintaining it. The seminal studies, which showed significant benefits and nonsignificant costs, combined with a theory that appeared to predict precisely that pattern, created a compelling and self-reinforcing scientific story that has dominated the field for 30 years. Indeed, as Wells et al. (2006) have noted, the theory, and its predicted no-cost pattern of results, "has permeated the literature on lineups" (p. 61). Also, for each of the four empirical comparisons, the first published study, although unrepresentative of the empirical results taken as a whole (see Tables 3 and 4), is cited more often than any other within its domain.[10]

This early marriage of theory and data may have created a primacy effect within the eyewitness literature. The importance of this arises from an extensive literature showing that strong beliefs, once established, persist and are resistant to change, even in the face of disconfirming empirical evidence (see Nickerson, 1998, for an

[10] We note two caveats regarding this point. First, one should be cautious regarding conclusions about the impact of an article based on the number of times it is cited, since papers may be cited for various reasons, incorrectly cited, or cited for the purpose of criticism. Second, one may ask whether the high number of citations is due to having more years in the literature. On that point, we conducted a year-by-year analysis of citations that showed that citation rates were either steady or increasing over the years.

Table 3 Effect sizes for the first published study and aggregate, for lineup instructions, lineup presentation, filler similarity, and filler selection

		$Correct_R - Correct_N$	$False_R - False_N$	$\log [(C/F)_R/(C/F)_N]$	$d'_R - d'_N$	Pearson's *r*
Lineup instructions	Malpass and Devine (1981)	.076	−.088	0.935	0.747	.146
	Aggregate	−.086	−.053	0.128	−0.024	.026
Lineup presentation	Lindsay and Wells (1985)	−.083	−.266	0.799	0.588	.180
	Aggregate	−.104	−.058	0.171	−0.029	.030
Filler similarity	Lindsay and Wells (1980)	−.130	−.390	0.613	0.669	.145
	Aggregate	−.088	−.145	0.326	0.172	.072
Filler selection	Wells, Rydell, and Seelau (1993)	.453	.000	1.137	1.225	.230
	Aggregate	.075	.082	−0.485	−0.185	−.079

extensive review of primacy effects and belief persistence and of confirmation biases more generally). Yet the question remains as to *how* the no-cost view has persisted. This persistence, we argue, was facilitated by a number of data-analytic confusions,[11] as well as a possible publication bias that kept disconfirming evidence out of the journals. These data-analytic and publication factors are considered in turn.

Nonsignificant versus nonexistent effects

In some writings, the losses of correct identifications associated with recommended procedures are correctly described as statistically nonsignificant. However, in other writings, nonsignificant effects are described as nonexistent—that is, that there was "no change in the rate of accurate identifications" (Lindsay & Wells, 1985, p. 562), that "correct identifications were unaffected" (Cutler & Penrod, 1995, p. 282), that the sequential lineup "did not reduce correct identifications" (Lindsay, 1999), and that "fitting distractors to the verbal description . . . does not interfere with recognition of the culprit" (Wells et al., 1998, p. 637). Perhaps the broadest statement of the no-cost claim was articulated in a white paper commissioned by the American Psychology and Law Society: "We have taken great care to recommend procedures that do not serve to reduce the chances that the guilty party will be identified" (Wells et al., 1998, p. 637). The inconsistency, confusion, and, in some cases, the equating of nonsignificant and nonexistent effects is, of course, not specific to eyewitness research but, rather, has a long history in statistics and psychology (Fisher, 1935; Neyman, 1950; and see Gigerenzer, 1993).

Reliance on a biased measure of accuracy

Accuracy in eyewitness identification experiments has typically been calculated in one way: as the ratio of correct to false identifications—that is, C/F. The C/F ratio, however, is a biased measure of accuracy. Clark et al. (2011) have shown, on the basis of simulations of the WITNESS model (Clark, 2003), that C/F increases when the accuracy of memory is held constant and only the decision criterion is raised. The reason is that C/F increases as the overall identification rate decreases (and the denominator goes toward zero). The same thing happens empirically as a result of increasing response confidence (Gronlund et al., 2012; Mickes et al., 2012). Thus, C/F will give preference to more conservative identification procedures, even if they are not more accurate. Unbiased instructions, sequential presentation, and the use of more similar fillers all produce more conservative responding.

Apples to oranges comparisons

In some cases, the no-cost conclusion was based on comparisons across different responses. Specifically, correct suspect identifications in guilty-suspect lineups have been compared with *all* identifications in innocent-suspect lineups. For example, Steblay (1997) reported, in her meta-analysis of lineup instruction studies, that the "mean accuracy rate for unbiased versus biased lineups (referring to the lineup instructions) in the [innocent-suspect] lineup condition was 60 % versus 35 %" (p. 289). This is a substantial decrease in correct rejections, but it does not accurately reflect the increase in risk to an innocent suspect. The complement of the correct rejection rate is, of course, the total identification rate from innocent-suspect lineups, which includes identifications

[11] One data-analytic issue has been addressed elsewhere (Clark, 2005) and will not be repeated here. In some studies, correct identification rates did not vary across conditions, due to ceiling effects. For example, if the identification rate is already high in the unbiased instructions condition, it has little room to increase in the biased instructions condition.

Table 4 Representativeness of seminal studies

	Lineup instructions Malpass and Devine (1980)		Lineup presentation Lindsay and Wells (1985)		Filler similarity Lindsay and Wells (1980)		Filler selection Wells et al. (1993)	
	z	*pr*	*z*	*pr*	*z*	*pr*	*z*	*pr*
Correct ID difference	1.00	.84	0.13	.55	0.38	.65	2.05	.98
False ID difference	0.35	.64	2.03	.98	1.65	.95	0.74	.77
Log of C/F ratios	0.95	.83	0.88	.81	0.68	.75	1.73	.96
d' difference	1.24	.89	1.13	.87	1.32	.91	1.88	.97
Pearson's *r*	0.85	.80	1.23	.89	0.88	.81	1.69	.95

Note. *pr* values represent the proportion of scores in each normalized distribution that are less extreme than that reported in each seminal study

of fillers who are known to be innocent. To estimate the false identification rate from these numbers, one has to divide the total identification rates by the number of lineup members. Thus, in Steblay's analysis, the 60 % and 35 % correct rejection rates translate (assuming six-person lineups) into false identification rates of (1 − .6)/6 = .07 and (1 − .35)/6 = .11.

This apples–oranges comparison is also woven into the meta-analysis of simultaneous and sequential lineups by Steblay et al. (2011), which reported an 8 % reduction in correct identifications and a 22 % reduction in "errors" in perpetrator-absent lineups. The comparison gives the impression that the benefits far outweigh the costs. However, the 22 % reduction refers to all errors, which includes not only false identifications of the innocent suspect, but also filler identifications. The 22 % reduction in error, when adjusted for the size of the lineup, is less than 4 % (as compared with an 8 % reduction in the correct identification rate).

Evidence for selective publication

Although the present article has focused on published studies, there is some evidence from the meta-analysis by Steblay et al. (2011) that studies that did not show a sequential lineup accuracy advantage had been filtered out of the publication process. Steblay et al. reported the effect sizes (Pearson's *r*) separately for guilty-suspect and innocent-suspect lineups, including those for several unpublished studies. For these 12 unpublished studies, the average effect size representing the loss of correct identifications ($r = -.17$) was only slightly smaller than the average effect size representing the reduction of false identifications ($r = .19$). Note that the sign on the *r* is negative for costs (correct identifications lost) and positive for benefits (false identifications avoided). The sum of the effect sizes (costs and benefits) for the *z*-transformed *r*s was not statistically different from zero, $t\ (11) = 0.51$, $p = .62$. We compared these results with those of studies that were published during the same period of time. Because we analyzed the published results differently than Steblay et al. did, we need to compare *their* unpublished effect sizes with *their* published effect sizes ($r = -.10$ for guilty-suspect lineups and $r = .25$ for innocent-suspect lineups), $t(22) = 3.17$, $p = .004$.[12] The unpublished studies signaled a trade-off between correct and false identifications, in contrast to the published studies.

Implications for Psychology and Public Policy

The analyses by Clark (2012a) and Clark et al. (2013), combined with the analyses presented here, suggest that eyewitness identification research has, for 30 years, perpetuated an empirical generalization that is contradicted by the same data that, for years, were thought to support it. In the remainder of the article, we briefly discuss the implications of these findings for theories of memory, for psychological science, and for public policy.

Theories of recognition memory and eyewitness identification

To the extent that procedural changes induce criterion shifts, one should expect covariation in correct and false identification rates. To the extent that an increase in the similarity of the fillers pulls choices away from an innocent suspect, that increase in similarity also should pull choices away from a guilty suspect.

The seminal studies that showed (or appeared to show) asymmetric variation in correct and false identification rates

[12] The set of published studies includes results that Steblay et al. (2011) excluded from their analysis for various reasons. Other rules for determining the comparison set of published studies produce the same pattern of results.

led to the premature demise of criterion-shift accounts of signal detection models for eyewitness identification. However, most eyewitness identification experiments do not produce the asymmetric no-cost pattern and are generally consistent with criterion-shift accounts. Indeed, the WITNESS model (Clark, 2003), which is a variant of signal detection matching models of recognition (Clark & Gronlund, 1996), has been fit to a wide range of data from eyewitness identification experiments (Clare & Lewandowsky, 2004; Clark, 2003; Goodsell et al., 2010). The attachment to the no-cost view led researchers to reject a theory that has now been shown to account for a wide range of data (signal detection theories of memory, including, but not limited to, the WITNESS model) in favor of a theory that does not (Wells, 1984, 1993). We should be clear in noting that the analyses presented here do not bear on the conceptual distinction between absolute and relative judgments. The analyses do, however, challenge the claim that shifts from relative to absolute judgments reduce false identifications with little or no loss of correct identifications, a claim that does not, in fact, follow from the absolute–relative distinction (see Clark et al., 2011; Goodsell et al., 2010).

Broader implications for psychological science

Researchers have recently expressed serious concerns regarding problems of replicability in psychological science (Simmons et al., 2011), and some have even described the problem as a "crisis" (Pashler & Harris, 2012; Pashler & Wagenmakers, 2012). In contrast to a replicability crisis, most sequential–simultaneous lineup comparisons, higher–lower similarity comparisons, and biased–unbiased instructions comparisons consistently show a reduction in both correct and false identification rates. However, if there is a crisis here, it arises from three sources: the inconsistency in the magnitude of the effects, particularly for changes in the false identification rates; the misinterpretation of changes in correct identification rates; and the unexplained variation in the *relative* magnitudes of the effects that underlie conclusions about accuracy and diagnosticity.

Related to the problem of replicability is the problem of selective publication, which may serve to keep "undead" theories alive by burying results, rather than the theories they falsify (Ferguson & Heene, 2012). This problem likely extends beyond simultaneous and sequential lineups.[13] We did not consider the recommendation that lineups be conducted by a blind lineup administrator, in part because there is only one published study with guilty- and innocent-suspect lineups (Greathouse & Kovera, 2009) and several unpublished studies (see Clark, 2012b).

Implications for criminal justice reform and public policy

The public policy implications of the no-cost pattern—and its disappearance—have been discussed at length elsewhere (see Clark, 2012a, 2012b; Laudan, 2012; Newman & Loftus, 2012; Wells et al., 2011). We will not repeat those arguments and counterarguments here. We note two important points, however. (1) None of the analyses presented here or elsewhere (i.e., Clark, 2012a) should be taken as "smoking gun" evidence to imply that the recommended procedures should not be implemented. Policy decisions about eyewitness identification should ultimately be left to policy-makers, who may appropriately consider normative and legal factors beyond correct and false identification rates. (2) There is an important role, not only for empirical data, but also for formal models of memory and decision making, in these policy considerations. If public policy is to be based on psychological science, it should be informed not only by empirical data, but also by rigorous theory development and evaluation (Clark, 2008; McQuiston-Surrett et al., 2006).

Final remarks

The present article examined the evolution of theory and data over time, showing how early, unrepresentative results and a compelling theoretical account of those results may have combined to establish and perpetuate an empirical generalization that, years later, has been shown to be false. In closing, it is important to not lose sight of the fact that over the same period of time during which empirical results were changing, thousands of innocent people were falsely identified by eyewitnesses, and many of those people were convicted of crimes that they did not commit. Efforts to develop and implement procedures to reduce the risk of false identification errors and false convictions constitute an appropriate and important application of psychological science to a real-world problem.

[13] A recent analysis by Francis (2012) suggests a publication bias in the research on verbal overshadowing (Schooler & Engstler-Schooler, 1990). Verbal overshadowing describes a decrease in identification accuracy for witnesses who previously generated a verbal description of the target (i.e., the perpetrator).

Acknowledgments The authors thank Roddy Roediger and several anonymous reviewers for their very helpful comments. The research was supported by a grant from the National Science Foundation, SES 1051183, and by the Presley Center for Crime and Justice Studies at the University of California, Riverside.

Appendix

The correct and false identification rates for all of the analyses are given in Tables 5, 6, 7, and 8, for biased and unbiased instructions, simultaneous and sequential lineups, more and less similar fillers, and suspect-matched and description-matched filler selection methods. The complete citations for studies in the analyses can be found as supplemental material. The correct and false identification rates provide the raw data from which the measures of accuracy (log of the C/F ratios, and *d'*) and accuracy change (Pearson's *r*) can be calculated.

The studies are restricted to those with adult participants, excluding child witnesses and older witnesses, and were published between 1980 and 2011. Previous analyses of these data were conducted by Clark (2012a) and Clark et al. (2013). The aggregation of data was based on a random effects design (Hedges & Vevea, 1998), in which each comparison constitutes the unit of analysis. Comparisons made across different levels of another independent variable in a given experiment were entered separately into the analyses. Thus, there was no within-study aggregation prior to the across-study aggregation.

Table 5 Correct and false identification rates, biased and unbiased instructions

Study	Publication year	Correct ID (Biased)	False ID (Biased)	Correct ID (Unbiased)	False ID (Unbiased)
Malpass & Devine	1981	.75	.16	.83	.07
Cutler, Penrod, & Martens	1987	.43	.11	.46	.06
Fleet et al.	1987	.63	.09	.65	.10
Cutler & Penrod	1988	.78	.05	.78	.05
Paley & Geiselman	1989	.53	.15	.40	.07
O'Rourke et al.	1989	.36	.10	.36	.04
Foster et al. Women, no consequence	1994	.66	.16	.49	.12
Foster et al. Women, consequence	1994	.65	.14	.55	.11
Foster et al. Men, no consequence	1994	.62	.14	.13	.13
Foster et al. Men, consequence	1994	.44	.11	.42	.16
Devenport & Fisher No authority	1996	.28	.12	.29	.08
Devenport & Fisher Authority	1996	.55	.13	.24	.08
Malpass and Devine (1980)	1997	.64	.10	.46	.07
Brewer & Wells High-similarity thief	2006	.43	.05	.37	.04
Brewer & Wells Low-similarity thief	2006	.38	.05	.30	.03
Brewer & Wells High-similarity waiter	2006	.55	.08	.58	.06
Brewer & Wells Low-similarity waiter	2006	.69	.09	.63	.04
Leippe et al. Low similarity	2009	.88	.52	.67	.15
Leippe et al. High similarity	2009	.70	.39	.70	.19
Greathouse & Kovera Double blind	2009	.64	.02	.43	.19
Greathouse & Kovera Double blind	2009	.50	.07	.56	.07
Greathouse & Kovera Single blind	2009	.86	.33	.47	.14
Greathouse & Kovera Single blind	2009	.57	.21	.79	.13

Table 6 Correct and false identification rates, simultaneous and sequential lineups

Study	Publication year	Correct ID simultaneous	False ID simultaneous	Correct ID sequential	False ID sequential
Lindsay & Wells	1985	.58	.43	.50	.17
Cutler & Penrod, Exp. 1	1988	.76	.07	.80	.03
Cutler & Penrod, Exp. 2	1988	.47	.07	.41	.04
Melara et al.	1989	.25	.17	.13	.04
Lindsay, Lea, & Fulford, Exp. 2	1991	.57	.33	.47	.07

Table 6 (continued)

Study	Publication year	Correct ID simultaneous	False ID simultaneous	Correct ID sequential	False ID sequential
Lindsay, Lea, Nosworthy, et al.	1991	.57	.20	.47	.05
Parker & Ryan	1993	.42	.25	.08	.08
Parker & Ryan	1993	.33	.02	.50	.02
Sporer (1993)	1993	.44	.12	.39	.07
Lindsay et al. (1997)	1997	.55	.06	.62	.04
Kneller et al. (2001)	2001	.61	.10	.50	.04
Smith et al., same race	2001	.46	.07	.23	.02
Smith et al., cross race	2001	.45	.15	.30	.15
Memon & Gabbert (2003)	2003	.47	.09	.17	.02
Haw and Fisher, high contact	2004	.63	.30	.50	.07
Haw and Fisher, low contact	2004	.60	.03	.47	.13
MacLin et al.	2005	.40	.11	.33	.07
MacLin et al.	2005	.47	.08	.27	.04
Rose et al.	2005	.75	.06	.46	.05
Wilcock et al.	2005	.67	.10	.63	.03
Clark & Davey (2005) Exp. 1 (next best pos. 2)	2005	.25	.07	.29	.05
Clark & Davey (2005) Exp. 1 (next-best pos. 4)	2005	.25	.08	.63	.14
Clark & Davey (2005) Exp. 2 (next-best pos. 2)	2005	.33	.09	.29	.05
Clark & Davey (2005) Exp. 2 (next-best pos. 4)	2005	.50	.08	.67	.08
Pozzulo & Marciniak, Appearance change	2006	.23	.07	.23	.07
Pozzulo & Marciniak, No appearance change	2006	.67	.08	.47	.10
Douglass & McQuiston	2006	.88	.16	.63	.15
Wells & Pozzulo (assailant)	2006	.24	.07	.12	.09
Wells & Pozzulo (accomplice)	2006	.40	.09	.20	.07
Levi	2006	.63	.08	.35	.05
MacLin & Phelan	2007	.48	.50	.23	.10
Carlson et al.. Exp. 2 Biased lineup	2008	.71	.64	.46	.33
Carlson et al. Exp. 2 Intermediate lineup	2008	.43	.30	.24	.38
Carlson et al. Exp 2 Fair lineups	2008	.31	.16	.41	.20
Carlson et al. Exp. 1	2008	.72	.02	.57	.05
Pozzulo et al..	2008	.48	.09	.40	.04
Greathouse & Kovera Blind/biased ins.	2009	.43	.19	.56	.07
Greathouse & Kovera Blind/unbiased ins.	2009	.64	.02	.50	.07
Greathouse & Kovera Nonblind/unbiased ins.	2009	.47	.14	.79	.13
Greathouse & Kovera Nonblind/biased ins.	2009	.86	.33	.57	.21
Gronlund et al. Susp. pos. 2 biased lineup	2009	.86	.29	.65	.20
Gronlund et al. Susp. pos. 2, mixed lineup	2009	.72	.20	.26	.21
Gronlund et al. Susp. pos. 2. fair lineup	2009	.76	.17	.47	.17
Gronlund et al. Susp. pos. 5, biased lineup	2009	.81	.41	.76	.15
Gronlund et al. Susp. pos. 5, mixed lineup	2009	.63	.04	.47	.04
Gronlund et al. Susp pos. 5 fair lineup	2009	.69	.16	.59	.09
Steblay & Philips No “don’t know” option	2010	.69	.08	.42	.04
Steblay & Philips “Don’t know” option	2010	.60	.08	.30	.02
Steblay et al. Table 1	2011	.10	.06	.08	.05
Steblay et al. Table 2	2011	.59	.01	.60	.02
Steblay et al. Table 3	2011	.52	.05	.62	.06

Note. *Pos.* position; *ins.* instructions; *susp.* suspect

Table 7 Correct and false identification rates, lineups with less similar or more similar fillers

Study	Publication year	Correct ID less similar fillers	False ID less similar fillers	Correct ID more similar fillers	False ID more similar fillers
Lindsay & Wells	1980	.71	.70	.58	.31
Lindsay et al. Exp. 1	1987	.78	.28	.67	.11
Lindsay et al. Exp. 2	1987	.78	.39	.83	.28
Lindsay et al. Exp. 3	1987	.53	.47	.44	.24
Wells et al.	1993	.71	.43	.67	.12
Lindsay, Martin, & Webber	1994	.81	.50	.79	.25
Brewer & Wells Unb. Ins. Thief	2006	.30	.03	.37	.04
Brewer & Wells Bias. Ins. Thief	2006	.38	.05	.43	.05
Brewer & Wells Unb. Ins. Waiter	2006	.63	.04	.58	.06
Brewer & Wells Bias. Ins. Waiter	2006	.69	.09	.55	.08
Carlson et al.	2008	.71	.64	.31	.16
Carlson et al.	2008	.46	.33	.41	.20
Leippe et al. Biased ins.	2009	.88	.52	.70	.39
Leippe et al. Unbiased ins.	2009	.67	.15	.70	.19
Gronlund et al. Sim, Susp pos. 2	2009	.86	.29	.76	.17
Gronlund et al. Sim, Susp pos. 5	2009	.81	.41	.69	.16
Gronlund et al. Seq., Susp pos. 2	2009	.66	.20	.47	.17
Gronlund et al. Seq. Susp pos. 5	2009	.76	.15	.59	.09

Note. *Unb.* unbiased; *ins.* instructions; *bias.* biased; *sim.* simultaneous; *seq.* sequential; *pos.* position

Table 8 Correct and false identification rates, suspect-matched versus description-matched foil selection

Study	Publication year	Correct ID Susp match	False ID Susp match	Correct ID Desc. match	False ID Desc. match
Wells et al.	1993	.21	.12	.67	.12
Lindsay et al.	1994	.66	.08	.79	.25
Juslin et al.	1996	.44	.09	.52	.09
Tunnicliff & Clark Exp. 1	2000	.53	.03	.53	.13
Tunnicliff & Clark Exp. 2	2000	.33	.08	.31	.38
Darling et al. Still images	2008	.58	.04	.48	.05
Darling et al. Moving images	2008	.44	.03	.42	.05

References

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 739–755. doi:10.1037/0278-7393.30.4.739

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. doi:10.1002/acp.891

Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior, 29,* 395–424. doi:10.1007/s10979-005-5690-7

Clark, S.E. (2008). The importance (necessity) of computational modeling for eyewitness identification research. *Applied Cognitive Psychology, 22*, 803–813. doi:10.1002/acp.1484

Clark, S. E. (2012a). Costs and benefits in eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238–259. doi:10.1177/174569161243958

Clark, S. E. (2012b). Eyewitness identification reform: Data, theory, and due process. *Perspectives on Psychological Science, 7,* 279–283. doi:10.1177/1745691612444136

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

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review, 16,* 22–42. doi:10.3758/PBR.16.1.22

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and Review, 3,* 37–60. doi:10.3758/BF03210740

Clark, S. E., Rush, R. A., & Moreland, M. B. (2013). Constructing the lineup: Law, reform, theory, and data. In B. Cutler (Ed.), *Reform of eyewitness identification procedures* (pp. 87–112). Washington, D.C.: American Psychological Association.

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology, and the law*. New York: Cambridge University Press.

Daubert v. Merrell Dow Pharmaceuticals (1993). 509 U.S. 579.

Egan, J. P. (1958). *Recognition memory and the operating characteristic*. Bloomington, IN: USAF Operational Applications Laboratory (Technical Note No. 58–51).

Fenster, J. (2012). Bill aims to reduce number of false identifications in police lineups. *New Haven Register*, March 17, 2012.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7,* 555–561. doi:10.1177/1745691612459059

Findley, K. A. (2008). Toward a new paradigm of criminal justice: How the innocence movement merges crime control and due process. *Texas Tech Law Review, 41,* 1–41.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin and Review, 19,* 151–156. doi:10.3758/s13423-012-0227-9

Garrett, B. L. (2008). Judging innocence. *Columbia Law Review, 108,* 55–142.

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

Gawande, A. (2001, January 8). Under suspicion. *The New Yorker*, pp. 50–53.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale NJ: Erlbaum.

Goodsell, C.A., Gronlund, S.D., & Carlson, C.A. (2010). Exploring the sequential lineup advantage using WITNESS. *Law and Human Behavior, 34*, 445–459. doi:10.1007/s10979-009-9215-7

Greathouse, S.M., & Kovera, M.B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. Law & *Human Behavior, 33*, 70–82. doi:10.1007/s10979-008-9136-x

Gronlund, S.D., Carlson, C.A., Neuschatz, J.S., Goodsell, C.A., Wetmore, S.A., Wooten, A., & Graham, J. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 1*, 221–228. doi:10.1016/j.jarmac.2012.09.003

Gross, S. R., Jacoby, K., Matheson, D. J., Montgomery, N., & Patil, S. (2005). Exonerations in the United States 1989 through 2003. *Journal of Criminal Law and Criminology, 95,* 523–560.

Gross, S.R., & Shaffer, M. (2012). Exonerations in the United States, 1989–2012: Report by the National Registry of Exonerations.

Hart, P. (2012). Can't count on eyewitnesses to ID criminals. *Houston Chronicle*, May 22, 2012.

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

Haw, R. M., & Fisher, R. P. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *Journal of Applied Psychology, 89,* 1106–1112. doi:10.1037/0021-9010.89.6.1106

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. doi:10.1037/1082-989X.3.4.486

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* 696–701. doi:10.1371/journal.pmed.0020124

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

Jennions, M. D., & Møller, A. P. (2002). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society, Biological Sciences, 269,* 43–48.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine, 327,* 248–254.

Laudan, L. (2012). Eyewitness identifications: One more lesson on the costs of excluding relevant evidence. *Perspectives on Psychological Science, 7,* 272–274. doi:10.1177/1745691612443065

Lehrer. J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? *The New Yorker*, pp. 52–57.

Leimu, R., & Koricheva, J. (2004). Cumulative meta-analysis: A new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society, Biological Sciences, 271,* 1961–1966. doi:10.1098/rspb.2008.2828

Lindsay, R. C. L. (1999). Applying applied research: Selling the sequential lineup. *Applied Cognitive Psychology, 13,* 219–225. doi:10.1002/(SICI)1099-0720(199906)13:3<219::AID-ACP562>3.0.CO;2-H

Lindsay, R. C., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law & Human Behavior, 4,* 303–313. doi:10.1007/BF01040622

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. doi:10.1037/0021-9010.70.3.556

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15,* 43–57. doi:10.1007/BF01044829

Malpass, R. S., & Devine, P. G. (1980). Realism and eyewitness identification research. *Law and Human Behavior, 4,* 347–358. doi:10.1007/BF01040626

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66,* 482–489. doi:10.1037/0021-9010.66.4.482

Malpass, R.S., Devine, P.G., & Bergen, G.T. (1980). Eyewitness identification: Realism vs. the laboratory. Unpublished manuscript.

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. doi:10.1037/1076-8971.12.2.137

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. doi:10.1037/a0030609

Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin, 27,* 1450–1462. doi:10.1177/01461672012711006

Munsterberg, H. (1908). *On the witness stand*. New York: S.S. McClure Company.

Newman, E., & Loftus, E. F. (2012). Clarkian logic on trial. *Perspectives on Psychological Science, 7,* 260–263. doi:10.1177/1745691612442907

Neyman, J. (1950). *First course in probability and statistics*. New York: Holt.

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220. doi:10.1037/1089-2680.2.2.175

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36,* 247–255.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536. doi:10.1177/1745691612463401

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530. doi:10.1177/1745691612465253

Sagan, C. (1980). *Cosmos*. New York: Random House.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470,* 437.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22,* 36–71. doi:10.1016/0010-0285(90)90003-M

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM – Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166. doi:10.3758/BF03209391

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. doi:10.1177/0956797611417632

Smith, S. M., Lindsay, R. C. L., Pryke, S., & Dysart, J. E. (2001). Postdictors of eyewitness errors: Can false identifications be diagnosed in the cross-race situation? *Psychology, Public Policy, and Law, 7,* 153–169. doi:10.1037/1076-8971.7.1.153

Steblay, N. M. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law and Human Behavior, 21,* 283–297. doi:10.1023/A:1024890732059

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. doi:10.1037/a0021650

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14,* 89–103. doi:10.1111/j.1559-1816.1984.tb02223.x

Wells, G.L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*, 553–571. doi:10.1037/0003-066X.48.5.553

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88,* 776–784.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*(2), 45–75. doi:10.1111/j.1529-1006.2006.00027.x

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78,* 835–844. doi:10.1037/0021-9010.78.5.835

Wells, G. L., & Seelau, E. P. (1995). Eyewitness identification: Psychological research and legal policy on lineups. *Psychology, Public Policy, and Law, 1,* 765–791. doi:10.1037/1076-8971.1.4.765

Wells, G. L., Small, M., Penrod, S., 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. doi:10.1023/A:1025750605807

Wells, G.L., Steblay, N.K., & Dysart, J.E. (2011). A test of the simultaneous vs. sequential lineup methods: An initial report of the AJS national eyewitness identification field studies.

Wisconsin Attorney General (2006). Model Policy and Procedure for Eyewitness Identification.

Wixted, J.T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176. doi:10.1037/0033-295X.114.1.152

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. doi:10.1177/1745691612442906