Psychon Bull Rev
DOI 10.3758/s13423-014-0759-2

THEORETICAL REVIEW

When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions

Caren M. Rotello · Evan Heit · Chad Dubé

Abstract There is a replication crisis in science, to which psychological research has not been immune: Many effects have proven uncomfortably difficult to reproduce. Although the reliability of data is a serious concern, we argue that there is a deeper and more insidious problem in the field: the persistent and dramatic misinterpretation of empirical results that replicate easily and consistently. Using a series of four highly studied "textbook" examples from different research domains (eyewitness memory, deductive reasoning, social psychology, and child welfare), we show how simple unrecognized incompatibilities among dependent measures, analysis tools, and the properties of data can lead to fundamental interpretive errors. These errors, which are not reduced by additional data collection, may lead to misguided research efforts and policy recommendations. We conclude with a set of recommended strategies and research tools to reduce the probability of these persistent and largely unrecognized errors. The use of receiver operating characteristic (ROC) curves is highlighted as one such recommendation.

Keywords Eyewitness memory · Signal detection theory · Statistical inference · Social cognition

There is a replication crisis in science, due to both outright fraud (e.g., Verfaellie & McGwin, 2011) and selective reporting of data and statistical tests (e.g., Francis, 2012). It has long been noted that replication and generalization are not as common in psychological science as they are in the physical sciences, leading some hard scientists to view psychology as a "Cargo Cult science" (Feynman, Leighton, & Hutchings, 1985). Indeed, the data from large ongoing replication efforts (Reproducibility Project: Psychology, Open Science Collaboration, 2012; "Many Labs" Replication Project, Klein et al., 2014) are troubling, revealing a 45 % overall failure to replicate. Results such as these strongly imply a need for replication as a matter of course in psychology, as in the physical sciences.

But is replication really enough to ensure the health and future progress of psychological science? We will argue that it is not. Though failure to replicate presents a serious problem, even highly replicable results may be consistently and dramatically misinterpreted if dependent measures are not carefully chosen. The fodder for our demonstration comes from a diverse collection of research domains: eyewitness memory, deductive reasoning, social psychology, and child welfare. Using such examples, we identify two potential pitfalls that can arise from replication: (1) When the analyses are based on a faulty dependent variable (DV), conceptual errors will occur, and replication simply leads to more of those errors; and (2) as the number of replications increases, researchers' confidence may grow unjustifiably, and less attention may be devoted to alternative hypotheses (Fiedler, Kutzner, & Krueger, 2012).

Consider a hypothetical situation illustrating the problem caused by choosing an inappropriate DV. For simplicity, imagine a two-condition, binary-choice experiment in which researchers compare the accuracy of subjects' judgments across conditions using percent correct. The experiment could be about any number of topics, including recognition memory, categorization, perception, or lie detection. Next, assume that there is truly no difference in subjects' ability to discriminate between the two

C. M. Rotello (✉)
Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, 135 Hicks Way, Amherst, MA 01003-9271, USA
e-mail: caren@psych.umass.edu

E. Heit
University of California, Merced, Merced, CA, USA

C. Dubé
University of South Florida, Tampa, FL, USA

Published online: 11 November 2014

classes of stimuli (generically, call them A and B) across the two conditions, but that subjects' preferences for one type of response ("A") over the other ("B") vary with condition. This bias difference could stem from experimental factors such as response reward contingencies, the number of trials deserving of an "A" response, or even the test room's temperature (Risen & Critcher, 2011). Under such conditions of differing biases, Rotello, Masson, and Verde (2008) showed that *t* tests frequently conclude that percent correct differs across conditions.

This outcome is deeply problematic, because subjects' ability to discriminate A from B does *not* differ. Percent correct is a measure that is typically confounded with response bias; thus, an effect that is truly one of response bias alone can be readily misconstrued as an accuracy effect.[1] Two practices that are typically viewed as safeguards against errors actually worsen the problem. Increasing the power of the experiment with a larger number of either subjects or trials per subject merely increases the probability of finding a misleading "difference" in percent correct across conditions.[2] Importantly, replication fails to redress this problem: The error rate does not decrease with replication. With more replications, the total number of errors simply increases. We suggest that a natural tendency of these replications is to unduly increase confidence in the veracity of false conclusions.

In the following four sections of this article, we will show how this hypothetical example plays out in the details of real research. Our examples are highly studied effects coming from the domains of eyewitness memory, deductive reasoning, social psychology, and studies of child welfare, which together use a variety of error-prone DVs. The measures include difference scores, response proportions, and eyewitness diagnosticity. In some cases, even measures derived from signal detection theory (SDT; Macmillan & Creelman, 2005) that are meant to overcome the pitfalls of these measures are also potentially incorrect. Finally, we will consider the DVs used in ongoing replication projects in light of the problems that we discuss.

Example situations in which more data may steer us wrong

Eyewitness memory: the sequential-superiority effect

We begin with an example of great practical significance, the *sequential-superiority effect* in eyewitnesses' lineup identifications. On the basis of numerous studies, psychologists have long argued that sequential lineups, in which a set of photos or suspects are shown to a witness one at a time, lead to identification decisions that are more diagnostic of guilt than those from simultaneous lineups, in which all of the photos are revealed at once. A meta-analysis of "72 tests of simultaneous and sequential lineups from 23 different labs involving 13,143 participant-witnesses" (Steblay, Dysart, & Wells, 2011) reported that the *diagnosticity ratio*, which is the ratio of correct to erroneous identifications (Wells & Lindsay, 1980), is higher for sequential lineups (7.72 vs. 5.78 for simultaneous lineups). Steblay et al. concluded that the sequential-lineup advantage is strong and that the identifications made in sequential lineups are 1.34 (=7.72/5.78) times as diagnostic as those made with simultaneous lineups. As a result, Wells et al. (2000, p. 586) argued that "the simplicity and robustness of the sequential-superiority effect has made it one of the most important of all the practical contributions of eyewitness . . . research." They noted that numerous police precincts have shifted from using simultaneous to sequential lineups: Fully one-third (32 %) of police precincts in the United States now use sequential lineups for eyewitness identifications (Police Executive Research Forum, 2013).

Does higher diagnosticity for sequential lineups imply that witness accuracy is higher? Unfortunately, it does not, because different combinations of correct and false identification rates can yield different values of diagnosticity, despite other measures showing constant accuracy (e.g., Wixted & Mickes, 2012).[3] Not only does the diagnosticity ratio lead to different conclusions than other measures, but—three decades of use notwithstanding—diagnosticity appears to be a poor choice for the data. To understand why, first consider how correct and erroneous decision rates (i.e., hit and false alarm rates, *H* and *F*) are assumed to covary. If we make a scatterplot of *H* as a function of *F*, and connect all of the (*F*, *H*) points that lead to the same diagnosticity value, the result is a straight line.[4] Figure 1 shows those lines, called receiver operating characteristics (ROCs), for several different diagnosticities. All of the points on any given ROC differ only in subjects' preferences for one response over another (i.e., response bias). Toward the lower left end of each line, subjects rarely respond positively, leading to few identifications. Toward the upper right end, they almost always choose from the lineup, leading to many hits but also many errors. If diagnosticity is a good measure of accuracy in the lineup task, then *H* and *F* from subjects with the same accuracy level but different response biases should fall on a straight line like those in Fig. 1.

[1] We use the term "accuracy" to generically reflect performance in a discrimination task, rather than as indicating a particular measure of that performance, such as percent correct or *d'*. Said differently, there is no single measure of accuracy called "accuracy."

[2] As larger samples are obtained, the data provide a better estimate of the underlying model, but the true model for percent correct is not usually consistent with the underlying evidence distributions from which the data are sampled.

[3] Indeed, Dusoir (1975) included diagnosticity as a potential measure of response bias.

[4] Diagnosticity = *H*/*F* so *H* = diagnosticity * *F*, a line with 0 intercept and slope equal to the diagnosticity value.

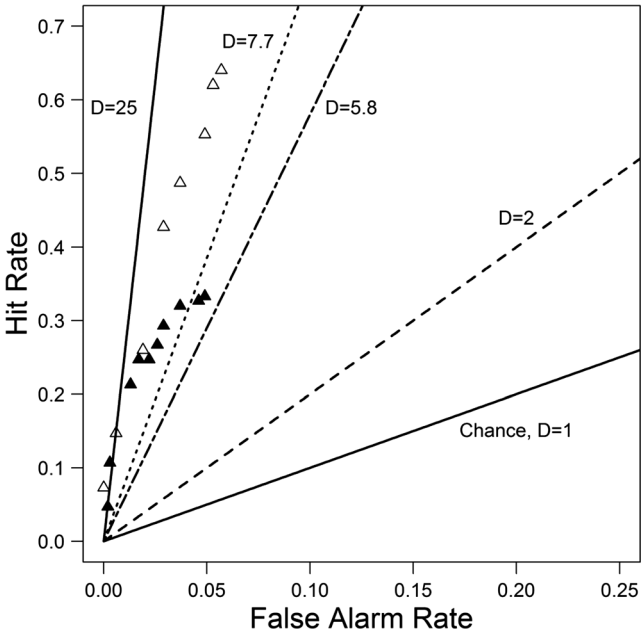

Fig. 1 Linear receiver operating characteristic curves (ROCs) implied by five different values of the diagnosticity measure, *D*, and empirical ROCs from the sequential (filled triangles) and simultaneous (open triangles) lineup conditions of Mickes et al. (2012, Exp. 1b). Note that the axes have been restricted to focus on the data

The observed ROCs for lineup identification do not match the assumptions of diagnosticity. The triangles in Fig. 1 reflect different observed points along empirical ROCs collected by asking witnesses to a simulated crime to rate their confidence in their identification decisions in either a simultaneously (open symbols) or a sequentially (filled symbols) presented lineup (Mickes, Flowe, & Wixted, 2012, Exp. 1b). Every triangle is a point on the same ROC, and thus reflects the same degree of witness accuracy, but a different overall willingness to choose. Notice that the empirical ROCs are curved, rather than linear as would be predicted by diagnosticity. The curved form of these empirical ROCs is commonly observed in memory and perception experiments (see Dube & Rotello, 2012; Pazzaglia, Dube, & Rotello, 2013), as well as in reasoning tasks (Heit & Rotello, 2014). We are not aware of any empirical ROCs that look like those predicted by diagnosticity.[5]

To date, several large-scale studies (total $N = 5{,}411$) have compared sequential and simultaneous lineups using ROCs as the primary analytic tool (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes et al., 2012). Notably, *none* of these experiments found higher accuracy with sequential lineups, and *all* of them found evidence of a reversed effect (simultaneous > sequential) when ROCs were used. The higher diagnosticity associated with sequential lineups is produced by differences in response tendencies, rather than an accuracy advantage: Witnesses are simply less likely to choose a suspect if the lineup is sequential. Thus, the ROC evidence clearly does not support the widely accepted conclusion of a sequential-superiority effect; on the contrary, it suggests instead a simultaneous-superiority effect. As a result of interpreting data with an accuracy measure whose assumptions are unsupported (the empirical ROCs are curved, whereas diagnosticity assumes linear ROCs), researchers reached the wrong conclusion about the optimal lineup procedure. Importantly, this occurred despite very extensive data collection using the same basic task. As a result, approximately one-third of police precincts now use the sequential identification method that leads to lower accuracy!

[5] See Swets (1986a) for a survey of empirical ROCs across a variety of research domains.

Reasoning: the belief bias effect

We turn next to a phenomenon studied in our own research, the *belief bias effect*. This effect is an increased tendency for people to accept the conclusion of an argument when that conclusion is believable rather than unbelievable. Although people prefer conclusions that are consistent with their prior beliefs in many tasks, this effect is commonly investigated in a syllogistic reasoning task. An example stimulus, from Evans, Barston, and Pollard (1983), is:

No addictive things are inexpensive.
Some cigarettes are inexpensive.
*Therefore, some addictive things are not cigarettes. (1)

Example 1 has an invalid but believable conclusion, and the subjects' task is simply to judge its validity. A reordering of the terms leads to another invalid conclusion, but one that is no longer believable:

No cigarettes are inexpensive.
Some addictive things are inexpensive.
*Therefore, some cigarettes are not addictive things. (2)

The same technique can be used to vary the believability of logically valid conclusions.

In the belief bias task, believable conclusions like Example 1 are falsely called "valid" about 70 % of the time, whereas structurally identical problems with unbelievable conclusions, like Example 2, are accepted as "valid" only 10 % of the time. Valid conclusions are usually accepted more often than invalid conclusions, but the factors of believability and validity interact, so that the difference in acceptance rates for valid and invalid problems is larger when the conclusion is unbelievable. A typical result, from Dube, Rotello, and Heit (2010, Exp. 2), is shown in Table 1; both main effects and the interaction are significant here.tgroup

The belief bias effect has been replicated many times (see Dube et al., 2010, for a summary), with the same basic result. For about 30 years, researchers have pursued a theoretical

explanation for the interaction, which was interpreted as showing higher reasoning accuracy when the conclusions are unbelievable. Although this conclusion appears quite believable, it may be invalid.

Until recently, belief bias data were analyzed almost exclusively with analysis of variance (ANOVA)-based comparisons of the response rates in the basic 2 × 2 design shown in Table 1. ANOVA appears perfect for the research questions: Are there main effects of conclusion believability and validity, and do those effects interact? Digging deeper unearths significant problems, however. In a 2 × 2 design like this, ANOVA results are based on difference scores. For the main effects, such tests assess whether the mean of Condition 1 equals that of Condition 2. For the interaction, the procedures test whether the differences between Conditions 1 and 2 are the same size when Treatment A is applied as when Treatment B is applied. For belief bias, the interaction tests whether the difference in "valid" response rates to valid and invalid problems (i.e., $H - F$) depends on the believability of the conclusion. Thus, the use of ANOVA to interpret such data implicitly assumes that $H - F$ (i.e., hits "corrected for" guessing) is an appropriate measure of accuracy.

We can use the same strategy that we applied to diagnosticity to understand the behavior of $H - F$. All combinations of H and F that lead to the same value of $H - F$ (call it k) can be plotted to generate the ROC. These turn out to be straight lines with an intercept of k and a slope of 1[6]; two examples for different values of k are shown as dashed lines in Fig. 2. As for the lines of constant diagnosticity in Fig. 1, each of the points on a straight line in Fig. 2 reflects the same accuracy (this time, as measured by $H - F$) but a different tendency to respond positively. Notice that the points labeled B and U correspond to the data from Table 1; the fact that these points fall on different $H - F$ ROCs implies that they lead to different accuracy levels, consistent with the significant ANOVA interaction term.

An alternative interpretation of points B and U is that they reflect the same accuracy level, but different response biases. The curve shown in Fig. 2 illustrates that possibility: It is the theoretical ROC for a different measure of accuracy, d_a, a measure similar to the detection theory measure d'. To distinguish which interpretation of the data is appropriate, one can compare the empirical ROCs to these theoretical predictions. If the form of the empirical ROC mismatches the form predicted by the DV (e.g., $H - F$), then response bias and accuracy are confounded, and a different DV should be chosen. As in the hypothetical example from the introduction, a failure to choose a different measure in such a scenario can result in significant differences in "accuracy" when in fact accuracy does not vary at all (Rotello et al., 2008), or in differences that are in the opposite direction from reality (as in our eyewitness example).

[6] Setting $k = H - F$, we see that $H = k + F$, a line with intercept k and slope 1.

Table 1 Probabilities of responding "valid" in a belief bias task

Conclusion Type	Valid (H = hit rate)	Invalid (F = false alarm rate)	Difference, $H - F$
Believable	.86	.61	.25
Unbelievable	.68	.32	.36

Data are from Exp. 2 of "Assessing the Belief Bias Effect With Rocs: It's a Response Bias Effect," by C. Dube, C. M. Rotello, and E. Heit, 2010, *Psychological Review*, *117*, pp. 831–863. Copyright 2010 by the American Psychological Association.

Dube et al. (2010; see also Trippas, Handley, & Verde, 2013) demonstrated that the empirical ROCs in belief bias tasks are curved and match those generated by d_a, thus demonstrating that the belief bias effect in their data (e.g., Table 1) was due simply to a response bias difference and *not* to a difference in reasoning accuracy. This analysis indicates that three decades of experimentation and theory development had been misdirected: The presumed accuracy effect of believability was actually a response bias effect.

This analysis of the belief bias literature, like that of the eyewitness identification literature, highlights our two critical concerns. First, increasing numbers of conceptual errors are likely to have occurred as researchers ran more studies with the wrong dependent measures, and second, researchers probably became increasingly confident in their conclusions drawn from these results. Major theories of reasoning (Evans & Curtis-Holmes, 2005; Johnson-Laird, 1983) have grown up alongside this accumulation of data in the belief bias task, yet these theories are all challenged by their faulty empirical foundation.

Heit and Rotello (2014) showed that this problem of using difference scores ($H - F$) and ANOVAs to measure

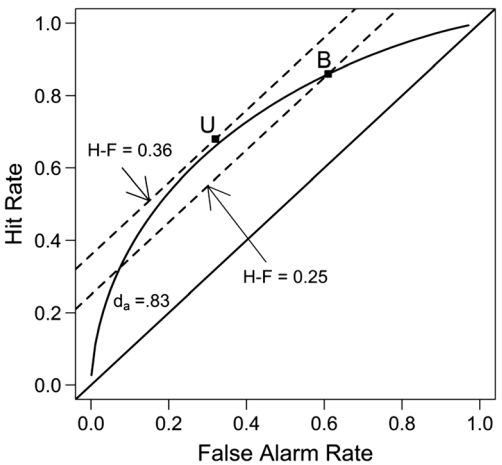

Fig. 2 Theoretical ROCs compatible with empirical data from two conditions (U, B). The dashed lines show the ROCs consistent with difference scores ($H - F$), contrasts, and ANOVA. The solid curve shows the ROC consistent with d_a

performance when the underlying ROCs are curved has also cropped up in other reasoning tasks, including conditional reasoning and induction (see also Heit & Rotello, 2010, 2012; Rotello & Heit, 2009). Thus, an even larger number of results may have been misinterpreted in the reasoning literature, simply because the assumptions of the DV were unsupported.

Social psychology: shooter bias

Perhaps the most replicated finding in social psychology is in-group favoritism: the increased tendency to make positive actions and judgments toward members of one's own social group, relative to out-group members. A dramatic example of this is *shooter bias* (or *weapon bias*; e.g., Correll, Park, Judd, & Wittenbrink, 2002; Greenwald, Oakes, & Hoffman, 2003; Payne, 2001). In a typical shooter bias study, White subjects see pictures of White and Black criminal suspects, paired with either a gun or a nonlethal object such as a cell phone. Subjects must decide quickly whether to shoot at the suspect (because he appears with a gun) or to avoid shooting (because he appears with a cell phone). All of these studies have shown in-group bias—namely, that White subjects are more likely to shoot at Black suspects than at White suspects. However, the detailed results have been somewhat inconsistent from study to study. Related to this problem, these studies have used a variety of analyses with various untested assumptions about the data.

Before reviewing the shooter bias studies in more depth, we note that, to our knowledge, no researcher has published ROC analyses of a shooter bias task. Because no study has collected confidence ratings that would allow the determination of the ROC's form, this section is necessarily speculative. However, we strongly suspect that because shooter bias studies are essentially visual detection tasks, the empirical ROCs are likely to be curved, as in virtually all perception tasks for which ROC data are available (Dube & Rotello, 2012). If the ROCs are indeed curved, then traditional analyses based on difference scores, including ANOVAs, are likely to lead to incorrect conclusions, just as we demonstrated with the belief bias studies.

Several studies of shooter bias (Correll et al., 2002; Lambert et al., 2003; Payne, Lambert, & Jacoby, 2002; Plant & Peruche, 2005; Plant, Peruche, & Butz, 2005) have reported ANOVA results, highlighting an interaction term, such that the error rate depends on an interaction between the suspect's race and the type of object he is holding.[7] However, if the ROCs for this task are curved rather than linear, then this oft-reported result is at risk, for the same reason that the belief bias interaction is problematic (see Fig. 2 and Dube et al., 2010). Moreover, the effects of moderating variables, such as public versus private settings (Lambert et al., 2003), instructions to ignore race (Payne et al., 2002), and practice effects (Plant et al., 2005) are also at risk of misinterpretation when ANOVAs are used.

Several studies have gone further than ANOVAs, by including model-based analyses. For example, Payne (2001), Payne et al. (2002), Lambert et al. (2003), and Plant et al. (2005) used Jacoby's (1991) process dissociation procedure (PDP) as a form of analysis. PDP assumes that the probability of response depends on the combined influences of automatic and controlled processes. Here, automatic processes are assumed to underlie the influence of the suspect's race on shooting decisions, whereas controlled processes underlie the influence of the object held by the suspect (weapon or not). Payne (2001; Payne et al., 2002) reported an automatic response bias to shoot, which was higher for Black than for White suspects. However, he found equal contributions of the controlled processes that respond to the presence of guns for Black and White suspects. Likewise, Lambert et al. (2003) replicated these result, reporting even stronger bias in public than in private settings. In contrast, Plant et al. (2005) concluded that shooter bias was moderated by the amount of practice at the task.

The studies using PDP analyses relied on the implicit assumption that PDP matches the underlying nature of the data. Notably, PDP modeling assumes linear, rather than curved, ROCs.[8] Hence, if the ROCs for the shooter task are curved rather than linear, then the PDP analyses are at risk of reaching incorrect conclusions regarding the nature of shooter bias and its moderators. Relatedly, Sherman et al. (2008) have applied a generalization of the PDP model, known as the Quad model, to the shooter task. As has been noted by Heit and Rotello (2014), the Quad model also generates linear ROCs; hence, conclusions based on that model may be faulty, as well.

Finally, there is some history of conducting signal-detection-based analyses on shooter bias data (Correll et al., 2002; Correll, Park, Judd, & Wittenbrink, 2007; Correll, Park, Judd, Wittenbrink, et al., 2007; Greenwald et al., 2003; Plant & Peruche, 2005). These studies reported the usual SDT accuracy measure d' and bias measure c, which assume curved ROCs. Four of the five studies concluded that the race of the suspect influenced response bias (i.e., shooting is more likely overall for Black suspects), but there was no difference in decision accuracy based on race (i.e., the ability to distinguish guns from other objects was the same for Black and White suspects). This overall conclusion is consistent with the findings based on PDP as well as the Quad model. However,

[7] Some purportedly implicit measures, such as the "bias" measure of response rates used by Correll (2008), also test exactly this interaction.

[8] The reason is a bit technical. The PDP model is formally equivalent to a high-threshold model (Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995), and high-threshold models yield linear ROCs for binary judgments like shoot/don't-shoot (e.g., Pazzaglia et al., 2013).

Greenwald et al. found a different result—namely, that *d′* depended on the race of the suspect. Importantly, the finding that SDT measures sometimes agree with PDP and Quad modeling analyses does not argue in favor of those model-based analyses. Our point is not that at-risk analyses will always lead to incorrect conclusions, but that they are likely to do so (see also the discussion in Heit & Rotello, 2014).

Indeed, the SDT measures may be problematic, as well. Unlike traditional analyses such as ANOVA, and model-based analyses such as PDP and Quad, the measures based on SDT assume curved rather than linear ROCs. Hence, they may provide a better match for the underlying nature of the data. However, these SDT measures assume symmetric ROCs, corresponding to equal variances for the underlying evidence distributions for weapon versus nonweapon, and herein lies the problem. If the ROCs are asymmetric (if the variances are unequal), then the *d′* measure is still prone to drawing incorrect conclusions—for example, failing to distinguish changes in accuracy from changes in response bias (Macmillan & Creelman, 2005; Rotello et al., 2008; Verde & Rotello, 2003).

We have no doubt that the oft-replicated shooter bias effect is real, in the sense that White subjects in these studies tend to shoot more at Black targets (whether or not they hold a weapon). However, the analyses reported in this literature make different, and contradictory, assumptions about the nature of the data. They cannot all be correct. Even if different measures sometimes reach the same conclusions, as has occurred for PDP-, QUAD model-, and SDT-based measures, the underlying models cannot simultaneously be true. Hence, it is difficult to draw conclusions about the potential effects of suspect race on discrimination accuracy (weapon presence), potential moderators, or theoretical explanations. As in our previous examples, it is likely that the combination of replications and inappropriate DVs has led to an accumulation of conceptual errors and confidently held but incorrect theories.

We argue that the only solution to this conundrum is to rerun the classic shooter bias studies using an improved design. Exact replication is highly *undesirable* for these studies, because the original designs do not allow for assessment of the form of the empirical ROC. Thus, the most appropriate measure of accuracy cannot be determined. Consider, for example, Correll (2008), a shooter bias study published in a volume of the *Journal of Personality and Social Psychology*, the entire contents of which have been selected as a target for replication (Open Science Collaboration, 2012). Among the dependent measures reported is something called "bias," which is simply a measure of the interaction between the "shoot" decision rates for guns versus tools and the race of the suspect. As such, "bias" is subject to all of the same criticisms that we raised for the belief bias example. In Fig. 2, point B could reflect the more liberal "shoot" response to Black suspects, and point U could indicate the more conservative response to White suspects. If this interpretation of the data is correct, then the appropriate response is to train police officers to shift their response biases rather than to undergo training designed to influence their discrimination accuracy. Only ROC data can address the question.

Child welfare: maltreatment referrals

As a final example, we turn to another matter of great societal importance: referrals for child maltreatment—namely, the situation in which an individual such as a doctor, teacher, or neighbor reports a case of suspected child abuse or neglect—as well as the investigation process, through which suspected cases may be substantiated (Sedlak et al., 2010). Naturally, there is great interest in assuring that these processes are accurate—for example, that true incidents of maltreatment are reported and nonincidents are not reported. Additionally, there is great interest in assessing racial, ethnic, and regional variations in both the referral process and the actual incidence of abuse. Such research has featured repeated rounds of massive amounts of data collection, worldwide. For example, in the United States, the most recent in a series of national studies of incidence, NIS-4 (the Fourth National Incidence Study: Sedlak et al., 2010) involved more than 1,500 public agencies, more than 11,000 agency staff, and nearly 12,000 detailed case reports, with the aim of drawing conclusions about more than a million cases of abuse per year.

As in the aforementioned experimental studies, the method of analysis is crucial. Examining the referral process, Mumpower (2010) and Mumpower and McClelland (2014) have suggested a variety of alternatives to the traditional approaches commonly taken in this literature (such as *t* tests on measures like *H* and *F*). Their various suggestions included percent correct, *d′*, and ROC analysis.

As we have seen, statistical conclusions depend on the DVs being analyzed. For example, maltreatment referrals are less accurate for Black than for White children when accuracy is measured with percent correct (Mumpower, 2010). However, conclusions differ when SDT measures are used (Mumpower & McClelland, 2014). Namely, the researchers found a substantial response bias difference: The overall tendency to refer was greater for Black than for White children. However, in terms of the *d′* measure of accuracy, an analysis of one type of data suggested *greater* accuracy in referring Black children, and another suggested only slighter greater accuracy for White children.

Drawing conclusions about the correctness of any of these possible analyses for the broader field of child welfare depends crucially on the shape of the empirical ROCs. Mumpower and McClelland (2014) plotted hypothetical ROCs, but their shapes were based on assumptions rather than observed ROCs. Mansell, Ota, Erasmus, and Marks (2011) analyzed two New Zealand–based data sets including initial referrals (with urgency ratings) and subsequent investigations,

and they found curvilinear ROCs; their results are reproduced in Fig. 3a. The Mansell et al. ROCs are based on actual cases, and they clearly indicate that percent correct is not an appropriate measure of referral accuracy. This conclusion is consistent with our own reanalysis of a published data set. Egu and Weiss (2003) studied 540 mandatory reporters (teachers) who were asked to read a vignette about an 8-year-old boy and to rate their agreement with a statement that the child was being abused or that his case should be reported to authorities. We used those agreement ratings to plot separate ROCs for responses to vignettes about children of different races (White, Black, and Hispanic).[9] As is clear in Fig. 3b, all three ROC functions are strongly curved.

Summing up, several large-scale child welfare studies have addressed the accuracy of maltreatment referrals and any possible racial and ethnic differences in referrals and true incidences. Without a doubt, conducting this real-world research involves complexities and subtleties beyond a typical laboratory experiment. Nonetheless, different analyses, assuming either linear or curved ROCs, can lead to different conclusions about these important issues. Once again, we stress that both assumptions cannot be true for a given data set. The limited ROC data available on child welfare referrals suggest that the ROCs are curved; hence, percent correct is an inappropriate measure of accuracy, but additional ROC-based work in this domain will be needed to more solidly establish this conclusion.

Discussion and recommendations

Our example effects are "textbook" effects; they are considered reliable enough to be described in survey courses. Despite the ease and regularity with which these example effects replicate, we have shown that two of them—belief bias and sequential superiority—have been dramatically misinterpreted; a third example strikes us as possibly misinterpreted, as well, and the fourth, child welfare referrals, urgently needs more ROC data for stronger interpretation. In all cases, the core problem stems from an early and persistent failure to consider the properties and assumptions of commonly used DVs (e.g., difference scores, diagnosticity) and the analytic tools (ANOVA, *t* tests), even as massive amounts of data are collected. When these measures and tools have assumptions that are inconsistent with the data—for example, when the DV or analysis assumes a linear ROC but the empirical ROCs are curved—the conclusions from the analysis are likely to be in error (Rotello et al., 2008; Swets, 1986b).

So far we have focused on the sort of binary choices (e.g., studied/new, "A"/"B," or shoot/don't shoot) that are ubiquitous in traditional psychological research. However, the general point—that more attention should be devoted to understanding performance measures and analyses—is broadly relevant. For example, Rotello and Heit (2014) reviewed neuroimaging studies of reasoning (e.g., Goel, Buchel, Frith, & Dolan, 2000; Goel & Dolan, 2003; Stollstorff, Vartanian, & Goel, 2012), showing that an analysis of the "accuracy" of responding to test trials that were congruent versus incongruent (i.e., the correct response did or did not agree with prior beliefs) actually amounted to an analysis of response bias differences rather than accuracy effects. This analysis put interpretations of localized brain function at risk. To move further afield, Tidwell, Dougherty, Chrabaszcz, Thomas, and Mendoza (2014) reviewed the measures used in brain-training studies (e.g., Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Miaskowski et al., 2007). In one common analysis, participants whose performance was affected by training ("responders"), and those whose performance does not change ("nonresponders") are treated as separate groups. These groups are then tested on another task, to assess whether they will show generalized cognitive improvements; the idea is that only responders should show posttraining transfer-task gains relative to pretraining. Tidwell et al. showed that a significant group difference in this responder analysis occurs because there is a correlation between training effects and transfer effects: The analysis is uninformative about whether the training itself leads to enhanced transfer gains, and the same statistical result can occur even when there are no transfer gains, or no training gains, at all. Although intuition may suggest that psychology researchers already have a good understanding of their dependent measures and analyses, these are among many examples that strongly indicate otherwise.

One challenge inherent in situations like those that we have described is that the misinterpreted results are not generally noisy or unreliable. The basic pattern of data may be quite consistent across studies and labs, as in the belief bias task and the sequential-superiority effect, and hence the interpretation of the data is similarly consistent. The interpretive errors are insidious, precisely because the effects are so systematically replicated. Unfortunately, the consequences can be severe. In the case of the belief bias effect, decades of research and theorizing targeted the apparent effect of a conclusion's believability on reasoning accuracy that Dube et al. (2010) concluded is actually a response bias effect. Worse, in the eyewitness domain, criminal defendants' lives have been affected: Police investigators now commonly use a sequential lineup procedure that may actually reduce the accuracy of eyewitness identifications.

This problem—of dramatically and consistently "getting it wrong"—is potentially a bigger problem for psychologists

[9] We treated strong vignettes as positive cases and weak vignettes as negative cases.

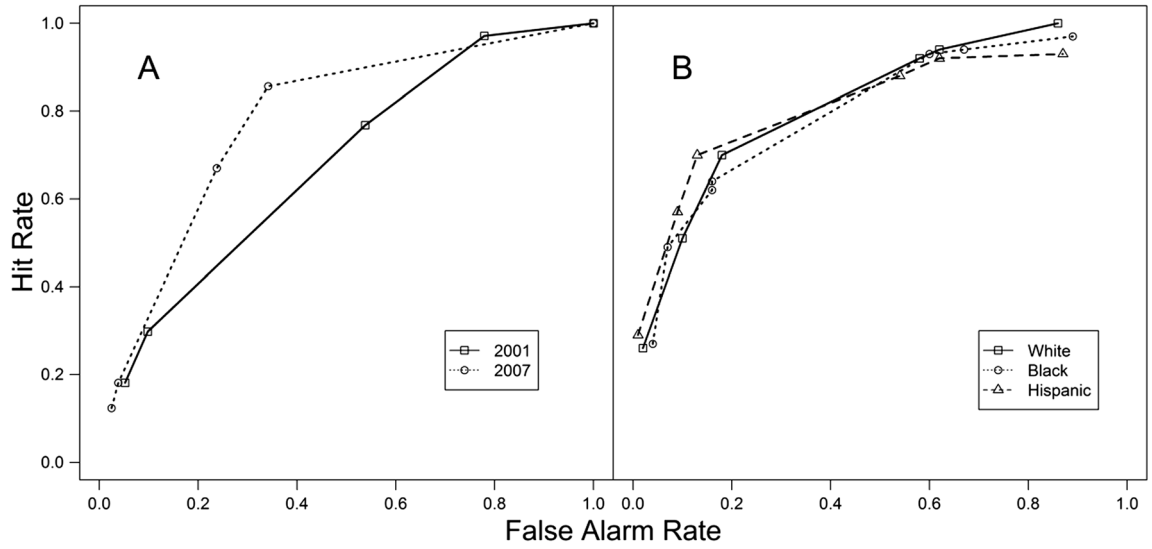

Fig. 3 (**a**) Child welfare ROCs from Mansell et al. (2011). From "Reframing Child Protection: A Response to a Constant Crisis of Confidence in Child Protection," by J. Mansell, R. Ota, R. Erasmus, and K. Marks, 2011, *Children and Youth Services Review*, *33*, pp. 2076–2086. Copyright 2011 by Elsevier. Reprinted with permission. (**b**) Child welfare ROCs from Egu and Weiss's (2003) vignette study of mandatory reporters

than the replication crisis, because the errors can easily go undetected for long periods of time. The probability of self-correction is low, even if ever larger numbers of researchers work on these same (and similar) problems. Indeed, as Rotello et al. (2008) showed, the probability of misinterpreting a response bias effect as an accuracy effect *increases* with sample size if an inappropriate measure of accuracy is applied to the data. Nor is peer review likely to provide a solution: Once an effect is "established," it may become challenging to persuade reviewers that the data should be analyzed differently.

Detection of these interpretive errors requires scientific discipline. It requires careful attention to the details of DVs, thorough awareness of their assumptions, and deliberate testing of their validity. Some of the most commonly used accuracy measures, such as percent correct and "bias-corrected" difference scores (like $H - F$), assume linear ROCs that simply have not been observed empirically (Dube & Rotello, 2012). Even the purported "nonparametric" measures, such as A' (Pollack & Norman, 1964) and the gamma coefficient (Goodman & Kruskal, 1954), have properties that are incompatible with existing empirical evidence and have led to misinterpreted data (e.g., Masson & Rotello, 2009; Verde, Macmillan, & Rotello, 2006).

To guard against data interpretations that may be as faulty as they are familiar, we recommend the following specific strategies:

1. *Know the properties of the chosen dependent measures*, particularly measures of decision accuracy. Repetition of the same analysis that was reported in another article is insufficient. Doing so without understanding the DVs and analytic assumptions that are involved risks perpetuating an erroneous conclusion.
2. *Compare the implied properties of DVs against subjects' actual behavior by collecting empirical ROCs*. Identifying the theoretical ROC for a DV often takes just a bit of algebra: the goal is to discover the function that defines H as a function of F (as we did for diagnosticity and $H - F$). Empirical ROCs are straightforward to generate by using confidence ratings. Macmillan and Creelman (2005) have provided detailed examples of how this is done; we have also provided a tutorial Excel sheet to simplify the task (see www.psych.umass.edu/memlab/roc_stats/). If the theoretical and empirical ROCs have different forms, then a different, and more appropriate, DV should be chosen. Failing to do so risks the reporting of accuracy "values [that] can vary from low to high, by >100 %, when, in fact, accuracy is constant" (Swets, 1986a, p. 196).
3. *Be cautious when response rates (decision biases) differ across conditions*. When response biases differ, accuracy is likely to be confounded with bias, making it easy for accuracy differences to be erroneously inferred across conditions (Rotello et al., 2008). This problem underlies the misinterpretation of both the belief bias and sequential-superiority effects, as well as other empirical phenomena (e.g., the revelation effect: Verde & Rotello, 2003; remember–know responses: Dunn, 2004; Rotello, Macmillan, Reeder, & Wong, 2005; memory for scenes: K. Evans, Rotello, Li, & Rayner, 2009; and memory for negatively valence stimuli: Dougal & Rotello, 2007; White, Kapucu, Bruno, Rotello & Ratcliff, 2014).

Implications for ongoing replication efforts

We began by questioning the value of replication in the absence of some consideration of the appropriateness of the dependent measure in the original study. We then identified several problematic measures of accuracy that have been used for decades, specifically focusing on diagnosticity, percent correct, and differences in response rates across conditions. Replication does not address these fundamental problems. Specifically, considering only the Reproducibility Project: Psychology (Open Science Collaboration, 2012), about 20 % of the experiments targeted for replication involve DVs that we would consider questionable for the reasons that we have outlined. Namely, these experiments rely on dependent measures such as percent correct, or focus on difference scores based on proportions, putting them at risk. Looking forward, we also see this issue arising as psychology journals increasingly encourage the publication of replication articles (Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014).

The emphasis on exact replications has another downside, of course, and that is that any design flaws in the original experiment are also repeated in the replications, and thus their consequences are perpetuated. A profound example of this replication challenge appears in the very first published registered replication report (Alogna et al., 2014), on the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990). Thirty-one labs methodically attempted to replicate Schooler and Engstler-Schooler's Study 1. Upon discovering that the sequence of tasks was not exactly replicated, 22 of those labs ran a second and higher-fidelity replication. Overall, both efforts—involving a total of more than 4,000 participants—produced replications, though with more modest effect sizes than in the original experiment. These data appear to indicate that describing a perpetrator impairs witnesses' subsequent ability to identify him in a lineup. One obvious implication is that jurors should discount any identification made following a verbal description. What all of these studies lacked, however, was a condition in which the lineup did not include the perpetrator. As have described, without information on the false alarm rate in this task, one cannot know whether the participant-witnesses' accuracy, their willingness to choose, or both, are affected by the description task. Indeed, Mickes and Wixted argued that it is possible that a witness's identification might deserve greater, rather than less, weight in court under analogous real-world conditions. Lacking such an important "target-absent" lineup condition, and thus evidence on the form of the empirical ROC that could guide the selection of an appropriate measure of accuracy, these data remain ambiguous.[10]

[10] Additional studies targeted for replication in the Reproducibility Project share this limitation—namely, focusing on hit rate without false alarm rate information.

Conclusion

The replication crisis is a black eye for science, and recent high-profile fraud cases have brought negative attention to psychological research in particular. Although a silver lining in this storm cloud has been an increase in attention paid to the reliability of experimental effects, we argue that a deep problem remains: commonly used, but faulty, DVs can and do produce systematic and consistent conceptual errors, and these errors cannot be remedied by more data collection. The recent focus on replication is necessary, but our examples show that it is not sufficient, and that our failures can affect psychological theorizing as well as real-world practices, with life-and-death consequences. It is time for psychological research to devote considerably more energy to its dependent measures and their assumptions.

Author note C.M.R. and E.H. wrote the manuscript, and C.D. provided critical revisions. All of the authors approved the final version of the manuscript for submission. This material is based on work done while E.H. was serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank David Weiss for providing the raw data from Egu and Weiss (2003), and John Wixted and an anonymous reviewer for their comments on an earlier draft.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., & Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556–578. doi:10.1177/1745691614545653

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes with the process dissociation framework. *Journal of Experimental Psychology: General, 124,* 137–160. doi:10.1037/0096-3445.124.2.137

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. *Journal of Applied Research in Memory and Cognition, 3,* 45–53.

Correll, J. (2008). 1/*f* noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology, 94,* 48–59. doi:10.1037/0022-3514.94.1.48

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007a). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92,* 1006–1023.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83,* 1314–1329.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007b). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology, 37,* 1102–1117.

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.
Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14,* 423–429. doi:10.3758/BF03194083
Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 130–151. doi:10.1037/a0024957
Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117,* 831–863.
Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review, 111,* 524–542. doi:10.1037/0033-295X.111.2.524
Dusoir, A. E. (1975). Treatments of bias in detection and recognition models: A review. *Perception & Psychophysics, 17,* 167–178.
Egu, C. L., & Weiss, D. J. (2003). The role of race and severity of abuse in teachers' recognition or reporting of child abuse. *Journal of Child and Family Studies, 12,* 465–474.
Evans, J. S. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11,* 295–306.
Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11,* 382–390.
Evans, K., Rotello, C. M., Li, X., & Rayner, K. (2009). Scene perception and memory revealed by eye movements and ROC analyses: Does a cultural difference truly exist? *Quarterly Journal of Experimental Psychology, 62,* 276–285.
Feynman, R. P., Leighton, R., & Hutchings, E. (1985). *Surely you're joking, Mr. Feynman! (Adventures of a curious character)*. New York, NY: W. W. Norton.
Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from a-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7,* 661–669. doi:10.1177/1745691612462587
Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19,* 151–156. doi:10.3758/s13423-012-0227-9
Goel, V., Buchel, C., Frith, C., & Dolan, R. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage, 12,* 504–514.
Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87,* 11–22.
Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–764.
Greenwald, A. G., Oakes, M. A., & Hoffman, H. G. (2003). Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology, 39,* 399–405.
Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 1,* 221–228.
Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 805–812.
Heit, E., & Rotello, C. M. (2012). The pervasive effects of argument length on inductive reasoning. *Thinking & Reasoning (Special Issue on Reasoning and Argumentation), 18,* 244–277.
Heit, E., & Rotello, C. M. (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition, 131,* 75–91.
Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18,* 235–241. doi:10.1016/j.tics.2014.02.010
Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541. doi:10.1016/0749-596X(91)90025-F
Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences, 108,* 10081–10086. doi:10.1073/pnas.1103228108
Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
Klein, R. A., Ratliff, K., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). *Investigating variation in replicability: A "many labs" replication project*. Retrieved from Open Science Framework, http://osf.io/wx7ck
Lambert, A. J., Payne, B. K., Jacoby, L. L., Shaffer, L. M., Chasteen, A. L., & Khan, S. R. (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology, 84,* 277–295.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
Mansell, J., Ota, R., Erasmus, R., & Marks, K. (2011). Reframing child protection: A response to a constant crisis of confidence in child protection. *Children and Youth Services Review, 33,* 2076–2086. doi:10.1016/j.childyouth.2011.04.019
Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509–527.
Miaskowski, C., Dodd, M., West, C., Paul, S. M., Schumacher, K., Tripathy, D., & Koo, P. (2007). The use of a responder analysis to identify differences in patient outcomes following a self-care intervention to improve cancer pain management. *Pain, 129,* 55–63. doi:10.1016/j.pain.2006.09.031
Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361–376.
Mumpower, J. L. (2010). Disproportionality at the "front end" of the child welfare services system: An analysis of rates of referrals, "hits", " misses", and "false alarms". *Journal of Health and Human Services Administration, 33,* 364–405.
Mumpower, J. L., & McClelland, G. H. (2014). A signal detection analysis of racial and ethnic disproportionality in the referral and substantiation processes of the U.S. child welfare services system. *Judgment and Decision Making, 9,* 114–128.
Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7,* 657–660. doi:10.1177/1745691612462588
Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181–192.
Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology, 38,* 384–396.
Pazzaglia, A. M., Dube, 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. doi:10.1037/a0033044
Plant, E. A., & Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science, 16,* 180–183.
Plant, E. A., Peruche, B. M., & Butz, D. A. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal

suspects. *Journal of Experimental Social Psychology, 41,* 141–156.

Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. Retrieved from www.policeforum.org/free-online-documents

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science, 1,* 125–126.

Risen, J. L., & Critcher, C. R. (2011). Visceral fit: While in a visceral state, associated states of the world seem more likely. *Journal of Personality and Social Psychology, 100,* 777–793.

Rotello, C. M., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1317–1330.

Rotello, C. M., & Heit, E. (2014). The neural correlates of belief bias: Activation in inferior frontal cortex reflects response rate differences. *Frontiers in Human Neuroscience, 8,* 862. doi:10.3389/fnhum.2014.00862

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review, 12,* 865–873. doi:10.3758/BF03196778

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70,* 389–401. doi:10.3758/PP.70.2.389

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

Sedlak, A. J., Mettenburg, J., Basena, M., Petta, I., McPherson, K., Greene, A., & Li, S. (2010). *Fourth National Incidence Study of Child Abuse and Neglect (NIS-4): Report to Congress*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115,* 314–335.

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

Stollstorff, M., Vartanian, O., & Goel, V. (2012). Levels of conflict in reasoning modulate right lateral prefrontal cortex. *Brain Research, 1428,* 24–32.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181–198. doi:10.1037/0033-2909.99.2.181

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117. doi:10.1037/0033-2909.99.1.100

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2014). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin & Review, 21,* 620–628.

Trippas, D., Handley, S. J., & Verde, M. F. (2013). The SDT model of belief bias: Complexity, time, and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1393–1402.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics, 68,* 643–654. doi:10.3758/BF03208765

Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 739–746. doi:10.1037/0278-7393.29.5.739

Verfaellie, M., & McGwin, J. (2011, December). The case of Diederik Stapel. *Psychological Science Agenda*. Retrieved March 7, 2014, from www.apa.org/science/about/psa/2011/12/diederik-stapel.aspx

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88,* 776–784. doi:10.1037/0033-2909.88.3.776

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, 55,* 581–598. doi:10.1037/0003-066X.55.6.581

White, C. N., Kapucu, A., Bruno, D., Rotello, C. M., & Ratcliff, R. (2014). Response bias for emotional words in immediate recognition memory is due to relatedness rather than emotional valence. *Cognition and Emotion, 28,* 867–880. doi:10.1080/02699931.2013.858028

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