Stop using d′ and start using d_a: Part I. Simulation explorations of single- and multi-point recognition measures of sensitivity

Adva Levi[1], Caren M. Rotello[2] and Yonatan Goshen-Gottstein[1]

[1] *School of Psychological Sciences, Tel Aviv University;*
[2] *Department of Psychological and Brain Sciences, University of Massachusetts Amherst*

Please address correspondence to Adva Levi,
advalevi@mail.tau.ac.il
Tel-Aviv University

Abstract

In this article, we address the seemingly high prevalence of false discoveries in recognition-memory research. Our challenge is to find a valid measure that effectively separates the contribution of sensitivity (accuracy) from that of bias. A stark realization emerged through Monte Carlo simulations, that in a myriad of tasks, sensitivity is confounded with bias. This is true for tasks that involve binary judgments for single items that are presented at test (Rotello et al., 2008). As a solution, we propose a version of a lesser-known measure, d-sub-a (d_a). Through comprehensive Monte Carlo simulations, we systematically evaluated the validity of common measures Pr = HR - FAR, A′, and d′, alongside d_a. We randomly sampled signals from Lure and Target distributions and used t-tests to compare iso-sensitive conditions differing in bias, across thousands of simulations iterations. For valid measures, significant results should be found at a rate of approximately 5%. We investigated the influence of different parameters, including the form of the distributions, their variability, the distance between them, the placement of response criteria, the sample size and the number of trials. Results revealed that common measures exhibited alarmingly high false discovery rates, exceeding 5%. The rates rose to 100% with larger sample sizes and a large number of trials. In contrast, with only a few minor exceptions, d_a was not affected by changes in bias. Our findings support the notion that d_a should be adopted as the default measure of sensitivity.

Introduction

Consider the following scenario: Researchers test patients on their memory performance compared to aged-matched controls, using a recognition-memory test. They conclude that patients' memory is significantly impaired. Treatment plans emphasizing memory improvement ensue. But what if patients' ability to encode and retrieve items from memory is like that of the control group, and they only differ in bias, a lower (or higher) tendency to classify items as "old"?

This scenario exemplifies the concerns that underlie this article—false discoveries in recognition memory (cf., Open Science Collaboration, 2015)—specifically, due to invalid measurement of participants' performance. In fact, a recent article described a 'crisis of measurement' facing the field of recognition memory (Brady et al., 2022). The crux of the matter is that with regard to commonly used sensitivity measures, none are valid measures of performance, able to tease apart the effects of sensitivity (accuracy) from bias in recognition memory. In the current article, we elucidate the concepts of sensitivity and bias alongside a brief review of the crisis. Importantly, we propose a possible solution that might enable researchers to continue running recognition-memory experiments without exposing their results to heightened risks of reporting false discoveries.

Measures that show invalid patterns of results, leading researchers to find false discoveries, have been previously documented in several research domains (Rotello et al., 2015). An example with an application to the legal domain is the sequential-superiority effect—the presumably better ability to distinguish guilty from innocent suspects in a sequential eyewitness lineup (with suspects presented one at a time in sequential order) as compared to a simultaneous lineup (where all suspects appear together for inspection). This effect is discovered when measuring lineup accuracy

with the diagnosticity ratio (for a meta-analysis, see Steblay et al., 2011). As it turns out, the diagnosticity ratio is higher as a function of the tendency to respond ‘old’ to faces (Mickes et al., 2012), thus confounding sensitivity with bias. Other effects too have been found to confound sensitivity with bias, including the “belief bias effect" in the domain of reasoning, the "shooter bias" in the domain of social psychology, and maltreatment referrals in the domain of child welfare (Rotello et al., 2015).

The current investigation is focused on recognition memory; still, it is likely relevant to tasks other than the recognition-memory task. Recognition is part of a family of tasks known as 'single-interval tasks' (Hautus et al., 2022). Such tasks are ubiquitous across all domains of cognitive research, including perception (e.g., visual-discrimination task), attention (e.g., spatial-cueing paradigm), decision-making (e.g., multi-element averaging task), language (e.g., semantic-relatedness decision task) and meta-cognition (e.g., judgment of learning task). In these tasks, single items are presented at test, and participants are asked to make binary judgments regarding them (e.g., seen-unseen; left-right). Old-new recognition falls into this category because participants are presented at test with single items and asked to make old/new judgments, that is, to judge whether it was an ‘old’ item that had been presented at study (i.e., a target) or whether it was a ‘new’ item that had not been presented at all (i.e., lure). Any conclusions drawn from this article regarding the recognition-memory task are likely relevant, at least to some degree, to other single-interval tasks (Rotello et al., 2015). Importantly, the current article provides a critical step in the direction of eliminating the sensitivity-confound in new research and in uncovering as-yet unknown false discoveries in our literature.

Typically, in single-interval tasks including the recognition task, researchers aim to find a difference in performance between two (or more) test conditions (e.g.,

shallow as compared to deep encoding, Craik & Tulving, 1975). The dependent variable, **sensitivity,** refers to how sensitive participants are in distinguishing between the mnemonic signals associated with targets as compared to those of lures (Hautus et al., 2022). Participants can correctly judge an item as a target (Hit – the item was indeed a target) or do so erroneously (False Alarm – the item was in fact a lure).

Though sensitivity is the latent variable researchers are interested in, **bias**, defined as a participant's proclivity for one choice over the other, also impacts participants' responses (Hautus et al., 2022). Participants with a more liberal level of bias will have a higher rate of 'old' responses, whereas participants with a more conservative level of bias will have a lower rate of 'old' responses, irrespective of whether the item is a target or a lure. The level of bias represents the tendency of an individual to respond one way or the other.

Crucially, a valid measure of sensitivity must be independent of bias. Thus, a measure of sensitivity should index the level of sensitivity but be impartial to changes in bias. Specifically, if two conditions do not differ in sensitivity but differ in bias, a valid measure should erroneously indicate a significant difference between the two conditions only 5% of the time due to type I errors (here and elsewhere, we assume an α level of 5%). We say that the measure is 'bias independent' when the Type I error rate is approximately 5%. In this article, we present Monte Carlo simulations that test these exact scenarios; we compared iso-sensitive conditions with a difference in bias and tested, across thousands of experiments, whether measures of sensitivity show erroneous significant differences at a 5% rate or higher.

Numerous single-point measures of sensitivity have been proposed, each balancing in a different way the magnitude of hit rate (HR; the proportion of "old" responses to targets) with that of the false alarm rate (FAR; the proportion of "old"

responses to lures). Each measure is premised on unique assumptions regarding the population from which the data are sampled (for details, see below, 'measurement model') [1]. The goal was to obtain a measure that indexes sensitivity irrespective of bias.

The idea that a sensitivity measure should be bias-independent is part of many researchers' tacit (and for some, explicit) knowledge. To illustrate, many if not most researchers would be hesitant in inferring changes in sensitivity between conditions due to differences in HR across the two conditions (e.g., 90% and 60%, under deep and shallow encoding, respectively). It is rather obvious that HR is not bias-independent; differences in HR between conditions can reflect changes in sensitivity, bias, or both. While it is obvious that HR is not bias-independent, HR does not balance HR with FAR. Still, unfortunately, it turns out that even measures that *do* balance FARs with HRs are not bias-independent.

The idea that changes in the value of a measure across experimental conditions can possibly be the outcome of changes in bias was brought home in the most powerful way through Monte Carlo simulations (Schooler & Shiffrin, 2005; Rotello et al., 2008). The research we describe in this article is mostly based on (and, in parts, is a replication of) those 2008 simulations, so we turn to a brief description of those simulations.

Rotello and colleagues simulated iso-sensitive aniso-bias conditions and indexed sensitivity using different measures, to test if Type I errors will emerge in a rate higher than 5%. The simulations tested the measures under different forms of lure and target distributions, different manipulation strengths (i.e., distance between the means

[1] The only known measure that is non-parametric, therefore does not entail assumptions regarding the population, is the geometric area under the ROC curve (AUC_g). However, even AUC has been shown to be bias-dependent (Levi et al., 2019).

of the lure and target distributions), different criterion placements, different sample sizes and different number of trials. Under the variety of simulated conditions and across a thousand iterations, the rate of Type I errors was tallied. The results revealed that as a rule, all common measures used for demonstrating different levels of sensitivity in single-interval tasks such as recognition are not bias independent[2].

The first goal of the current article was to replicate these simulations. The 2008 simulations put to test the validity of measures such as Percent correct, A′ and d′ (see details, below). The replication of the 2008 simulations is necessary for at least two reasons. First, like the need for a replication of empirical experiments, simulations too must be independently replicated. Were there mistakes in the code? Did the simulations sample the parameter space correctly? If we wish to consider the implications of the simulations—high rates of Type I errors—the findings must first be shown to replicate.

Second, the 2008 findings did not receive the attention they deserved from the scientific community. At the time the current article was written, the 2008 simulations were cited only 97 times. Researchers continue to use invalid measures until today. Thus, the impact to the previous simulations was negligible, more so when considering the far-reaching implications of the simulations to so many cognitive tasks.

The research community was only prompted to reflect on the manner that experiments were conducted and analyzed following the 'replication crisis' (Open Science Collaboration, 2015). The alarmingly low rates of replications that were found resulted in an influx of articles describing the ailments of current procedures

[2] In the following sections, we specify several scenarios for which specific measures are valid. We will also explain why these scenarios are unlikely for recognition data.

and possible cures (Smith & Little, 2018). New journals were created, and scientific conferences devoted entire sections to addressing issues of methodology. Today's researchers are probably better prepared to process the alarming news that the 2008 simulations bear. It is time, therefore, that they be replicated. We thus examined whether common measures (Pr = HR - FAR, A′, d′; see below for the different formulas) were indeed not bias independent. To anticipate the results, we replicated the findings of the 2008 simulations.

A second, more important goal of this article, was to demonstrate a remedy to this precarious situation. Such a remedy required finding a measure—either a lesser-known measure or an entirely new one—that could be shown to be bias independent. We will promote a little-known measure called d-sub-a (denoted d_a). Specifically, we promote our version of it (see below for detailed explanation). If successful, adopting this measure will provide a possible solution to the measurement crisis.

To better understand the measurement crisis, we now remind readers of the psychometric concepts of reliability and validity. **Reliability** relates to the ability to repeat experiments and obtain similar results. A measuring tape is reliable for measuring the length of different tables but would be less reliable (or not reliable at all) if measuring the width of pieces of paper. **Validity** describes the extent to which changes in the values of the dependent variable are related to the (typically latent) construct it is supposed to measure. Thus, if a measuring tape were to be used to measure the color of a table, the values read from the measuring tape would not capture the hue of the table. This is even though these values might be highly reliable. A measure or test may have a high level of reliability yet not be valid.

The replicability crisis has focused on enhancing reliability (e.g., on running high powered experiments and using large sample sizes; Smith & Little, 2018).

Importantly, highly replicable results do not necessarily represent true discoveries. If a study is reliable (easily replicated) but not valid, it might constitute a false discovery and would thus advance our knowledge in no way whatsoever.

Alas, this seems to be the case for recognition-memory research. The results of the simulations that we wished to replicate here (Rotello et al., 2008) suggest that false discoveries in recognition research are easily replicated and more so with large samples and a large number of trials.

To understand why false discoveries are so prevalent in recognition memory, it is helpful to distinguish between two types of models: the 'data-generating model' and the 'measurement model'. *We argue that for all common recognition-memory sensitivity measures the two models are incongruent. Importantly, model incongruity likely leads to a high prevalence of false discoveries.*

Data-Generating Model. An often-overlooked fact is that empirical data do not come to the world in a haphazard way. Rather, data are sampled from a population with some form. The form may be Gaussian or Uniform or any other imaginable (or unimaginable) distribution. Whatever the distribution, it has certain parameters (e.g., equal or unequal variance). In the best-case scenario, heated debates may exist regarding the population characteristics (e.g., Mickes et al., 2007; Rouder et al., 2010; Wixted & Mickes, 2010). In the worst-case scenario, researchers are not even aware that the data are generated from a distribution. Irrespective of the epistemological knowledge concerning the population characteristics, the data are always sampled from a specific population. The term *'data-generating model*' refers to the population characteristics from which the data are sampled. With regard to recognition memory, our reading of the literature suggests that the leading data-generating model for

recognition memory is that of Unequal-Variance Gaussian distributions. Still, this is by no way consensual (more on this, below).

Measurement Model. Each of the different measures that have been proposed to index performance in single-interval tasks is formulated based on unique assumptions regarding the nature of a (likely unknown) data-generating model (notwithstanding measures that are linear transformations of each other). The term *'measurement model'* refers to the set of assumptions that a measure is based on. For example, the oft-used measure d′ has a measurement model of equal-variance Gaussian distributions of the lure and target items (Tanner & Swets, 1954). Below we list the measurement models for the measures we investigate in this article.

When measuring recognition memory, the equal-variance assumption of d′ is incongruent with the leading data-generating model that assumes unequal variances of these distributions. *Such discrepancies are the root of the measurement crisis: the measurement model is not congruent with the data-generating model.* Note that for any given data set sampled from a particular data-generating model, one measure at most can have a measurement model that reflects the true data-generating model (Brady et al., 2022). Each of the other measures that have different measurement models will necessarily be incongruent with the data-generating model[3]. Unfortunately, there is no guarantee that the research community has identified the true data-generating model. Even if it has, there is no guarantee that any known measure is congruent with this true data-generating model nor that such a measure exists.

[3] Even if the discrepancies between the two models are small, the incongruity between a measurement model and the data-generating model can still be meaningful and render a measure invalid.

But why would a discrepancy between the data-generating model and the measurement model lead to a measurement crisis? In theory, such discrepancies may only influence the validity of the measures to a negligible degree. Indeed, high correlations have been documented between different measures of sensitivity. Given these high correlations, how consequential can using a measure that is premised on the wrong data-generating model be?

It turns out that the consequences of using such assumption-incongruent measures are disastrous, yielding high rates of Type-I errors for non-existent differences of sensitivity. Indeed, the Rotello et al. (2008) simulations reveal that common measures of sensitivity are not robust to incongruities between their measurement model and the data-generating model. Specifically, when the experimental conditions differ only in bias, the measures will show significant differences in sensitivity at a high rate (i.e., high false-discovery rates). Moreover, and unintuitively, as sample size and with number of trials increase, the Type I error rates (T1ER) rise to 100%. Ponder this – a 100% T1ER translates to results that will <u>always</u> replicate, despite being false. As such, common sensitivity measures are invalid measures of sensitivity, and scientific discoveries made with these measures cannot be inferred from the data.

We propose that d_a may be a valid measure of sensitivity yielding low T1ERs when indexing iso-sensitive conditions. Most sensitivity measures are based on only a single (FAR, HR) pair of observations (labeled 'operating points'), and are thus called single-point sensitivity measures (SPSMs). In contrast, d_a can rely on multiple operating points. We now elaborate on these two approaches to measuring sensitivity.

Multi-point sensitivity measures and ROCs. Researchers can use multiple (FAR, HR) points to index sensitivity. One common multi-point measure is AUC_g

(but see Footnote 1). The application of d_a we propose here is also based on multiple operating points (see below).

The most common method for obtaining multiple operating points is by receiver-operating characteristic (ROC) curves (Green & Swets, 1988). ROC curves graphically depict multiple operating points by presenting the HR on the ordinate and FAR on the abscissa across different levels of response bias. The most frequently used method for obtaining ROC data is through the use of confidence ratings. To obtain a confidence ROC (Hautus et al., 2008), judgments are made on a confidence scale (e.g., on a scale ranging from 1-6, '1'= highly confident new; '6'= highly confident old). Given confidence ratings, it is possible to create different bias cutoffs computed for each participant across all responses[4] (Levi et al., 2021). As described in Footnote 4, for a six-point confidence scale, five (FAR, HR) operating points can be computed (For further explanation, see Mickes, 2015). Each operating point reflects the same sensitivity level, but a different level of bias (Hautus et al., 2022). The ROC curve can be transformed into Z space (e.g., if H = .95, then zH = 1.645), yielding a zROC curve.

The zROC curve is an important tool from which the data-generating model of a given task can be inferred (Swets, 1986; Wixted, 2007; Wixted & Mickes, 2010; Dube & Rotello, in press). Specifically, if the underlying distributions are Gaussian, the zROC curve has a linear slope (Hautus et al., 2022). For recognition data, the slope is indeed linear (Wixted, 2007; For a review see Wixted 2020). Importantly, the slope of the zROC curve—S—is an estimate of the lure-to-target ratio of the standard

[4] The most liberal cutoff is created by regarding as 'new' responses of '1' and comparing them to Responses '2'-'6', regarded as 'old'. HR and FAR can thus be computed for this scheme. A slightly less liberal cutoff scheme would be represented by tallying Responses 1 and 2, regarding them as 'new', and pitting them against Responses 3-6 viewed as 'old'. This procedure continues until Responses 1-5 are regarded as ‘new’ and compared to Response 6, for a total of 5 HR-FAR pairs.

deviations (SDs) of the data-generating model. Empirical studies have consistently found that for recognition, S = 0.8 on average. This value represents a target distribution that is 25% wider than the lure distribution (Lockhart & Murdock, 1970; Ratcliff et al., 1992; Dougal & Rotello, 2007; Mickes et al., 2007; Wixted & Mickes, 2010). In summary, according to the analysis of the zROC curve, the best data-generating model for the recognition-memory task is a Gaussian unequal-variance signal detection (UVSD) model. Critically, below we suggest that the measure, d_a, may have a measurement model that is congruent with that of the UVSD data-generating model.

Single-point sensitivity measures. The second approach to measuring sensitivity relies on *a single* (FAR, HR) pair, therefore only requiring participants' binary old/new judgment. Three of the most common SPSM are Pr, A′ and d′. All three measures are based on different measurement models.

Pr = HR-FAR. Pr, also known as "corrected hit probability" or "corrected hit rate" because HR is ‘corrected’ by subtracting the FAR from it. This measure is premised on a double high-threshold model (Hautus et al., 2022), in which the mnemonic information is assumed to be in a discrete, binary state (but see Rotello et al., 2006; Pazzaglia et al., 2013; Starns et al., 2014). Importantly, previous simulations (Rotello et al, 2008) tested Percent correct. Percent correct is a linear transformation of Pr and is based on the same measurement model, therefore the results for Pr and Percent correct are interchangeable. In the current simulations, we tested Pr instead of Percent correct because both Pr and Percent correct are regularly used by researchers in the field, and showing the low validity of both measures in simulations (Percent correct in the 2008 simulations, Rotello et al.; Pr in the current simulations) can

benefit the establishment of a better measurement standard with a larger audience of researchers.

A′. Initially introduced as an assumption-free measure (Pollack & Norman, 1964), A′ was proven to implicitly assume a measurement model of either Rectangular or logistic distributions, depending on the magnitude of sensitivity (Macmillan & Creelman, 2004).

$$\boldsymbol{A'} = \begin{cases} \dfrac{\mathbf{1}}{\mathbf{2}} + \dfrac{(\boldsymbol{H} - \boldsymbol{FA}) \cdot (\mathbf{1} + \boldsymbol{H} - \boldsymbol{FA})}{\mathbf{4}\boldsymbol{H}(\mathbf{1} - \boldsymbol{FA})}, & \boldsymbol{FA} \leq \boldsymbol{H} \\ \dfrac{\mathbf{1}}{\mathbf{2}} + \dfrac{(\boldsymbol{FA} - \boldsymbol{H}) \cdot (\mathbf{1} + \boldsymbol{FA} - \boldsymbol{H})}{\mathbf{4}\boldsymbol{FA}(\mathbf{1} - \boldsymbol{H})}, & \boldsymbol{FA} \geq \boldsymbol{H} \end{cases} \qquad (\mathbf{1})$$

d′. d′ is based on signal-detection theory. It measures the distance between the means of the lure and target distributions, in units of their shared standard deviation. The underlying assumptions of d′ are that the lure and target distributions are equal-variance Gaussians (Tanner & Swets, 1954).

$$\boldsymbol{d'} = \boldsymbol{z}(\boldsymbol{HR}) - \boldsymbol{z}(\boldsymbol{FAR}) \qquad \mathbf{(2)}$$

Importantly, all the SPSM are assumption-incongruent for the UVSD model – the model that perhaps best describes the data-generating model of recognition memory. Pr and A′ assume distributions that are not Gaussian. In contrast, d′ is premised on a Gaussian distribution. However, it assumes the distributions to be equal variance. Critically, if the distributions underlying recognition memory are indeed Gaussian, multiple studies have shown that they are unequivocally not equal variance (Ratcliff et al., 1992; Dougal & Rotello, 2007; Mickes et al., 2007). Stated differently, conditional on the distributions being Gaussian (the measurement model of d′), they are not equal variance—an assumption of the measurement model of d′. Thus, the equal-variance assumption of d′ is indefensible for recognition memory, wherein the data-generating model is reflected by a zROC slope of 0.8.

d_a. Fortuitously, one measure of sensitivity, a multi-point measure, is congruent with the UVSD model. This measure is d_a (Swets & Pickett, 1982). Like d′, d_a is also based on signal-detection theory. Unlike d′, d_a is not premised on an equal-variance assumption. d_a is defined as the distance between the means of the lure and target distributions, measured in the root-mean square average of their SDs (Swets & Pickett, 1982). d_a is based on the subtraction of Z-transformations of HR and FAR, and is defined as:

$$d_a = \sqrt{\frac{2}{1 + \frac{SD_{lures}}{SD_{targets}}^2}} \left(z(HR) - \frac{SD_{lures}}{SD_{targets}} \cdot z(FAR) \right) \quad \textbf{(3)}$$

The measurement model of d_a can support the UVSD[5] data-generating model. By substituting the lure-to-target SD ratio (see Formula 3) with the zROC slope, S, we get a unique (multi-point) way to use d_a as a measure.

Potentially, d_a might be a valid measure of recognition memory, when inserting a fixed value of 0.8 as the SD ratio. However, placing a mean value previously found in experiments as the SD ratio does not consider the variability in the magnitude of this ratio, and of S, across participants. It turns out that the value of S varies across participants, with values as low as 0.5 for some participants and higher than 1 for others[6] (Ratcliff et al., 1992; Macmillan et al., 2004). Here we propose taking this variability into account by estimating S Individually for each participant and then placing the obtained value as the SD ratio in computing d_a for that individual participant.

To reiterate, one goal of this article was to replicate the Rotello et al. (2008) simulations. Like Rotello et al (2008), we too simulated data that were generated from

[5] Note that for a Gaussian equal variance model, S = 1, and d_a reduces to d′.
[6] S=0.5 and S=1.5 reflect target distributions with SDs that are 100% wider and 50% narrower, respectively, of the SD of lure distribution.

equal-variance Gaussian, unequal-variance Gaussian, and Rectangular distributions, in experiments with different manipulation strengths (i.e., the distance between the means of the lure and target distributions), sample sizes and the number of trials. Across thousands of iterations, two simulated experimental conditions were always compared, differing in bias but not in sensitivity. We measured sensitivity with the three common sensitivity measures: Pr, A′ and d′. Because the two conditions were iso-sensitive, significant differences in sensitivity should not be found above the 5% level. As a reminder, Rotello et al. found rates higher than 5%, reaching levels as high as 100% as the number of participants and trials increased.

More importantly, we tested the validity of d_a as a measure of sensitivity, using S as an estimate of the lure-to-target SD ratio. The zROC curves we plotted for the computation of S were based on confidence ROC curves, created from 5 criteria we placed over the lure and target distributions (see Methods).

For our proposed version of d_a, based on estimating the mean value of S across the two conditions for each participant, we included only participants with a minimum of three operating points in each condition. Some participants have fewer than five operating points because the Z transformation of the operating points[7] is not possible for values of HR=1 or FAR=0. As a result, some simulated participants will have fewer than five operating points on their zROC curve (see Footnote 4).

Our choice of a minimum of three operating points was based on the fact that with at least three points on the zROC, no correction method is required (for more detail, see Method). The problem with applying correction methods is that they all require additional assumptions regarding the data-generating model. By excluding

[7] Although the rates of 1 or 0 can be altered using correction methods, we chose to refrain from using such methods (see Methods, below).

participants with fewer than three operating points we were able to refrain from using any correction methods in our analyses. An additional benefit was that we would have a potentially better estimate of the SD ratio (by excluding participants with only two operating points, ostensibly yielding a worse estimation of the SD lure-to-target ratio).

For all measures, we used the same lure and target distributions for both conditions and created a difference in bias by setting Condition 1 to have more liberal criteria (the liberal condition) and Condition 2 to have more conservative criteria (the conservative condition). If a measure of sensitivity is valid, we should obtain the same index of sensitivity for both conditions, reflected in a T1ER of approximately 5%.

We predicted two main results. First, common SPSMs are bias dependent and would yield high rates of Type I error, increasing with sample size and number of trials. Second, that d_a is a bias-independent measure of sensitivity.

Methods

We ran 10,000 iterations for each of 1,476 Monte-Carlo simulations in the goal of testing the validity of Pr, A′ and d′, and importantly, d_a. We simulated experiments of two-level within-participant designs in which we compared sensitivity for iso-sensitive conditions.

To simulate our data, we randomly sampled values of signals from lure and target distributions with means and SDs defined by parameters we chose (see "Distribution variables"). The signals sampled from the target distributions represented participants' responses to targets, both hits and misses. The signals sampled from the lure distribution represented participants' responses to lures, either false alarms (FAs) or correct rejections.

Following Rotello et al. (2008), for each participant in a particular simulation, we used the identical sampled signal values for the two iso-sensitive conditions. By choosing this approach, we matched the two samples thereby simulating data from a within-participant design[8]. The only difference between the conditions was the placement of the criteria (see "*Criteria variables*" for details).

Five criteria were needed to simulate a 6-point confidence scale, henceforth labeled C1-C5 (see Figure 1; see Footnote 4 for theoretical explanation). In both conditions, responses were categorized as hits or FAs (for signals sampled from the target and lure distribution, respectively) when the value of signals was higher than that of the middle criterion (C3, see Figure 1). HRs and FARs were used to calculate Pr, A′ and d′ (see Formula 2 and 3), as well as for d_a.

[8] This sampling procedure yields the optimal conditions for demonstrating a measure's validity, because given the identical observations, the same index of sensitivity should be obtained (within statistical error). Hence, if SPSMs are shown to *not* be valid under our sampling procedure, they cannot be valid when measuring sensitivity in real-world empirical data. The reverse is not true; measures that are found to be valid under our procedure which tests the measure under optimal conditions will still have to be tested in empirical setting to be endorsed as valid measures.

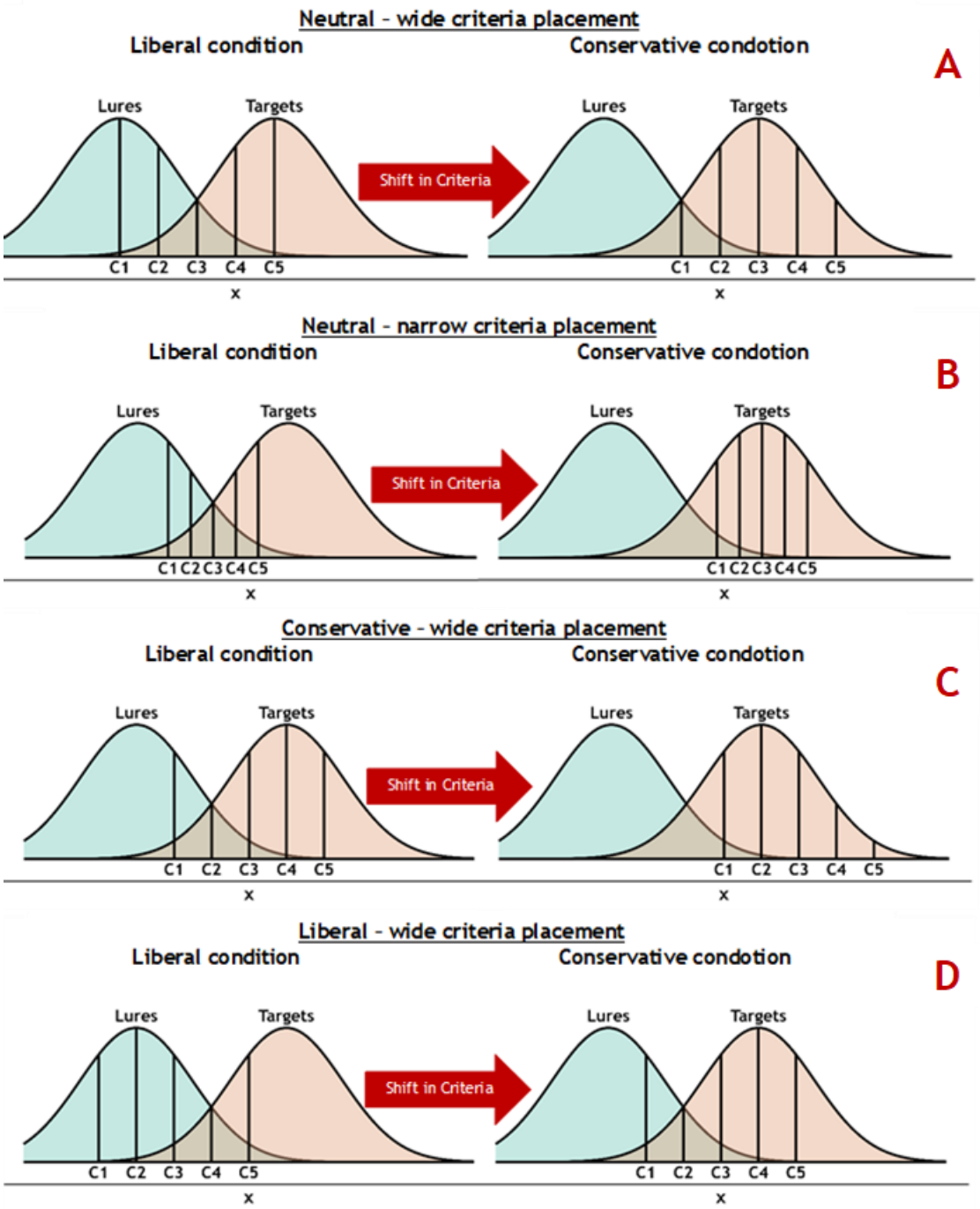

Figure 1. Illustration of different criteria placements in the simulations, and the shift between the Liberal (left) and Conservative (right) conditions for the different placement. The two top illustrations represent a neutral criteria placement, wide (Panel A) and narrow (Panel B). The bottom illustrations represent a Conservative (Panel C) and a Liberal (Panel D) placement. The illustrations depict cartoons of the criteria placement and do not reflect the true values used in the simulations.

For the calculation of S, multiple operating points were needed. See Appendix B, for an illustration of how different points are extracted from each of the criteria. The points were subsequently converted to z-space.

For two reasons, several simulated participants had fewer than five operating points[9]. As mentioned, we included participants with at least three operating points in each condition. To get a more reliable estimate of the true lure-to-target SD ratio, for each participant we computed S separately per condition and used the mean of the two S values for the estimation of d_a. The same value of S was used for calculating d_a in the liberal and in the conservative conditions. Participants who did not have the predefined minimum of operating points for computing S in both conditions, were excluded from analysis.

When using signal-detection sensitivity measures, the conversion of the rates to z-space requires correction methods for values of HR and FAR of 100% and 0%, respectively (e.g., log-linear correction methods, Snodgrass & Corwin, 1988). However, no correction methods were needed in the analysis of our simulated data when requiring at least three operating points in each condition. When analyzing the data of participants with a minimum of three operating points, C3 will always yield HR values lower than 100% and FAR values higher than 0%. This is so because participants for whom 100% of the simulated signals sampled from the target distribution exceed C3 will be discarded because they will have fewer than three operating points. For these participants, the HR for C3 will be 100%. The HR for C1 and C2, the lower criteria, will inevitably also be 100%, because all 100% of the signals that exceed C3 will also exceed the lower criteria[10]. At bottom, under the algorithm requiring at least three operating points per condition, the HR and FAR we

9 First, HR of 100% and FAR of 0% yield z-scores equal to infinity. Therefore, all such operating points were eliminated from the computation of S. Second, some participants had two or more identical operating points. For example, if the sampled responses for a participant all exceeded the second criterion, then the HR and FAR for the first and second criterion would be the same. The upshot of having two (or more) identical operating points was fewer points for the computation of S.

10 Likewise, participants with FARs of 0% for C3 will have fewer than three operating points and will be discarded from the analysis. FARs of 0% represent data for which no responses from the lure distribution exceed C3. Inevitably, for these participants, 0% of signals will exceed the higher criteria, C4 and C5.

used for calculating sensitivity (see Formulas 2 and 3) are never equal to 100% and 0%, and no correction method is needed neither for the calculation of d_a nor for the calculation of d′.

Following Rotello et al. (2008), we sampled the parameter space with multiple experimental scenarios. The different experimental scenarios varied across three types of study variables, with their levels completely crossed. The variable types were *Experimenter variables*—variables that are controlled by the experimenter in an empirical setting, *Distribution variables*—variables that describe the data-generating model, usually unknown by the experimenter, and *Criteria variables*—the placement and arrangement of the different confidence levels, corresponding to participants' possible levels of bias. See Figure 2 for a summary of the variables and their levels.

Experimenter variables	Sample size	25, 50 or 100 participants
	Number of trials per condition	64, 128 or 256 trials per condition
Distribution variables	Type of distribution	Gaussian (Equal-variance, Unequal-variance with 0.8 SD ratio, Unequal-variance with 0.6 SD ratio) Rectangular (Equal-variance)
	Distance between distributions	0.5, 1, or 2 SDs for the Gaussian distributions, 0.37 SDs for the rectangular distributions (relative to the Lure distribution)
Criteria variables	Placment	how liberal the placement is
	Spread	wide or narrow
	Shift size	0.3, 0.5 or 0.7 SDs for the Gaussian distributions, 0.08 SDs for the rectangular distributions (relative to the Lure distribution)

Figure 2. Summary of all the variables, and their levels, used in the simulations to explore the parameter space. All variables are categorized by variable types.

Experimenter variables. Twenty combinations of five sample sizes (N = 15, 25, 50, 75 or 100) were crossed with four numbers of trials (T = 32, 64, 128 or 256).

Distribution variables. We simulated experimental scenarios of both equal and unequal-variance Gaussian lure and target distributions. We also simulated equal variance Rectangular distributions. We simulated scenarios of different distances between the lure and target distributions (For the Gaussian distributions, d = 0.25, 0.5, 1, 1.5, 2 and 2.5 SDs relative to the lure distribution. For the Rectangular distributions, d = 0.37 SDs relative to the lure distribution [11]). A larger distance between those distributions reflects a higher sensitivity for distinguishing between targets and lures.

Criteria variables. We set criteria under different schemes for the liberal condition and defined the conservative condition relative to the liberal condition (see Figure 1). For simplicity, the criteria were always equally spaced. For the liberal condition we manipulated the locations of the criteria (Liberal, Neutral or Conservative) as well as their spread, that is, the distance between the criteria (Wide or Narrow).

For the conservative condition, we shifted the five liberal criteria by a fixed amount. We created different experimental scenarios by varying the magnitude of the shift (For the Gaussian distributions, 0.3, 0.5 or 0.7 SDs relative to the lure distribution. For the Rectangular distributions, 0.08 SDs relative to the lure distribution). This manipulation allowed us to find boundary conditions under which a measure might not be bias-independent. For measures that are affected by bias, a

[11] When running rectangular lure and target distributions, we wanted to cover the parameter space in a manner that is comparable to the Gaussian distributions (e.g., small and large distances between distributions have approximately the same meaning in regards to the lure SD). However, some distribution and criteria variables led to a very large exclusion of participants, such that in several simulations we did not have enough participants to run a t-test. We therefore limited the experimental scenarios for Rectangular-distributions simulations. We still obtained results from 18 different simulations of rectangular distributions, each containing 10,000 iterations. All the results from those simulations appear in Appendix A.

bigger shift in bias might lead to more erroneously significant results (i.e., higher T1ER).

Results

We investigated 1,476 parameter combinations – each parameter combination will be referred to as a simulation. A simulation was defined as identical values across all the parameters (sample size, number of trials, distance, and criteria placement) [12]. We ran 10,000 iterations of each simulation. The dependent variable was T1ER — rate of erroneous significant results— across the 10,000 iterations. The detailed results per simulation appear in Appendix A. Here we provide representative findings. Both the data and the simulations script will be made available through the OSF link – https://osf.io/d5jub.

Type I error rates for SPSMs. Pr, A′ and d′ and d_a were computed for each of the two experimental conditions and the data were then submitted to a dependent sample t-test. T1ERs were calculated as the proportion of significant tests out of the 10,000 iterations, per SPSM. Figure 3 presents T1ERs of Pr (Panel A), A′ (Panel B) and d′ (Panel C) as a function of sample size and number of trials for simulations assuming Gaussian distributions (equal variance, left column; unequal variance, right column). Figure 4 presents T1ERs of Pr (Panel A), A′ (Panel B) and d′ (Panel C) as a function of sample size and number of trials, assuming Rectangular distributions.

[12] Note that the degrees of freedom for the t-tests was not always equal to the number of simulated participants. In addition to the exclusion of participants explained in the Methods section, some simulated participants yielded values of H=0% or FA=100%, reflecting wrong responses for *all* target or lure trials, respectively. These participants were also excluded (so as to best mimic the analysis of empirical data). Indeed, in empirical setting, H=0% or FA=100% suggests that the participant did not understand task instructions or did not pay attention at either encoding or retrieval. Thus, their data will probably be eliminated from analysis, a procedure we endorse.

Our results completely replicated those of Rotello et al. (2008). Thus, T1ERs for SPSMs were approximately 5% under only two conditions, across their parameter space. First, for d′ under equal-variance Gaussian distributions (Appendix A). Second, for Pr under Rectangular distributions (also see Appendix A). Notwithstanding these two conditions, for all other experimental conditions, SPSMs were higher than 5% and grew to 100% with increases in sample size and in the number of trials. We note in particular the simulations of unequal-variance Gaussian distributions (likely the distributions underlying recognition memory) for which all SPSMs showed high T1ERs. Together, the results are consistent with the idea that T1ERs are approximately 5% if and only if the data-generating model is congruent with the measurement model (i.e., T1ER of d′ for Equal-variance Gaussian distributions; T1ER of Pr for Rectangular distributions).

An additional finding, again replicating Rotello et al. (2008), was that under all scenarios for which T1ERs were higher than 5%, the rates were positively correlated with both the number of trials and sample size. For example, when computing TIERs of d′ for unequal-variance simulations, T1ERs unfailingly reached 100% for simulations with 256 trials and sample size of 100 participants[13].

[13] This was the case for all simulations where the size of shift in criteria between conditions was 0.5 or 0.7. However, for a very small shift in criteria (0.3), T1ER sometimes reached values of 'only' 98% or 99%. This is because the small size of shift resulted in a small difference in bias, so few T1ERs occurred due to the (small) changes in bias.

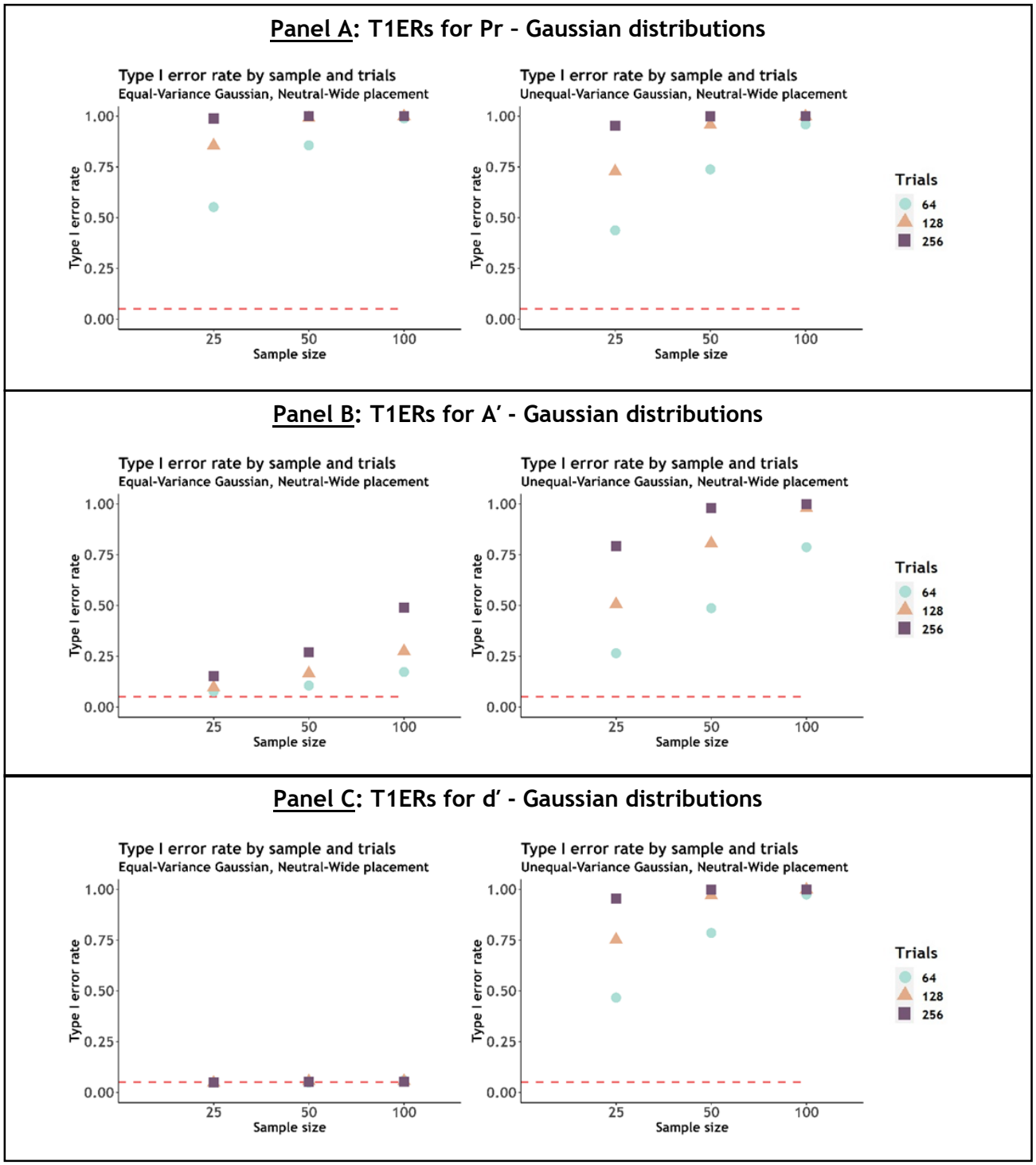

Figure 3. Type-I error rates of Pr, A′ and d′ under Gaussian-distributions (equal and unequal variance; left-side and right-side figures, respectively) as a function of sample size and number of trials. The spread of criteria was neutral and wide. The distance between distributions was 1 SD relative to the lure distribution and the size of shift in criteria between conditions was 0.5 SD relative to the lure distribution. For Unequal-variance Gaussian distributions, S = was 0.8 (reflecting the lure-to-target SD ratio of recognition memory).

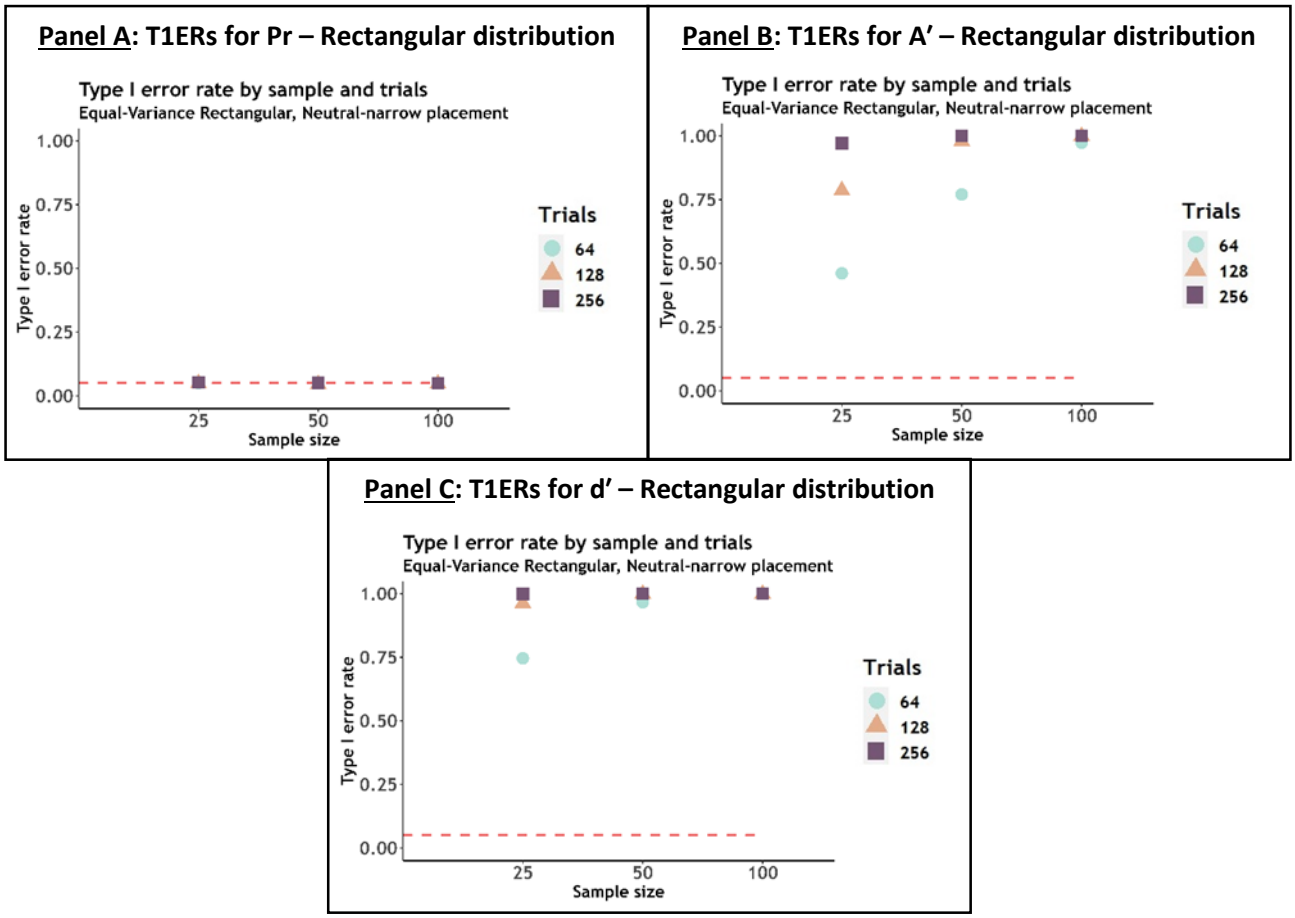

Figure 4. Type-I error rates of Pr, A′ and d′ under Rectangular-distributions as a function of sample size and number of trials. The spread of criteria was neutral and narrow. The distance between distributions was 0.37 SD (relative to the lure distribution), and the size of shift in criteria between conditions was 0.08 SD (relative to the lure distribution).

Type I error rates for d_a. d_a was the one measure that stood out from the crowd, with T1ERs of only 5% for both equal and unequal Gaussian distributions. Figure 5 illustrates T1ERs of d_a as a function of sample size and number of trials. For Gaussian distributions, both equal and unequal variance, d_a presented the acceptable 5% levels of Type I errors, across all values of sample size and number of trials.

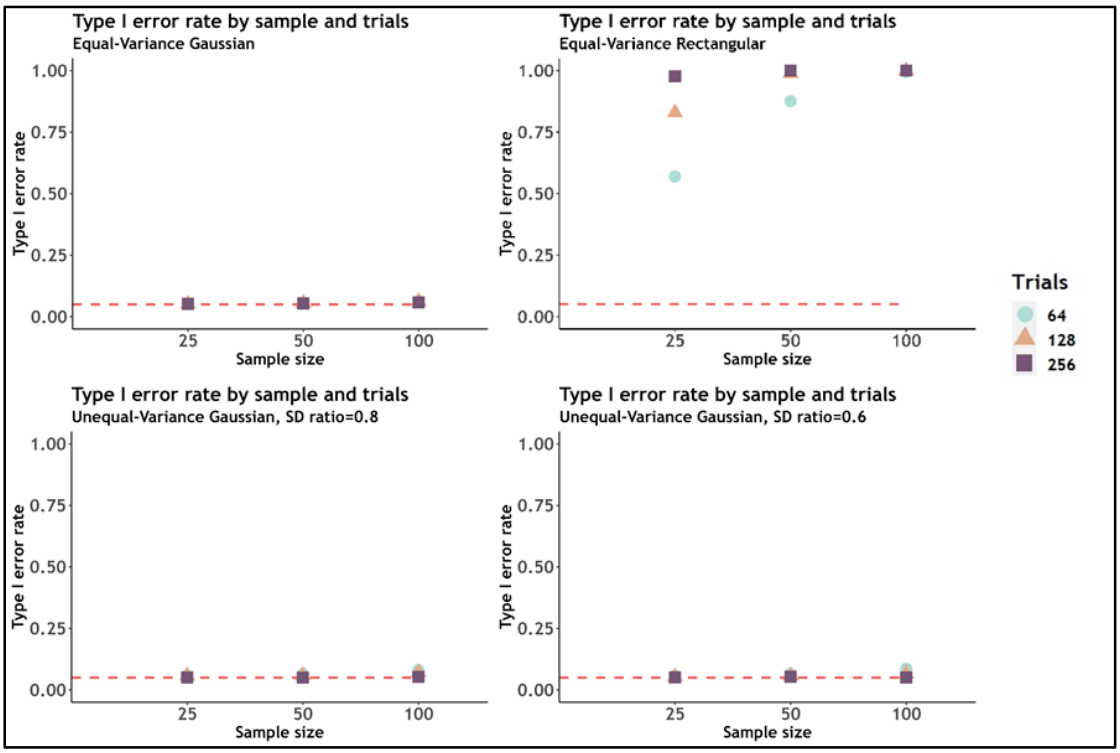

Figure 5. Type-I error rates of d_a for different simulated distributions (Top left: Equal-variance Gaussian, Top right: Equal-variance Rectangular, Bottom left: Unequal-variance Gaussians, S = 0.8, Bottom right: Unequal-variance Gaussians S = 0.6), as a function of sample size and number of trials. For Gaussian distributions, the spread of criteria was neutral and wide. The distance between distributions was 1 SD (relative to the lure distribution), and the size of shift in criteria between conditions was 0.5 SD (relative to the lure distribution). For Rectangular distributions, the spread of criteria was neutral and narrow. The distance between distributions was 0.37 SD (relative to the lure distribution), and the size of shift in criteria between conditions was 0.08 SD (relative to the lure distribution).

Still, we identified three experimental conditions for which the rate of d_a were consistently higher than 5% (see Appendix A)[14]. First, for Rectangular distribution, T1ERs of d_a were usually higher than 5%. This finding was predicted because a data-generating model which assumes Rectangular distributions is incongruent with the Gaussian assumption of the measurement model underlying d_a.

Second, for Gaussian distributions, high T1ERs were observed when two cumulative conditions were met: a low number of trials in conjunction with a large

[14] Because of the large number of simulations, idiosyncratic combinations of parameters sometimes yielded T1ES higher than 5%. We did not document these patterns because they did not seem to illuminate the conditions under which d_a cannot be trusted to yield 5% error rates.

sample size. For example, in Figure 5, T1ERs reached approximately 10% when the sample size was 100 (the largest simulated sample size) with 64 trials per condition (the fewest simulated number of trials; see Figure 5 – Bottom Panels, teal circles for 64 trials). Note that the sample size of 100 did not produce a high T1ER when the number of trials per condition was higher (128 or 256; see Figure 5 – Bottom Panels, orange triangles and purple squares, respectively). In the Discussion, we provide an account for this pattern of results.

Third, for Gaussian distributions T1ERs were higher than 5% when a large number of simulated participants were excluded from analysis. This was most often observed for simulations with few trials. As described in the Method section, participants were excluded when fewer than three operating points could be extracted in either of the two conditions. As a result, the *actual* sample sizes (i.e., sample sizes de facto) were smaller than the *simulated* sample sizes (i.e., the sample size that was set in the computer simulation). See Appendix C for a detailed tabulation of the actual sample sizes as a function of the simulated samples sizes and as a function of number of trials across participants.

Figure 6 depicts the T1ER as a function of the actual sample size (as a proportion of the simulated sample size) and as a function of criteria placement. Examination of the figure reveals that overall, decreases in the proportion of the actual sample size out of the simulated sample size are associated with a higher frequency of T1ERs greater than 5%. Note, however, that high T1ERs were typically observed for conservative criteria-placement which produced a small proportion of actual sample size (out of simulated sample size). Also note that even when considering simulations where the mean actual sample size was small, higher T1ERs were observed only for a small number of those simulations with most simulations associated with T1ERs of

approximately 5%. In fact, most simulations did not include exclusions in any of the 10,000 iterations[15].

In Appendix C, we investigate the actual sample size, presenting the effects of simulated sample size and the number of simulated trials per condition on the median size of the actual sample size. In Appendix D, we explore an additional pattern of results, testing the computation of S for different Gaussian simulations. We show how the overall similarity between S and the true SD ratio is high, and how a decrease in sample size were associated with estimates of S that deviated to greater extents from the true SD ratio. We return to these results in the Discussion.

[15] When exploring the parameter space in simulations, levels of different variables can create extreme experimental scenarios (see Discussion, Footnote 16, for an example). These scenarios are usually very unlikely in empirical settings. If a pattern of results is not observed across several levels of distributional and criteria variables, it should be taken with a grain of salt.

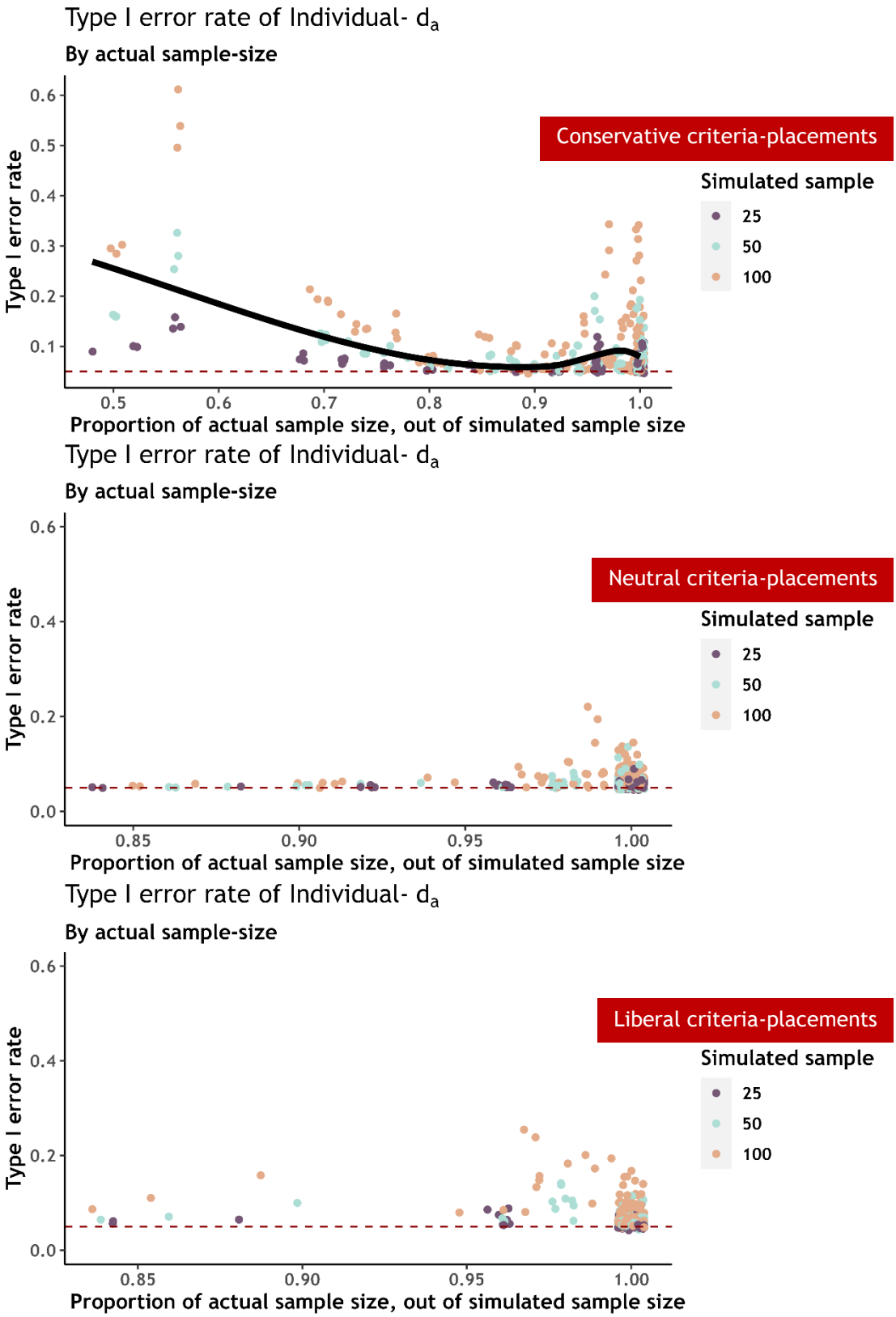

Figure 6. Type I error rate of d_a, as a function of the proportion of actual sample-size out of the simulated sample size. All the simulations described in this figure are Gaussian. Conservative criteria-placements are presented in the top panel, Neutral criteria-placements are presented in the middle panel, Liberal criteria-placements are presented in the bottom panel. Different sample sizes are depicted as different colors.

Discussion

We ran Monte-Carlo simulations and explored the parameter space of possible data-generating models to investigate the validity of common SPSMs (Pr, A′ and d′) in indexing sensitivity, as well as of the multi-point measure of d_a. The simulations revealed that d_a is the best contender for indexing sensitivity in recognition memory. *For Gaussian lure and target distributions, T1ERs of d_a were approximately 5% across different sample sizes, number of trials, response criteria, and distances between distributions.* Importantly, the 5% T1ER was found even for unequal-variance Gaussian distributions, the data-generating model that many researchers argue best reflects recognition memory. Assuming that for recognition memory the lure and target distributions are Gaussian with an SD ratio of 0.8, d_a is the only sensitivity measure that is demonstrably able to index recognition data in a valid manner (Figure 5, bottom-left panel).

Unlike d_a, more often than not, SPSMs showed T1ERs that were higher than 5%. This pattern of results replicated previous simulations (Rotello et al., 2008) that found T1ERs higher than 5% for several SPSMs. The only exception was when the measurement model was congruent with the data-generating model. Specifically, d′ was associated with only 5% T1ERs when measuring sensitivity in data from equal-variance Gaussian distributions. Likewise, T1ERs were 5% when Pr was used to measure data from equal-variance Rectangular distributions. However, *whenever the data-generating models were incongruent with the measurement models, T1ERs were higher than 5%, reaching 100% with increases in the number of trials or in sample size.* Notably, for unequal-variance Gaussian distributions with SD ratio of 0.8—arguably the true data-generating model for recognition memory—T1ERs for all SPSMs were higher than 5%.

Both here and in previous simulations (Rotello et al., 2008), the high T1ERs obtained for SPSMs show that computing sensitivity measures when the data-generating model is incongruent with the measurement model can lead to wrong inferences about the data. Note that the fact there is a high positive correlation between results from Pr and d′ does not mean we can use Pr for equal-variance Gaussian distributions and expect the valid results obtained by using d′. Using a wrong sensitivity measure can be thought of as analogous to using a wrong measure of central tendency when performing a t-test. Thus, for example, although means and medians are highly correlated, replacing means with medians will yield invalid conclusions in a t-test.

Like Rotello et al. (2008), we too found a positive association for SPSMs between the T1ERs and both sample size and number of trials. This association is a cause of great concern. To increase statistical power, researchers recruit large samples, as recommended (Smith & Little, 2018) following the replication crisis (Open Science Collaboration, 2015). Our results suggest that large sample sizes, while increasing power, can also increase the rate of Type I errors in recognition memory. Therefore, finding a measure that has a measurement model congruent with the true data-generating model of recognition memory is vital for the ability to sustain high-powered, reproducible, and (most importantly) valid recognition research.

To compute d_a, we simulated confidence ratings. From the confidence ratings, we were able to estimate the ratio of the lure-to-target SD (To be referred to as "the SD ratio"). The values of the SD ratio and their estimation in the simulated data (S) were approximately the same, in that the mean S values in simulations deviated from the true SD ratio by approximately 0.05 or less. The largest deviations of mean S from the true SD ratio were observed in simulations in which the placement of criteria was

conservative (see Appendix D). Indeed, conservative criteria-placement simulations were also the simulations in which most cases of T1ER higher than 5% were observed for d_a (see Results, Figure 6).

Presumably, because of the location of criteria, conservative placements yielded more FAR = 0% than other placements, yielding fewer than five operating points for many participants and therefore a less accurate estimations of the SD ratio – less accurate S[16]. Because the value of S was based on less data and was therefore less accurate, the ensuing estimates were more distant from the true SD ratio. Eventually, less accurate S leads to a larger discrepancy between the simulated data-generating model and the measurement model of d_a (which is based on S), therefore, to higher T1ERs (see explanation below).

Nevertheless, in most simulations, the values of S were similar to the true SD ratio and seemed to provide a good estimate for the computation of d_a. This result is not surprising in that the simulations did not include any noise and as such were ideal for the estimation of any parameter, including the true SD ratio. Future research is required to test the estimation of S using empirical data, where the control over noise is limited.

Importantly, to get the multiple operating points necessary for the estimation of S, confidence ratings are needed. It is not sufficient to obtain binary old-new

[16] In the conservative placements, the middle criterion was set at 0.5. The mean of the lure and target distributions were set at equal distances from 0 (see Methods). Therefore, in large-distance simulations (distance = 2 SDs relative to the lure distribution), described in the Results as the simulations that included the largest deviations of S from the SD ratio, the mean of the lure distributions was set at -1. If the mean of the lure distribution is -1, many signals sampled from this distribution are around -1. If the middle criterion (C3; see Methods) was set at 0.5 for the liberal condition and shifted by 0.3-0.7 SDs (relative to the lure distribution; see Methods) for the conservative condition, many signals will not exceed this criterion, resulting in much fewer "old" responses for lures. Ultimately, a large portion of participants will have a 0% FAR for at least one operating point. That is, they will have fewer operating points for the computation of S. This was not the case for the liberal criteria placement- the middle criterion was -0.5, but the mean of the lure and target distributions were like the conservative placement, therefore a larger portion of participants had all five operating points.

judgments. Researchers might be reluctant to move from binary responses to confidence responses. Indeed, one lab has reported recognition accuracy to be lower following responses given on a six-point confidence scale as compared to binary responses (Benjamin et al., 2009). Still, re-analyses of those data revealed no difference in accuracy between the two models of responding (Kellen et al., 2012). We believe that the requirement for using confidence ratings should not turn researchers away from using d_a, a measure that we have demonstrated here to be the only valid sensitivity measure for recognition memory.

Against the backdrop of most Gaussian simulations that revealed 5% T1ERs with d_a, higher rates were found under two scenarios. Both scenarios (described below) have a common feature – the less accurate estimation of S under specific conditions of our simulations. The first, for large sample sizes (N=100) and a small number of trials (64 trials), T1ERs were higher than 5% (see Appendix A). The high T1ERs for many participants and few trials can be explained in the following way. For a measure to be valid and have 5% T1ER, the measurement model must be congruent with the data-generating model. One element of the measurement model of d_a is the SD ratio, estimated by S. If participants have few responses, for example, because of a low number of trials per condition, then the estimate of S will be less accurate (see Figure 6 and Appendix A) and consequently, the data-generating model will be less congruent with the measurement model. Though the discrepancies between the data-generating and the measurement models are likely very small, even small discrepancies can yield significant differences in highly powered experiments, such as experiments with large sample sizes. It is for this reason, that T1ERs higher than 5% were found for the conjunction of large sample size with few trials.

If a small number of trials can lead highly powered experiments to show false significant results at a rate higher than 5%, a larger number of trials might help in reducing the T1ER back to 5%. By running more trials for each condition, a more accurate SD ratio estimation is obtained, resulting in higher congruency between the data-generating model and the measurement model. Therefore, T1ERs will likely be reduced to the desired 5%. This is indeed the case, as we discovered in our results. When the number of trials was higher (128 or 256 trials), T1ERs were approximately 5% for d_a, even with large sample sizes (see Results, Figure 5). We assume that d_a can still be used in large-sample experiments so long as the number of trials each participant undergoes is closer to 128 rather than 64. Fortunately, that is already the case for most recognition-memory studies where researchers typically run tasks with considerably more than 64 trials per condition.

The second scenario under which high T1ERs for d_a were observed was when the actual sample size was considerably smaller than the simulated sample size (see Results). The higher T1ERs seem, again, to be the outcome of the less accurate parameter estimation (S). As before, this pattern was observed primarily for simulations with a small number of trials.

The two exceptions notwithstanding, we argue that d_a is a better alternative than SPSMs for measuring sensitivity and is, in fact, a very good alternative. This measure is bias-independent for experiments with a large number of trials, an attribute of the experiment that is often under the researcher's control. When such experiments are run, a cautionary signal for T1ERs higher than 5% is a large portion of exclusions. Results with many exclusions should be viewed with caution.

Importantly, the use of S introduces a data-based estimation of a parameter (the true SD ratio) into the calculation of d_a. As the estimation is based on five (or fewer)

operating points, the SD ratio is often not estimated precisely by S. When calculating d_a, if we could have access to the true parameter, we suggest that the ensuing T1ERs would not be higher than 5%. This suggestion is supported by the results of equal-variance Gaussian distributions simulations. We compared T1ERs for d′ which has a built-in assumption of SD ratio = 1 (which is a true assumption for those simulations) to d_a, which estimates the SD ratio from the data via S (and inevitably obtains slightly inaccurate estimates for some participants in each simulated experiment). This comparison revealed that for most of those simulations, d′ is associated with lower T1ERs than d_a. Unfortunately, outside of such simulations, we don't have access to the true population SD ratio.

To evaluate the implications of our simulations to future recognition experiments, we wish to point out an asymmetry in the potential of simulations to advance the validity of any measure. Simulations alone provide a robust tool in showing a measure to be invalid, in our case—in establishing the SPSMs are not bias-independent. However, to establish that a measure, any measure, is bias-independent – simulations are not enough.

For SPSMs, simulations represent possible experimental scenarios under which T1ERs are higher than 5%. Because there is a possibility that such experimental scenarios do exist (e.g., there is a strong possibility that the lure and target distributions are indeed unequal-variance Gaussian), and because of the severity of Type I errors in impairing our ability to infer from our data, these results should be sufficient for researchers to renounce those measures and seek alternatives.

For d_a, however, 5% T1ERs in simulations are not sufficient to prove this measure is valid. Simulations always include overt and covert assumptions regarding the computer-generated data. Specifically, in our simulations, we knowingly made

assumptions regarding the data-generating model of recognition memory. As such, we ran simulations based on Gaussian (both equal and unequal variance) and Rectangular distributions. d_a only showed low T1ERs when data were simulated from Gaussian distributions. The debate in the recognition literature regarding the data-generating model of recognition memory (e.g., Mickes et al., 2007; Rouder et al., 2010; Wixted & Mickes, 2010) is ongoing, and results from our simulations are orthogonal to the question of the true recognition memory data-generating model. Thus, for example, our simulations reveal that if the true data-generating model is Rectangular, d_a will not be valid. Because we have only indirect access to the true data-generating model in recognition memory, the only way to guarantee the validity of d_a will be by analyzing real-world empirical data.

Moreover, many other aspects of data generation in our simulations were based on simplistic assumptions. For example, the confidence ratings in each trial were sampled independently. However, confidence carry-over effects have been demonstrated, such that the confidence in trial N is correlated with the confidence in trial N-1 (Kantner et al., 2019). This confidence carry-over effect was not taken into consideration in our sampling process, and we do not know the extent to which such carry-over effects will affect the validity of d_a. As such, in an empirical setting, there can be many other effects we are unaware of. We do not know the extent to which these effects impact the validity of our measures (Treisman & Faulkner, 1984; Van Zandt, 2000; For examples in perceptual tasks, see Treisman & Williams, 1984). The only way to prove the validity of d_a is to put it to the test using empirical data from iso-sensitive conditions.

While future research is needed to establish the validity of d_a, our current investigation points at a promising direction in support of a little-known sensitivity

measure for measuring accuracy in recognition tasks. Importantly, as a replication of previous simulations (Rotello et al., 2008), our results also confirm the importance of choosing a valid measure, not only in recognition memory but for other cognitive tasks as well. The choice of a measure must always consider the data-generating model of the task, and the fit between the data-generating model and the measurement model, so conclusions can be drawn from valid results. The choice of a valid measure is an inherent part of the manufacturing processes of empirical data.

Bibliography

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*(1), 84-115.

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2022). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review, 30*(2), 421–449.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294.

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14*(3), 423–429.

Dubé, C., & Rotello, C. M. (in press). Signal detection theories of recognition memory. In J. T. Wixted (Ed.), *Learning and Memory: A Comprehensive Reference, 3rd edition* (Vol. 4: Memory and Cognition). Elsevier.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics* (2nd ed.). Peninsula Publishing.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2022). *Detection theory: A user's guide* (3 ed.). Routledge.

Hautus, M. J., Macmillan, N. A., & Rotello, C. B. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review, 15*(5), 889–905.

Kantner, J., Solinger, L. A., Grybinas, D., & Dobbins, I. G. (2019). Confidence carryover during interleaved memory and perception judgments. *Memory & Cognition*, *47*(2), 195–211.

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review*, *119*(3), 457-479.

Levi, A., Rotello, C. M., & Goshen-Gottstein, Y. (2019, November 16). *It's Really tough to Measure Discriminability: Frequency of Type I Error Rates for the (Geometric) Area Under the Roc Curve*. 2019 Psychonomics, Montreal, Canada.

Levi, A., Mickes, L., & Goshen-Gottstein, Y. (2021). The new hypothesis of everyday amnesia: An effect of criterion placement, not memory. *Neuropsychologia, 166,* 108114–108114.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*(2), 100–109.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A user's guide* (2 ed.). Psychology Press.

Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & psychophysics*, *66*(3), 406-421.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93–102.

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*(4), 361-376.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*(5), 858–865.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

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.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1–12), 125–126.

Ratcliff, R., Sheu, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518–535.

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.

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember-know judgments using signal detection and threshold models. *Memory & Cognition, 34*(8), 1598–1614.

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*(2), 389–401.

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin & Review*, *17*(3), 427–435.

Schooler, L. J., & Shiffrin, R. M. (2005). Efficiently measuring recognition performance with sparse data. *Behavior Research Methods*, *37*(1), 3-10.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*, 2083–2101.

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.

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*(1), 99-139.

Starns, J. J., Rotello, C. M., & Hautus, M. J. (2014). Recognition memory zROC slopes for items with correct versus incorrect source decisions discriminate the dual process and unequal variance signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1205–1225.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, *99*(1), 100-117.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory Academic.* Academic press.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409.

Treisman, M., & Faulkner, A. (1984). The effect of signal probability on the slope of the receiver operating characteristic given by the rating procedure. *British Journal of Mathematical and Statistical Psychology, 37*(2), 199-215.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological review, 91*(1), 68-111.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 582-600.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233.

Wixted, J. T., & Mickes, L. (2010). Useful scientific theories are useful: A reply to Rouder, Pratte, and Morey (2010). *Psychonomic Bulletin & Review*, *17*(3), 436–442.