# Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis

**Running Head:** Blinded inference

Jeffrey J. Starns*[1], Andrea M. Cataldo[1], Caren M. Rotello[1], Jeffrey Annis[2], Andrew Aschenbrenner[2], Arndt Bröder[2], Gregory Cox[2], Amy Criss[2], Ryan A. Curl[2], Ian G. Dobbins[2], John Dunn[2], Tasnuva Enam[2], Nathan J. Evans[2], Simon Farrell[2], Scott H. Fraundorf[2], Scott D. Gronlund[2], Andrew Heathcote[2], Daniel W. Heck[2], Jason L. Hicks[2], Mark J. Huff[2], David Kellen[2], Kylie N. Key[2], Asli Kilic[2], Karl Christoph Klauer[2], Kyle R. Kraemer[2], Fábio P. Leite[2], Marianne E. Lloyd[2], Simone Malejka[2], Alice Mason[2], Ryan M. McAdoo[2], Ian M. McDonough[2], Robert B. Michael[2], Laura Mickes[2], Eda Mizrak[2], David P. Morgan[2], Shane T. Mueller[2], Adam Osth[2], Angus Reynolds[2], Travis M. Seale-Carlisle[2], Henrik Singmann[2], Jennifer F. Sloane[2], Andrew M. Smith[2], Gabriel Tillman[2], Don van Ravenzwaaij[2], Christoph T. Weidemann[2], Gary L. Wells[2], Corey N. White[2], Jack Wilson[2]

[1]Organizing Authors (University of Massachusetts, Amherst)

[2]Contributing Authors (Multiple Institutions)

*Correspondence to: jstarns@psych.umass.edu

**Abstract:** Scientific advances across a range of disciplines hinge on our ability to make inferences about unobservable theoretical entities based on empirical data patterns. Accurate inferences rely on both a) discovering valid, replicable data patterns, and b) accurately interpreting those patterns in terms of their implications for theoretical constructs. The replication crisis in science has led to widespread efforts to improve the reliability of research findings, but comparatively little attention has been devoted to the validity of inferences based on those findings. Using an example from cognitive psychology, we demonstrate a blinded inference paradigm for assessing the quality of theoretical inferences from data. Our results reveal substantial variability in expert judgements on the very same data, hinting at a possible *inference crisis*.

**Data and materials availability:** Data and analyses are available at https://osf.io/92ahy/?view_only=2f6d9b285c2d4e279f144b6fed363142.

# Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis

At the most fundamental level, science is the process of creating, testing, and refining ideas that explain and predict natural phenomena. Two core components are necessary for this process to be effective: First, researchers must be able to produce reliable data patterns. Second, researchers must be able to reach sound theoretical conclusions based on those patterns. Scientists in a variety of fields have developed techniques to minimize failure in the first component, that is, to correct the surprisingly high rate of unreliable data patterns reported in the scientific literature, often referred to as the *replication crisis* (Open Science Collaboration, 2015). These techniques, including pre-registration (Miguel et al., 2014), an increased emphasis on direct replication (Open Science Collaboration, 2015), and blinded analysis (MacCoun & Perlmutter, 2015), are crucial for promoting reliable scientific findings. However, we suggest that researchers looking to reform the scientific process should broaden the scope of their investigation to assess whether researchers can make valid theoretical conclusions by analyzing empirical outcomes. This broader perspective could reveal whether some fields suffer from an *inference crisis*; that is, a situation in which researchers have a surprisingly high likelihood of making incorrect theoretical conclusions even if they are working with reliable, replicable data patterns (Rotello, Heit, & Dubé, 2015).

The most direct way to assess inference quality is to create data sets for which the correct inferences are known and to determine whether researchers can discover these correct inferences through blinded data analysis. This *blinded inference* procedure represents an extension of blinding techniques already in common practice. As outlined in 1, blinding techniques applied during data collection and analysis are used routinely to reduce the tendency of researchers and/or participants to promote desired outcomes. Specifically, “blinded data collection” refers to experimental designs that blind the experimental participant, the researcher, or both to the assigned condition (e.g., placebo v. drug), minimizing the ability of these agents to change their behavior according to their beliefs about the assigned condition. “Blinded analysis” techniques, increasingly common in physics (MacCoun & Perlmutter, 2015), hide from the data analyst either the true experimental condition from which each observation is drawn (e.g., scrambled conditions) or the true value of the observation itself (e.g., addition of removable random noise), thereby limiting the ability of analysts to promote desired outcomes with their analysis choices, such as in the well-documented practice of *p*-hacking (Simmons, Nelson, & Simonsohn, 2011). These blinding procedures are valuable tools to limit the malign effects of “researcher degrees of freedom (*df*),” a term that describes the wide range of design and analysis choices researchers can use to address the same research question (Simmons et al., 2011). A recent study (Silberzahn et al., 2018) highlighted the influence of researcher degrees of freedom by sending the same data set to 29 teams of researchers and asking each team to determine whether soccer referees disproportionately “red-card” darker-skinner players. The results showed substantial variability in analysis techniques and conclusions across the research teams.

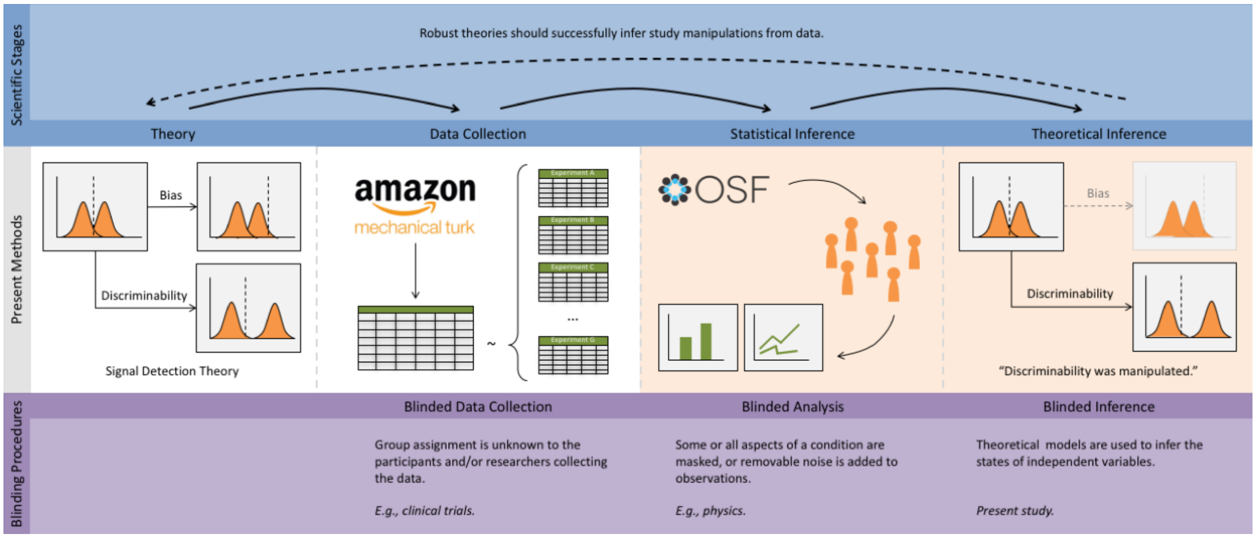


**Fig. 1**. Diagram of the scientific process. The top panel denotes the main stages. The middle panel outlines the methods used in the present paper. The bottom panel denotes common blinding techniques applied in each of the scientific stages, including the blinded inference paradigm advocated for in the present paper.

These blinding methods are excellent strategies to limit the influence of researcher degrees of freedom and/or to assess the consistency of inferences across researchers, but they do not address the *validity* of those inferences. This extra step is crucial because researchers might make inference errors even if they are not promoting a desired outcome with their analysis choices, and these errors could be consistent across researchers who make similar choices (for examples, see Rotello et al., 2015). To assess the validity of theoretical inference, we advocate widespread use of a blinded inference design to supplement traditional approaches. In such a design, researchers who are blinded to condition assignment make inferences about the state of independent variables that are linked to theoretical constructs. Our characterization of the blinded inference technique is heavily influenced by a recent study by Dutilh et al. (2018) in which condition-blinded data sets were sent to response-time modelers who were asked to infer whether the conditions differed in terms of psychological constructs such as response caution and evidence strength. Our general charaterization of the blinded inference approach relies on

Dutilh et al.'s innovative design with two modifications: (1) analysts should be asked to make inferences about empirically manipulated factors rather than latent constructs so that the correct inferences can be unambiguously defined, and (2) analysts should be required to communicate the level of uncertainty associated with their inferences in terms of a probability distribution.

As characterized here, blinded inference can be used in any scenario in which researchers claim that they can (a) measure a theoretical construct based on data patterns and (b) manipulate that theoretical construct with independent variables. If both of these claims are true, then researchers should be able to make accurate inferences about the state of independent variables specifically linked to the theoretical construct by analyzing data. If researchers fail in this task, then it suggests that at least one of the claims is false, i.e., researchers either lack valid techniques for measuring the theoretical construct, lack valid ways to manipulate it, or both. In turn, failures to validly measure theoretical constructs could arise from a variety of problems. One class of problems applies to the process of selecting a measurement model to map patterns of data to underlying processes. Different models might suggest different inferences even if they have a similar ability to match observed data patterns. Another class of problems applies to the process of applying the model, and includes malign factors like parameter estimation biases and mishandling of data.

Concretely, consider a famous example: Mendel and his peas. Mendel recorded systematic patterns of variables, i.e., the relationship between the traits of parents and offspring, and linked them to unobservable theoretical constructs, i.e., hereditary "factors" that obeyed certain laws. His data have been described as being too clean, with too few extreme observations, which may be a result of "unconscious bias in classifying ambiguous phenotypes, stopping the counts when satisfied with the results, recounting when results seem suspicious, and repeating experiments whose outcome is mistrusted" (Hartl & Fairbanks, 2007). Thus, Mendel's conclusions might represent the first documented case of *p(ea)*-hacking. Clearly, Mendel would have benefitted from using blinded analyses to eliminate researcher biases, but we wish to demonstrate how he could have gone further.

By applying his theory of genetics, Mendel claimed to be able to (a) measure underlying hereditable factors by evaluating the phenotype of a plant and (b) manipulate hereditable factors in offspring by selecting parents with certain phenotypes. These are precisely the claims related to the validity of theoretical inference that can be tested in a blinded inference paradigm. For example, someone could have given Mendel a number of plants produced by mating parents with certain traits (unknown to Mendel) and asked him to use his laws of hereditability to predict the likely traits of the *parent* plants by interpreting the traits of the offspring. Mendel would not have been able to make perfect inferences, of course, given that some phenotypes can be produced by multiple genotypes, but he should have been able to make substantially more accurate inferences than someone without a valid theory linking the phenotypes of parents and offspring. We claim that a procedure like this one would have provided a more compelling demonstration of the predictive value of Mendel's laws than unblinded data that could be "massaged." Moreover, by revealing specific offspring phenotypes for which the parents' phenotypes were particularly difficult to predict accurately, it might have allowed the limitations in Mendel's basic theory to be identified more quickly.

Many modern scientists share with Mendel the challenge of making inferences about theoretical constructs on the basis of indirect evidence. For example, modern geoscientists infer the composition and dynamics of Earth's interior from a variety of indirect methods, including radar and magnetic fields. Likewise, cosmologists have inferred that dark matter exists in the absence of direct observation. In the authors' discipline, cognitive processes are inferred from observable behaviors such as decision accuracy or response times. Thus, a critical step in establishing the validity of many scientific claims is to test the inferential power of the data, and this is precisely what the blinded inference procedure achieves: If the researcher is blind to the nature of the manipulation(s), conclusions about what experimental factor was manipulated depend entirely on the data and not on the expectations or unconscious biases of the researcher.

In what follows, we demonstrate the blinded inference paradigm using an example study from recognition memory research. The scheme in the middle of 1 summarizes the design. We conducted a study in which we sent recognition memory researchers ("contributors") seven data sets generated with common experimental manipulations and asked them to make inferences about memory performance. In a recognition memory task, participants are asked to indicate whether they previously encountered a stimulus (often a word) in a certain context (typically a study list). A common question is whether, and to what extent, an independent variable produces changes in discriminability (the ability to distinguish stimuli that were and were not seen in the target context), and in many cases this determination is obscured by differences in response bias (the overall predilection for saying "studied"). Signal detection theory (SDT; Macmillan & Creelman, 2005) was developed in the 1950s with the goal of separating discriminability and bias, and SDT-based measures have been in common use throughout psychology and other disciplines ever since. Several other models or measurement techniques have been developed as alternatives to SDT (Ratcliff, 1978; Riefer & Batchelder, 1988), and some of these also achieved wide popularity throughout psychology (e.g., Erdfelder et al., 2009). Thus, researchers have had nearly seven decades to hone their ability to distinguish discriminability and bias as theoretical constructs, and thousands of papers have been published using models and measures that claim to be able to do so. We tested published memory researchers on their ability to detect whether memory discriminability varied between experimental conditions that might have also varied in terms of response biases.

We have two primary research questions: First, how variable are inferences across researchers? Finding high variability across researchers would be unsettling, given that they all analyzed the same data. Second, and more importantly, how accurate are researcher inferences? If recognition memory researchers have effective methods for manipulating and measuring discriminability and bias based on seven decades of investigating these constructs, then they should be able to make accurate inferences about whether conditions come from the same level or from different levels of a discriminability manipulation.

To preview, we found surprisingly high variability in the inferences of memory researchers asked to interpret the same data, and we also found that many researchers made more inferential errors than would be expected from sampling variability in the data. Given that our task required a relatively simple inference, we suspect that this pattern of surprisingly low inferential accuracy is likely to be found in other research areas. Broadly, however, we emphasize key positive outcomes of this study. Our study exemplifies scientists' commitment to improving the research process, in that many respected memory researchers had the courage to put their conclusions to a public test. Moreover, despite the troubling error rate of the group, our framework identified multiple researchers as having made highly accurate inferences. We therefore believe that our study demonstrates a promising methodology for the future goal of improving inference quality by identifying best practices.

## Methods

### Experimental Design

There were two main phases of data collection. In Phase 1, we collected experimental data in a large-scale recognition memory experiment that used standard study materials and included orthogonally-varied factors known to influence memory discriminability and response bias.[1] The between-subjects design of Phase 1 is analogous to any comparison of memory performance between a special population (e.g., Alzheimer's patients) and a control group, except that our participants were randomly assigned to conditions. In Phase 2, subsets of the full data set were selected to generate seven two-condition experiments in which only the factor affecting discriminability varied (2 experiments), only the factor affecting response bias varied (2 experiments), both factors varied (2 experiments), or neither varied (1 experiment). The conditions in these seven experiments were masked and the data were shared with researchers who had published papers investigating recognition memory, and these experts (or "contributors") were asked to rate the probability that each experiment had only a memory discriminability manipulation, only a response bias manipulation, both, or neither. Contributors were not told how many experiments of each type were included in the data sets, and they were free to select their preferred strategy for distinguishing memory discriminability and response bias.

### Phase 1

[1] All study procedures were approved by the Institutional Review Board at the University of Massachusetts Amherst.

**Participants.** A total of 459 participants were recruited through Amazon's Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011) using psiTurk (Gureckis et al., 2016). Participants earned $1.00 for completing the experiment.

**Materials.** The experiment utilized 104 high-frequency (at least 100 occurrences/million in Kučera & Francis, 1967) English nouns that were 3-7 letters long. Four words were used in the practice block, and the remaining 100 were equally divided into two study lists, A and B. Participants were randomly assigned to study either list A or list B. All participants were tested on the combined list of all 100 words, resulting in complete counterbalancing of stimulus status (studied or unstudied) across participants.

**Procedure.** The experiment was coded in javascript using the jsPsych library (de Leeuw, 2015). Participants were given detailed instructions that included comprehension checks for key components, and they completed a brief practice block before beginning the main task. Word order in the study and test phases was independently randomized for each participant. On each trial of the study phase, participants were asked to report whether the presented word represented an animate object. All of the stimulus words represented clearly animate or inanimate objects, as judged by four independent raters. Each word remained on the screen until the participant entered a response for the animacy question. On each trial of the test phase, participants were first asked to report whether or not they had seen the presented word in the study phase. Participants were then asked to report how confident they were in their response on a 1-3 scale, in which a "1" meant "Not Sure" and a "3" meant "Very Sure". All responses were made via key press, and participants were asked to balance speed and accuracy throughout the experiment.

Memory discriminability and bias were manipulated between participants. Discriminability was manipulated by varying the number of times each word was presented in the study phase (1, 2, or 3). Bias was manipulated by instructing participants to avoid making particular kinds of errors in the test phase. Specifically, conservative participants were told to particularly avoid false alarms (“old” responses to unstudied items), liberal participants were told to particularly avoid misses (“new” responses to studied items), and neutral participants were told to avoid both errors equally. This manipulation was reinforced by varying the quality of the error feedback in the test phase, such that conservative participants saw a “BAD ERROR!” message after false alarms and standard “ERROR” message after misses, liberal participants saw a standard “ERROR” message after false alarms and a “BAD ERROR!” message after misses, and neutral participants saw a standard “ERROR” message in both cases. The “BAD ERROR!” message was accompanied by a reminder of the type of error to particularly avoid and was presented longer than the standard message (2500ms vs. 500ms).

**Phase 1 results.** Complete data are available at the OSF site. A summary of the data analyzed in each of the seven experiments appears in Table 1. We offer no statistical interpretation of these data, given our goal of crowd-sourcing that interpretation in Phase 2 (described next). However, we note that the outcome of this experiment is very consistent with decades of recognition memory literature. For example, hit rates increased and false alarm rates decreased with repeated learning opportunities (as in, e.g., Lachman & Field, 1965; Ratcliff, Clark, & Shiffrin, 1990; Stretch & Wixted, 1998; Verde & Rotello, 2007). We also observed typical effects of response bias manipulations: both hit and false alarm rates tended to increase as increasingly liberal responding was encouraged (e.g., Dube, Starns, Rotello, & Ratcliff, 2012; Han & Dobbins, 2009; Starns, Hicks, Brown, & Martin, 2008; Swets, Tanner, & Birdsall, 1961) and the effects of bias appeared weaker when encoding strength was greater (e.g., Ratcliff, Sheu, & Gronlund, 1992).

| Experiment | Condition | Discriminability | Bias | N | Hit Rate | False Alarm Rate |
|---|---|---|---|---|---|---|
| A | 1 | 3 | Liberal | 24 | .873 | .239 |
| | 2 | 3 | Conservative | 24 | .875 | .126 |
| B | 1 | 1 | Liberal | 27 | .865 | .266 |
| | 2 | 2 | Conservative | 25 | .840 | .191 |
| C | 1 | 2 | Neutral | 27 | .861 | .205 |
| | 2 | 3 | Neutral | 24 | .911 | .174 |
| D | 1 | 1 | Neutral | 27 | .781 | .256 |
| | 2 | 1 | Conservative | 26 | .739 | .195 |
| E | 1 | 1 | Conservative | 26 | .742 | .192 |
| | 2 | 3 | Neutral | 24 | .815 | .190 |
| F | 1 | 1 | Liberal | 26 | .812 | .287 |
| | 2 | 3 | Liberal | 26 | .935 | .164 |
| G | 1 | 2 | Liberal | 26 | .847 | .208 |
| | 2 | 2 | Liberal | 26 | .913 | .208 |

*Notes:* Discriminability represents the number of times each target word was presented in the study phase (1, 2, or 3). Liberal and conservative biases refer to instructions to particularly avoid missing studied items and false alarms to unstudied memory probes, respectively, in the test phase; neutral bias emphasized both errors equally. N indicates sample size, and hit and false alarm rates indicate the proportion of correct and erroneous “old” judgments.

**Table 1.** Definition and summary statistics of the seven experiments sent to contestants.

### Phase 2

**Participants.** Contributors were recruited through targeted e-mails to researchers with a background in recognition memory and/or models of memory and decision making. These individuals were encouraged to forward our invitation to other experts. Out of the 121 researchers who were initially contacted, a total of 46 contributors (comprising 27 PIs and 19 members of their labs) submitted analyses. The data were available in two phases, one for which the confidence-rating data were withheld and another that included the confidence ratings. The purpose of the phases was to investigate whether or not confidence ratings improved inference quality. Of the 27 groups of contributors, 14 also submitted new analyses when the confidence rating data were released. Two contributors declined authorship, and their inferences are de-identified. Of the 44 contributors who accepted authorship, 33 (representing 19 labs) opted to have their inferences associated with their identities; the others chose to remain anonymous. The 27 PIs had an average of 14.7 years of post-Ph.D. experience.

**Materials.** Subsets of data collected in Phase 1 were sampled to form seven "experiments" for the contributors to analyze, summarized in Table 1. Each experiment was designed to have two between-participant conditions that differed in terms of either a memory discriminability manipulation, a response bias manipulation, both, or neither. The data for each condition were created by taking separate random samples of participants who studied list A and participants who studied list B and combining them. Each condition had either an equal number of participants from the two lists or very close to equal (off by one). The data sets that contributors received for the binary analyses included data from the test phase with variables for participant ID, condition (1 or 2), study list (A or B), trial (1-100), test word, whether or not the tested word had been studied (target or lure), the participant's binary response ("old" or "new"), and response time for the binary response. The data sets that contributors received for the confidence rating analyses additionally included the participant's confidence rating, both on the original 1-3 scale and on a recoded 1-6 scale that ranged from "Very Sure New" to "Very Sure Old", and response time for the confidence rating response.

Each contributor completed a submission template summarizing their analyses (see OSF site for an example). The template asked contributors to report the authors collaborating on the submission, accept or decline authorship, and indicate whether they would prefer their conclusions be de-identified. Contributors were then asked to provide a description of their process for analyzing the data in sufficient detail for external replication, a description of any exclusion criteria that were applied, and any code that they were comfortable sharing. All shared code is available at the OSF site. Contributors were lastly asked to report four probabilities for the four possible types of experiment; namely, experiments for which the two conditions were from (1) different levels of a memory strength (discriminability) manipulation but not different levels of a bias manipulation, (2) different levels of a bias manipulation but not different levels of a memory strength manipulation, (3) different levels of both a memory strength and a bias manipulation, or (4) the same levels of memory strength and bias (i.e., null data sets).

**Procedure.** Materials for the binary and confidence rating data analyses were posted to separate private OSF pages. The materials for the binary data analyses were made accessible to contributors on July 7, 2017 and analyses were due August 31, 2017. The materials for the confidence rating data analyses were made accessible on September 9, 2017 and analyses were due on November 1, 2017. No changes to the binary data contributions were allowed after the confidence rating data were released. To support the independence of contributors' inferences, all communication of the coordinating team with contributors was conducted via individually-generated emails, contributors' identities were not shared until mid-November of 2017, and contributors were strongly discouraged from discussing their interpretations of the data with one another in case they accidentally discovered their common participation.

## Results

Our response format was designed to highlight the fact that contributors needed to distinguish the effects of discriminability and bias, but we are primarily interested in conclusions about whether there was a discriminability manipulation. A wide range of research questions in the recognition memory literature require conclusions about discriminability, whereas bias is more often considered a “nuisance” process. Moreover, focusing on discriminability gives our contributors the best chance to succeed because discriminability is better understood and less theoretically contentious than bias (Macmillan & Creelman, 2005). To isolate discriminability inferences, we collapsed the “memory alone” and “both” categories to represent the reported probability of a discriminability manipulation and the “bias alone” and “neither” categories to represent the reported probability of no discriminability manipulation (see OSF for bias results, which unsurprisingly showed poorer inference performance than the discriminability results).

Fig. **2**A shows histograms of the reported probability of a discriminability manipulation across contributors for each of the seven experiments, with regions reflecting correct and incorrect inferences marked in green and red, respectively. The most striking finding shown in Fig. **2**A is the extremely high variability across contributors, with responses spanning a wide range of probabilities for all experiments. For example, some contributors reported a 0% chance that the conditions in Experiment A came from different levels of a memory discriminability manipulation, some reported a 100% chance, and the rest follow an essentially uniform distribution of probability estimates between these two extremes. Responses were concentrated on the correct side for some experiments (e.g., D, F), but not for others (A, B). The high level of variability is surprising given that all researchers received the same data sets. Note that Fig. **2**B, addressed in greater detail below, shows the data that informed the researchers’ inferences, namely the proportion of studied and non-studied items called “studied” (or the “hit rate” and “false alarm rate” in signal detection terms). The dark symbols show results with no participants or trials excluded and grey symbols show results of applying the exclusion criteria used by each contributor. A priori, some experiments seemed likely to be easier to interpret, for example, when both the hit and false alarm rate effects were large and consistent with the same theoretical inference (e.g., in Exp. F, the higher hit rate and lower false alarm rate for Cond. 2 both indicate higher memory discriminability in this condition).

The variability in inferences was matched by high variability in the analysis methods selected by our contributors. These methods, identified on the *y*-axis of Fig. **2**D and described in the Supplemental Materials, are purportedly capable of distinguishing memory discriminability and response bias. Within most of these techniques, some contributors used traditional frequentist statistical methods (e.g., maximum likelihood estimation, significance tests) and others used Bayesian methods (e.g., posterior distributions of parameters or model selection via Bayes Factors). When all analysis choices were considered, no two contributors used exactly the same analysis approach (e.g., same exclusion criteria, measurement technique, and statistical approach).

To summarize inferential accuracy, we counted the number of times across experiments that each contributor reported the true discriminability effect status as the most likely outcome, that is, reported a greater than 50% chance of a discriminability manipulation when discriminability was in fact manipulated or reported a less than 50% chance of a discriminability manipulation when it was not. A histogram of these results appears in Fig. **2**C. Slightly over half of the contributors performed well by this measure, correctly describing five or six of the seven data sets, but the other contributors performed more poorly. We note that the contributor with zero correct inferences estimated a 50% chance of a discriminability manipulation for every experiment, so in fairness, this contributor did not make any *incorrect* inferences either.

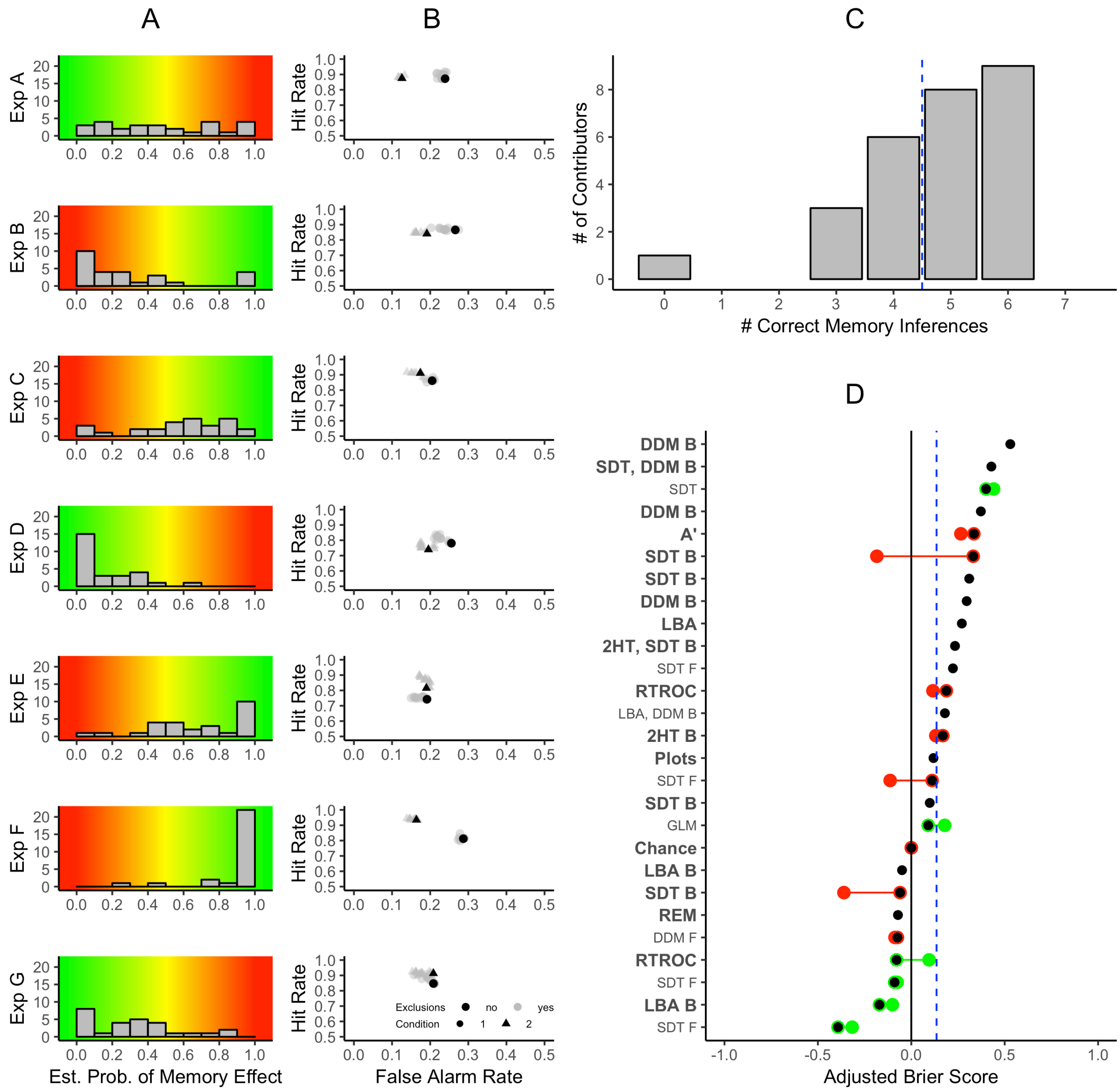


**Fig. 2**. Discriminability inference performance. Panel A: Histograms of contributors' estimated probabilities of an effect in each of the seven experiments. Red denotes incorrect estimates whereas green denotes correct estimates. Panel B: Hit and false alarm rates for each of the seven experiments. Black points represent original values. Grey points represent values after applying each contributor's specified exclusion criteria. Panel C: Histogram of the number of correct inferences out of the seven experiments analyzed for each contributor. The blue dashed line denotes simulation-based benchmark for reasonable performance. Panel D: The adjusted Brier score for each contributor, labelled by their chosen method of analysis. (Note that 19 contributor groups, highlighted in bold, were willing to have their names associated with their responses. The OSF page includes a figure that identifies these contributors.) Black points represent scores for the binary data analysis. Red and green points represent scores for the data analysis with confidence ratings where performance decreased or increased, respectively. The black vertical line denotes chance performance; The blue dashed line denotes simulation-based benchmark for

reasonable performance. Labels on the y-axis denote analysis strategies (defined in the Supplemental Materials) and statistical choices (B = Bayesian; F = frequentist).

Even a valid inference procedure will sometimes reach inaccurate conclusions due to sampling variability, so we needed to identify a benchmark accuracy level below which it would be reasonable to conclude that an invalid inference technique had been applied. We performed model simulations to identify this benchmark. In the simulations, we generated data sets by randomly sampling data from a signal detection model and analyzing those data sets with measures derived from the same model (see the Supplementary Materials for details). Each simulated data set contained the same type of information as the data sets sent to contributors with no labeling to identify the experimental manipulation. Thus, the simulation code performed blinded inference just like our contributors. The key difference between the simulation code and the contributors' analyses is that the former uses an inference procedure that is known to be valid (i.e., consistent with the process that generated the data), so the results represent expected performance levels when sampling variability is the only source of inaccuracy. We set performance benchmarks such that only 10% of the simulated studies fell below the value, meaning that performance is rarely that bad when a valid inference method is applied.

The benchmark for number correct is indicated with a dashed line in Fig. **2**C. Nearly half of the contributors fell below this benchmark, suggesting that some aspect of their inference method was ineffective. To assess whether our empirical data sets were a particularly misleading sample (like the 10% of simulated data sets that produced accuracy below our benchmark even when a valid inference technique was applied), we used the analysis technique from the simulation on the actual data sets sent to contributors and obtained correct inferences for 6 of the 7 data sets. Thus, the empirical data sets do not seem to be a "bad" or misleading sample.

Scientists should be able to express appropriate degrees of certainty in their conclusions, so we also assessed accuracy with a measure that is sensitive to the contributors' probability estimates: the Brier score (Brier, 1950). Brier scores compute the variance between the predicted probability that an outcome will occur and the actual outcome (coded as a 0 or 1). In our case, the outcome is whether or not the two conditions in an experiment come from different levels of a discriminability manipulation. Therefore, the best possible performance is produced by reporting a 0% predicted chance of a discriminability manipulation for all data sets without a discriminability manipulation and a 100% predicted chance for all data sets with a discriminability manipulation, the worst possible performance is the converse, and "chance" performance means reporting a 50% chance for all data sets (meaning that estimates provide no information about which data sets have discriminability manipulations). We adjusted our Brier scores such that 0 represents chance performance, 1 represents the best possible performance, and –1 represents the worst possible performance (see the Supplementary Materials for details). In our simulations to explore performance levels for a valid inference technique, the median adjusted Brier score was .44 and 10% of scores fell below .13, which will thus serve as our benchmark for problematic inferences. Applying the analysis technique from the simulations to the empirical data sets sent to contributors produced a Brier score of .38, which is well above our benchmark.

Fig. **2**D shows ranked Brier scores for our contributors (contributions are labeled by their inference technique). The contributor who reported 50% for every data set is on the chance line. Although this contributor returned no correct inferences, their probability estimates outperformed about one-third of contributors in terms of Brier scores. The contributors who are below chance made multiple incorrect inferences with high confidence levels; in other words, their reported probabilities provided *misinformation* as to which data sets were likely to have a discriminability manipulation. Roughly half of contributors were below the benchmark for problematic inferences, shown by the dashed vertical line, demonstrating that researchers fairly commonly made the mistake of being inappropriately confident in their incorrect inferences. Reassuringly, some contributors achieved Brier scores that are basically as high as can be expected given sampling variability in the data, suggesting that they applied appropriate inference methods. Given the poor overall performance, one might wonder whether these high-performing contributors were simply lucky, indicating that none of our contributors truly succeeded in the inference task. The Supplementary material includes analyses that strongly support the conclusion that at least some of our contributors applied valid inference procedures.

Inference errors were not associated with the choice of any particular analysis technique. The *y*-axis of Fig. **2**D reveals no clear pattern. Methods used by multiple contributors tend to be distributed among the top, middle, and bottom rankings, as are techniques relying on frequentist and Bayesian approaches. Our simulation results also showed that inferences about discriminability are generally robust to different measurement methods, at least for data patterns similar to those in our experiments. Specifically, we reanalyzed all of the simulated data sets using a different measure of discriminability ($P_r$ = hit rate – false alarm rate) that is consistent with a different class of models (Pazzaglia, Dube, & Rotello, 2013; Snodgrass & Corwin, 1988) than the data-generating signal detection model. The $P_r$ analyses achieved accuracy levels that were well above our benchmarks for problematic inferences in terms of number correct and Brier scores (see the Supplementary materials for details). $P_r$ depends on different processing assumptions than the signal-detection model used to sample the simulated data sets, but the two models often make similar discriminability inferences for data set like the ones we sent to contributors (inferences start to diverge for data sets that have large bias effects, but our bias effects were moderate). Thus, it is possible to make appropriate conclusions about discriminability when using a measurement model that does not exactly match the processes generating the data, and selecting an incorrect measurement model cannot entirely explain the poor inference performance revealed in Figure 2.

Variability in inferences was not predictable from contributors' rules for censoring data. Recall that the grey symbols in Fig. **2**B show the mean hit and false alarm rates for each condition with the exclusion criteria used by each contributor. Although these censoring rules clearly resulted in different hit and false alarm rates, we were unable to identify any systematic relationship between these rules and inference accuracy. Moreover, seven contributors did not exclude any data, yet they used different analytic tools and reached different conclusions about the probability of a discriminability effect.

Theoretically, discriminability and bias effects are more easily distinguished with receiver operating characteristics (ROCs) formed from confidence-rating data than with binary old/new response data (Rotello et al., 2015). In a second round of blinded inference, we re-sent the data sets with an addition column for the reported confidence level on each trial, and 14 contributors offered new probability ratings based on the ROCs in each experiment. The resulting Brier scores appear in Fig. **2**D with lines to mark the difference from the corresponding Brier scores based on the binary-response data. The largest changes were actually negative, reflecting reduced inferential accuracy with ROC data.

## Discussion

Distinguishing memory discriminability effects from bias effects is a common empirical issue for recognition memory researchers that has important theoretical and practical implications; for example, understanding memory processes in a special population (e.g., older adults) hinges on the ability to determine if differences from a control group reflect a memory discriminability effect. The available tools to interpret discriminability are well-established, and some have been in use for nearly 70 years (Macmillan & Creelman, 2005). Despite these truths, our expert contributors had mixed success when faced with the task of inferring whether discriminability had been manipulated across conditions that might have also had different levels of response bias. Strikingly, the reported probability of a discriminability effect was highly variable across contributors even though they all received the same data sets. One natural interpretation of these results is that the data themselves were too noisy to allow clear inference. Our simulations are inconsistent with that conclusion: 90% of simulated sets of experiments yielded five or more (of seven possible) correct inferences about discriminability. Thus, we view the outcome of this blinded inference study as a challenge to recognition memory researchers; one which should result in a re-evaluation of our methods, and in humbler presentation of future conclusions that rely on the ability to distinguish discriminability and bias effects. The fact that we found generally low inference quality when researchers used decades-old analysis tools shows that the normal practice of science is not sufficient to ensure effective analysis techniques. Indeed, some examples of systematically problematic inferences have survived decades of scientific review, to the detriment of theoretical progress in those domains (see, e.g., Dube, Rotello, & Heit, 2010, for a specific example and Rotello et al., 2015, for a more general treatment). Widespread use of the blinded inference procedure will help to quickly identify these inference problems and refine analysis methods to optimize inference quality.

**Constraints on generality.** Our study only provides information about a single research scenario – assessing discriminability changes based on recognition memory data – but the fact that we found surprisingly low accuracy for this relatively simple inferential task suggests that problematic inference procedures may plague a broad range of research domains. However, these different domains must be assessed individually in future work, and our results should not be used to make general conclusions about general validity of scientific research. Even within the field of recognition memory, our results are only directly troubling for studies that attempt to make conclusions about discriminability and bias when both processes can potentially vary. Although this is an unavoidable situation for some research questions (e.g., comparing memory across different populations), for other questions memory researchers can substantially simplify the inferential process by experimentally controlling bias when evaluating discriminability, or vice versa. Moreover, memory researchers use a wide range of different types of paradigms and data beyond the recognition tasks that we investigated.

The blinded inference paradigm demonstrated here is also not a substitute for good theory testing and development. A theory that makes correct assumptions could perform poorly in blinded inference based on limitations in the analysis tools available to implement the measurement properties of the theory, and a theory that makes incorrect assumptions might nevertheless serve as a useful tool in some situations (e.g., Newton's Laws are sufficient for many applications despite being incomplete). Our results show that inference problems are not limited to particular theoretical approaches in recognition memory: even researchers who relied on the same measurement model were highly variable in their inferences. Good theory development should run on several parallel tracks simultaneously – empirical assessment, quantitative modeling or analysis, and, we argue, blinded inference studies – to establish that applications of the theory can truly measure what they are intended to measure.

Another potential limitation of our results is that contributors might have applied different analysis standards for our project than they would in a "real" study conducted in their labs. We cannot rule out the possibility that our contributors might have made better inferences if they were analyzing their own data for their own purposes, but there are many good reasons to consider this unlikely. The vast majority of our contributors elected to be co-authors on this manuscript, and a majority (19/27) agreed to have their name directly linked to their performance level in presentations and publications (note that while inference methods were used as labels in Figure 2, results identified by contributor are available on OSF). Thus, one could argue that our contributors had a stronger incentive for rigor compared to typical studies in which no one is likely to re-run the analyses and conclusions are never compared to an "answer key." Indeed, our contributors generally displayed a remarkable level of motivation and dedication to the project, with some applying state-of-the-art techniques like hierarchical Bayesian modeling and/or analyzing the data with multiple measurement models to inform their conclusions. Moreover, the majority of contributors (14/27) agreed to make their analysis code publicly available (see OSF). Thus, we are confident that the inference problems that we observed are not based on a simple lack of effort, and although we cannot rule out the possibility that some contributors made careless, easily correctable mistakes, we seriously doubt that these mistakes can fully explain the inference problems that we observed.

**Comparison to similar studies.** Our results are similar to those of Silberzahn et al. (2018) in that both reveal high variability in inferences across contributors who all received the same data. In many ways, though, the high variability in our contributors' inferences is even more surprising – and troubling – given that our inference task represented a fairly common research scenario. Whereas Silberzahn et al. (2018) asked contributors to address the novel research question of whether referees are biased against darker-skinned players by analyzing real-world data that lacked an experimental control, we asked our contributors to address a research question that has been a focus of recognition memory research for decades and to do so with data from controlled experiments.

Our results are also similar in some respects to the previous blinded inference study reported by Dutilh et al. (2018), but direct comparisons are difficult based on procedural differences between the two studies. In that study, response-time (RT) modelers analyzed unlabeled data sets with the goal of inferring whether the conditions differed in psychological constructs represented in RT models. Unfortunately, contributors disagreed about which cognitive processes should theoretically vary as a function of certain experimental manipulations; in other words, they had different views about what the "answer key" should be. Different scoring rules were developed in light of this disagreement, making it difficult to characterize overall performance. Using the originally planned scoring, at least, the proportion of correct inferences (71%) was similar to our overall accuracy rate (68%). We recommend that future blinded inference studies adopt our strategy of asking contributors to make inferences about experimental manipulations as opposed to underlying theoretical processes to avoid scoring ambiguities. A second difference between our study and Dutilh et al. (2018) also limits our ability to compare the results: Their contributors were not required to express their uncertainty with probability distributions. As a result, we do not know if their contributors' inferences varied as dramatically as ours, with contributors reporting effect probabilities ranging from 0% to 100% for some data sets, and we cannot compare Brier score results between the two studies.

**Refining analysis quality.** Blinded inference can be a method to not only assess inference quality, but also to improve it. Many of our contributors expressed surprise when they learned of their performance level and conveyed that they would carefully re-evaluate their chosen analysis techniques. Our results show that inference problems in recognition memory are not a simple matter of choosing poor measurement techniques, as there are many instances of the same technique being used by both high- and low-performing contributors. Defining the characteristics of effective inference will require additional research, but for now we recommend that analysts try a variety of analysis techniques and, ideally, have multiple researchers independently analyze the data, reserving high confidence for consistent inferences.

**Conclusion.** We will end by again emphasizing that all of our contributors drew inferences about the same data. Thus, the disparate conclusions reached by our contributors are not another example of the replication crisis. Contributors were allowed to use any analysis and any data censoring criteria they preferred, but those researcher degrees of freedom could not systematically influence their conclusions because contributors were blind to the nature of the experimental manipulation. Thus, our findings suggest that current efforts to improve research quality are incomplete, in that they largely focus on limiting researchers' ability to bias results by promoting desired outcomes (whether implicitly or explicitly). Even unbiased analysis techniques can be *ineffective*, so it is critical for scientists to put their skills as analysts to direct (and public) tests. The blinded inference paradigm is a promising method of assessing inference quality and improving analysis procedures, so any field that uses analysis techniques to link data patterns to unobserved theoretical constructs will benefit from applying this method. Our results suggest that even well-established areas of research may be facing an inference crisis that warrants equal consideration with the replication crisis.

## References

Aarts, A., & Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. https://doi.org/10.1126/science.aac4716

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *117*(3), 831–863. https://doi.org/10.1037/a0019634

Dube, C., Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and Language*, *67*(3), 389–406. https://doi.org/10.1016/J.JML.2012.06.002

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., … Donkin, C. (2018). The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychonomic Bulletin & Review*, 1–19. https://doi.org/10.3758/s13423-017-1417-2

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial Processing Tree Models. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8

Han, S., & Dobbins, I. G. (2009). Regulating recognition decisions through incremental reinforcement learning. *Psychonomic Bulletin & Review*, *16*(3), 469–474. https://doi.org/10.3758/PBR.16.3.469

Hartl, D. L., & Fairbanks, D. J. (2007). Mud sticks: on the alleged falsification of Mendel's data. *Genetics*, *175*(3), 975–979. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17384156

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Lachman, R., & Field, W. H. (1965). Recognition and recall of verbal material as a function of degree of training. *Psychonomic Science*, *2*(1–12), 225–226. https://doi.org/10.3758/BF03343418

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*(7572), 187–189. https://doi.org/10.1038/526187a

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van der Laan, M. (2014). Promoting Transparency in Social Science Research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

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*(6), 1173–1203. https://doi.org/10.1037/a0033044

Ratcliff, R. (1978). Theory of Memory Retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037//0033-295X.85.2.59

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*(2), 163–178. https://doi.org/10.1037//0278-7393.16.2.179

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518–535. https://doi.org/10.1037/0033-295X.99.3.518

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. https://doi.org/10.1037//0033-295X.95.3.318

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944–954. https://doi.org/10.3758/s13423-014-0759-2

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. C., … Nosek, B. A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.17605/OSF.IO/QKWST

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, *117*(1), 34–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2966230

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition*, *36*(1), 1–8. Retrieved from http://silk.library.umass.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2008-03226-001&site=ehost-live&scope=site

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1379–1396. https://doi.org/10.1037/0278-7393.24.6.1379

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*(5), 301–340. https://doi.org/10.1037/h0040547

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*(2), 254–262. https://doi.org/10.3758/BF03193446

**Author contributions:**

Conceptualization: Starns, Rotello

Data Curation: Cataldo

Formal Analysis: Starns, Cataldo, Measuring Memory Project contributors

Investigation: Cataldo

Methodology: Starns, Rotello, Cataldo

Project Administration: Starns, Cataldo, Rotello

Software: All authors

Supervision: Starns, Rotello

Visualization: Cataldo

Writing: Starns, Cataldo, Rotello

Inference Contributions: Jeffrey Annis, Andrew Aschenbrenner, Arndt Bröder, Gregory Cox, Amy Criss, Ryan A. Curl, Ian G. Dobbins, John Dunn, Tasnuva Enam, Nathan J. Evans, Simon Farrell, Scott H. Fraundorf, Scott D. Gronlund, Andrew Heathcote, Daniel W. Heck, Jason L. Hicks, Mark J. Huff, David Kellen, Kylie N. Key, Asli Kilic, Karl Christoph Klauer, Kyle R. Kraemer, Fábio P. Leite, Marianne E. Lloyd, Simone Malejka, Alice Mason, Ryan M. McAdoo, Ian M. McDonough, Robert B. Michael, Laura Mickes, Eda Mizrak, David P. Morgan, Shane T. Mueller, Adam Osth, Angus Reynolds, Travis M. Seale-Carlisle, Henrik Singmann, Jennifer F. Sloane, Andrew M. Smith, Gabriel Tillman, Don van Ravenzwaaij, Christoph T. Weidemann, Gary L. Wells, Corey N. White, Jack Wilson

**Competing interests:** Authors declare no competing interests.

**Data and materials availability:** Data and analyses are available at https://osf.io/92ahy/?view_only=2f6d9b285c2d4e279f144b6fed363142.

**List of Supplementary Materials:**

Supplementary Text

Figures S1-S5

Table S1