# Stop using d′ and start using $d_a$: Part II. Empirical Recognition Memory Data Reveal Type-I Error Rates of Different Sensitivity Measures

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

The selection of an accuracy (sensitivity) measure is a pivotal decision faced by recognition-memory researchers. Despite an abundance of measures developed over the years with the intention of measuring sensitivity independent of participants' bias, Monte Carlo simulations found all commonly used sensitivity measures to be invalid because they are confounded with bias (Rotello et al., 2008; Levi et al., 2024). The one valid measure was our proposed version of the signal-detection measure, d-sub-a (denoted $d_a$). Empirical confirmation in real-world experiments is critical, however, to establish the validity of any measure, including $d_a$. The goal of the current investigation was to use empirical data to test the validity of $d_a$, as well as other common sensitivity measures, as indices of sensitivity. We ran a large-scale recognition experiment, where sensitivity was not manipulated at encoding. Using implied base-rate, bias was manipulated at test. For valid measures, erroneous significant results should be observed at a rate of approximately 5% when testing iso-sensitive conditions. We repeatedly ran the experiment to derive *rates* of Type I errors. $d_a$, but no other measure (Pr, A′, and d′), demonstrated the desired 5% false-discovery rate across different sample sizes. Together with the results of our simulations, these new findings provide unassailable evidence that $d_a$ is a valid sensitivity measure for recognition-memory tasks, offering an easy solution to the prevailing measurement crisis in recognition memory.

## **Introduction**

Designing and employing a recognition-memory test (or, for that matter, any cognitive test) is much like designing and employing a production process in an industrial plant. It is a well-thought-out process, requiring the planning and evaluation of multiple factors prior to manufacturing the goods: choosing the category of study items—words, objects, faces; choosing the specific items to be presented during the task; deciding list length and stimulus presentation durations; deciding how to measure participants' decision accuracy. Mistakes in any of these steps can result in deficient outcomes. In the manufacturing context, for example, allowing tiny grains of sand to slip into the assembly line rather than maintaining a sterile environment may result in non-functional electronic chips.  Analogously, overlooking the tiniest detail in a cognitive testing process may render the final product—the experimental conclusions – completely worthless. Alas, this seems to be the case for recognition memory.

Here we focus on a decision that researchers often assume to be trivial—the choice of a measure of recognition-memory accuracy. Different researchers may choose different measures to index accuracy. Indeed, even the same researcher may sometimes use different measures across different studies. A justification for these different choices is rarely, if ever, provided. Unfortunately, it turns out that the decision of which measure to use may well be the equivalent of the 'grain of sand' in the recognition memory task. A wrong choice of measure may yield false discoveries brought about by an alarmingly high probability of Type-I errors.

The problem of choosing appropriate measures is not uniquely related to the recognition memory task. Examples of similar 'grains-of-sand' in other research domains include the sequential-superiority effect (in Eyewitness memory research), the belief bias effect (in Reasoning research), shooter bias (in Social psychology), and even referrals for child maltreatment (Rotello et al., 2015). While the measures are different, the main concern remains the same – when researchers choose the wrong measures, false scientific findings can be discovered. In this article, we focus on the problem of measurement in recognition memory which is relevant to tasks used across several domains of research, including cognitive psychology, social psychology and experimental psychopathology (e.g., recognition-memory tested on patients with schizophrenia, Danion et al., 1999; recognition-memory tested in social psychology to explore cross-race facial recognition, Hehman et al., 2010).

We collected empirical data from a recognition task to investigate the validity of different measures (i.e., the rate of Type-I errors), showing whether they are good or poor choices for indexing recognition memory data. These empirical data complement computer-generated data obtained in previous simulations (Levi et al., 2024), with the hope of converging on a single valid measure for recognition memory. To understand our predictions and results, we must clarify the basic concepts that led us to the current project. To this end, we first define the measured psychological trait, accuracy, in the context of the recognition task (see "Measurement of the recognition-memory performance"). Second, we briefly summarize the findings from previous simulations that explored measurement problems in recognition memory (see "Precursors to the current project: Monte Carlo simulations"). Third, we discuss the assumptions of different accuracy measures and compare them to the underlying models previously suggested for recognition memory (see "SPSMs are invalid measures: Incongruence between measurement models and data-generating models"). Finally, we explain why the validity of a measure cannot be proved via computer simulations.  Rather, such simulations must be supplemented by empirical data (see "The essential role of empirical data").

## Measurement of recognition-memory performance

In the recognition task, participants are presented with single items during test and are asked to make binary judgments regarding their status in the experiment (Hautus et al., 2022): are they ‘old’ (presented during the study phase) or ‘new’ (not previously presented). The recognition task is but one example, representing a wider set of tasks known as 'single-interval tasks' (Hautus et al., 2022). These tasks share the common feature that a single item is presented at test, and a binary decision is required from participants, with one choice being objectively correct (e.g., objectively, the item was either previously presented or not). While we focus on recognition memory, our arguments in this article can easily be generalized (with caution) to most single-interval tasks, including meta-cognition tasks such as the judgment of learning task, perceptual tasks such as the visual-discrimination task and attention tasks such as the spatial-cueing paradigm.

Similar to other single-interval tasks, accuracy performance in recognition tasks is indexed by the degree to which participants are sensitive to the difference between old items (targets) and new items (lures) during a test phase, and is therefore referred to as ‘*sensitivity’* (Hautus et al., 2022). If all aspects of the experimental ‘manufacturing process’ are well thought out, a difference in sensitivity between conditions can be attributed exclusively to changes in memory and can allow the researcher to advance conclusions regarding the underlying mnemonic mechanisms.

However, participants’ judgments may be affected not only by mnemonic changes across the conditions but also by their preference for one binary choice over another— their ‘response bias’ (Hautus et al., 2022). A good measure must index sensitivity irrespective of bias—to what degree participants are liberal or conservative in their proclivity to respond ‘old’. That is, if two conditions are iso-sensitive, a significant difference between the conditions should be observed in only about 5% of the cases (assuming a statistical $\alpha = .05$) even if participants respond with a different level of bias in each condition. We can then say the measure is ‘bias independent’. If a measure is not bias-independent, then sensitivity and bias are confounded, and researchers cannot infer changes in sensitivity from significant differences in accuracy.

To illustrate, consider the fact that for a condition with truly higher sensitivity, higher rates of correct identifications of studied items will be observed (e.g., a higher hit rate, HR – the percentage of correct ‘old’ responses to targets, out of all responses to targets). Still, the measure of HR is not bias-independent because higher levels of HR may just as easily reflect a more liberal response bias, irrespective of whether the item was a target or a lure.

Numerous measures have been developed over the decades with the intention of measuring sensitivity independent of bias. These measures offer different ways to balance HR with the false alarm rate (FAR—the percentage of incorrect ‘old’ responses to lures, out of all responses to lures). Examples of such measures are Pr = HR - FAR, $A'$ and $d'$ (for details, see below; Levi et al., 2024). Because these measures are all based on a single pair of (FAR, HR) observations, they are called *single-point sensitivity measures* (SPSMs).

In this article, we provide unequivocal evidence that SPSMs are invalid indices of sensitivity in recognition memory because they are not bias-independent. To this end, we present **empirical results** obtained using a novel methodology for gauging an estimate of Type-I error rates (T1ERs) in empirical experiments. The methodology entails repeatedly running an experiment wherein the null hypothesis is known to be true and calculating the T1ER across the many replications of the experiment. Importantly, using our methodology, we show how a different measure, $d_a$, is associated with accepted levels of T1ERs. We will establish, therefore, that $d_a$ is bias-independent and argue that it should be adopted as the gold-standard to measure accuracy in recognition memory.

**Precursors to the current project: Monte Carlo simulations**

Theoretical arguments against indexing sensitivity with the SPSMs were initially advanced by Swets (1986a; Rotello et al., 2015). The conclusions against using SPSMs were cemented by Monte Carlo simulations (Rotello et al., 2008; Levi et al., 2024; cf., Schooler & Shiffrin, 2005). In the simulations, for each participant, two iso-sensitive conditions were created by using the same data (sampled only once from the lure and target distributions) in the two conditions. This resulted in a within-participant design. The two iso-sensitive conditions differed in the placement of the response criterion (bias). Though calculated from the same data sample, the difference in bias between conditions (see Levi et al., 2024) resulted in different HR and FAR values. Note that because no difference in sensitivity was simulated, valid measures should yield significant results at a rate of 5% (reflecting the Type–I error rate of $\alpha$ =5%, the rate used throughout this article).

Different SPSMs were used to measure sensitivity. A within-participant t-test for each experiment subsequently tested for significant differences between the conditions. By manipulating the values of several parameters across simulations, multiple experimental scenarios were generated. The parameters included the form of the lure and target distributions, sample size, number of trials per condition, distance between the means of the lure and target distributions, and criteria placement (i.e., bias). The 2008 simulations tested percent correct, $\gamma$, $\gamma_c$, A′ and d′. Our recent simulations examined Pr, A′ and d′.

A high rate of false significant results was apparent across most parameter combinations in the simulated experiments. For example, in the recent simulations, SPSMs showed erroneous significant differences between the conditions, in rates higher than 5% (see Figure 1), for almost all assumed forms of target and lure distributions (i.e, equal-variance Gaussian distributions, unequal-variance Gaussian distributions, equal-variance Rectangular distributions). Exceptions to the high rates of significant results are described below.

In addition, and *counterintuitively*, as the sample size or the number of trials increased, T1ERs reached levels as high as 100% (Rotello et al., 2008; Levi et al., 2024). The gravity of this finding cannot be overstated—the replication process does not guarantee that the herd of false discoveries will be culled. When discoveries are false, they can be replicated easily and reliably. Alas, SPSMs—the very measures that are typically used in recognition research—likely produce false discoveries not only in simulations but also in the empirical literature. The high incidence of false discoveries led to a recent proclamation that memory researchers are undergoing a 'crisis of measurement' (Brady et al., 2022; also see Rotello et al., 2015).

Recently, we ran simulations of our version of the lesser-known sensitivity measure d-sub-a (Levi et al., 2024; see below for a detailed account of this measure). Across many experimental scenarios, $d_a$ was associated with T1ERs approximating 5% (see Figure 1). This opens the possibility for a potential solution to the measurement crisis. However, this possibility required verification in empirical setting. In this article, we describe a large-scale recognition memory test where we empirically investigated bias independence of common SPSMs and, more importantly, of $d_a$. We explored real-world data because of possible differences that may exist between the actual empirical recognition memory data and the assumptions made when running simulations about how data are generated, as well as how recognition is performed. We elaborate on this below.

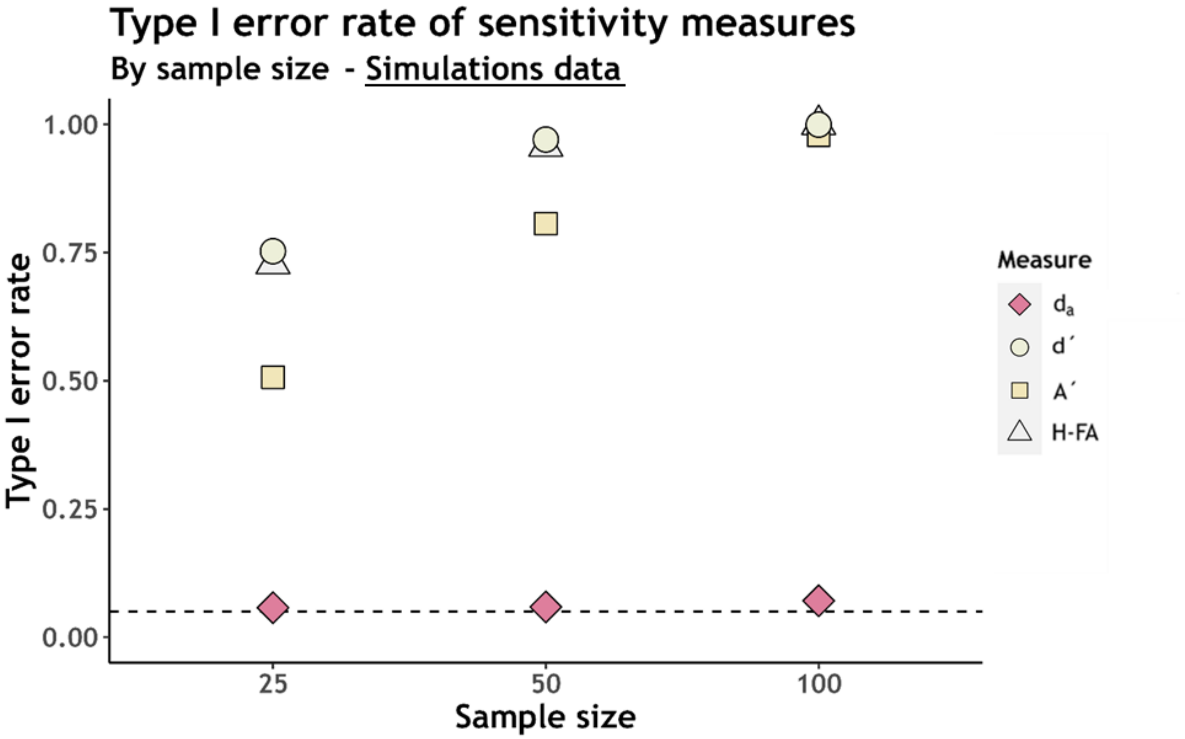


**Figure 1**
Type-I error rates as a function of sample size and sensitivity measure, using simulation data (Levi et al, 2024). The simulations sampled data from unequal-variance Gaussian lure and target distributions, with the target distribution being 25% wider than the lure distribution (0.8 lure-to-target standard-deviation ratio, see below for more details). The simulations included 128 trials per condition. All other assumptions made for the simulations are detailed in Levi et al. (2024, see Figure 3).

In the next sections, we introduce the terms ‘measurement model’ and ‘data-generating model’. We then argue that in recognition, these two models are incongruent for all SPSMs and that because of this incongruity, all SPSMs cannot be bias-independent measures of recognition. We then suggest that $d_a$ may likely be congruent with the recognition-memory data-generating model and thus provides a promise of bias independence.

**SPSMs are invalid measures: Incongruence between measurement models and data-generating models**

For scientists to describe systematic laws of nature, they assume that data are generated from some distribution and do not come into the world arbitrarily, in a haphazard way. To make valid inferences scientists must uncover the nature of those distributions. The term 'data-generating model' refers to the form of the distribution from which the data are generated (e.g., Gaussian, Rectangular) and its associated parameters or assumptions (e.g., equal variance; unequal variance). Therefore, a valid measure of a task must be developed under assumptions that are congruent with the data-generating model[1].

The different SPSMs were developed under different assumptions regarding the nature of the data-generating model[2]. The set of assumptions upon which a measure is premised is its 'measurement model'. The simulations described above found all SPSMs to not be bias-independent (yielding high T1ERs) when their measurement models were incongruent with the simulated data-generating model. This included Pr[3], A′, and d′ (Rotello et al., 2008; Levi et al., 2024). The conditions in our simulations for which T1ERs were curtailed at 5% were those for which the measurement model and data-generating model were congruent. For example, d′ has a

[1] In recognition memory, the data-generating model cannot be retrieved from the binary responses of participants without manipulating bias between conditions, therefore the choice of an appropriate measure is quite an arduous task (Rouder et al., 2010; but see Wixted & Mickes, 2010).

[2] **Pr** (also referred to as "corrected hit probability" or "corrected hit rate" or "corrected recognition") was developed under a double high-threshold measurement model (Pazzaglia, Dube, & Rotello, 2013; Hautus et al., 2022). This model assumes that mnemonic information exists in a discrete, binary state. **A′** was developed with the intention of offering a non-parametric alternative for indexing sensitivity (Pollack & Norman, 1964). However, it turns out that the measurement model of A′ entails the assumptions that the lure and target distributions are rectangular or logistic (depending on the level of sensitivity; Macmillan & Creelman, 2004). **d′** is a signal-detection measure of sensitivity, developed under the assumptions of continuous, equal-variance Gaussian lure and target distributions (Tanner & Swets, 1954).

[3] Rotello et al. (2008) investigated percent correct. Our recent simulations (Levi et al., 2024), however, tested Pr due to its greater popularity in the recognition literature. Importantly, if two measures are linear transformations of one another, they will have an identical measurement model. In the case of Pr, it is a linear transformation of percent correct, $P_c$= (HR + CRR)/2, where CRR is the correct rejection rate (Hautus et al., 2022). Because CRR = 1- FAR, testing Pr or percent correct is interchangeable with regards to our simulations.

measurement model of equal-variance Gaussian lure and target distributions. Indeed, although d′ was not bias-independent under most scenarios, when the simulated data were generated from equal-variance Gaussian distributions, T1ERs were only 5%.

The crux of the measurement crisis in recognition memory is that much evidence supports the idea that recognition memory is mediated by *un*equal-variance Gaussian distributions (known as the unequal variance signal detection, UVSD, model; see Ratcliff et al., 1992). This is challenging because no SPSM has a measurement model of unequal-variance Gaussian distributions. To reiterate, the simulations reveal that if the data-generating model of recognition is unequal variance Gaussian, all SPSMs are not bias-independent, with T1ERs as high as 100%. What is the evidence in support of the UVSD model as the data-generating model?

While it is difficult to ascertain the data-generating model from the rates of hits and false alarms, sophisticated tools have been developed to point out good potential candidates. One such tool is the receiver-operating characteristic (ROC) curve. The ROC curve plots multiple (FAR, HR) points, known as ‘operating points’ (Green & Swets, 1988), displaying the HR on the ordinate and the FAR on the abscissa. Researchers typically compute multiple operating points by asking participants at test to rate their confidence that an item was ‘old’, for example, on a scale that ranges from 1 to 6 (with '1' indicating high confidence in the item being new and '6' indicating high confidence in the item being old; Hautus et al., 2008). Using confidence ratings, different bias cutoffs can be established[4]. This allows the computation of five unique operating

[4] The most liberal cutoff involves treating responses of '1' as 'new' and comparing them to responses of '2' to '6', which are considered 'old'. By computing HRs and FARs using this scheme, one can assess performance of each participant across all responses. A slightly less lenient cutoff scheme involves considering responses of '1' and '2' as 'new' and comparing them against responses of '3' to '6', which are regarded as 'old'. The most conservative cutoff involves regarding '1' to '5' as 'new', and only '6' responses as 'old'.

points, each with a different level of bias but the same level of sensitivity (Mickes, 2015; Hautus et al., 2022).

Different measurement models make unique predictions regarding the shape of the theoretical ROC curve (Swets, 1986a), and some make predictions regarding the zROC curve. The zROC curve is the transformation of the ROC curve into z-space. A linear zROC curve is predicted if the underlying lure and target distributions are Gaussian, and the slope (S) of that linear curve is an estimate of the ratio of the lure-to-target standard deviations (SDs). That is, if the distributions are equal variance, then S=1 (Hautus et al., 2022). Importantly, in support of the Gaussian assumption, a linear zROC has repeatedly been documented (Wixted, 2007, for a review see Wixted 2020; Dube & Rotello, in press). Note, however, that a linear-like zROC curve does not *prove* that the underlying distributions are Gaussians (Lockhart & Murdock, 1970).

The linear zROC slope in recognition memory is consistent with continuous, Gaussian underlying distributions (Swets, 1986b; Wixted, 2007; Wixted & Mickes, 2010) as assumed by signal-detection theory (SDT). Such a slope is inconsistent with the measurement models of Pr and A′ regarding the lure and target distributions. Still, a linear slope does not refute other theoretical frameworks (Rouder et al., 2010). Importantly, if the lure and target distributions are indeed Gaussian, the assumptions of both Pr and A´ are violated while one of two assumptions of d´ is met.

The second assumption under which d´ is premised is that of equal-variance of the lure and target distributions. It turns out, however, that under the assumption of Gaussian distributions, the equal-variance assumption is blatantly wrong (Mickes et al., 2007; Wixted & Mickes, 2010). To reiterate, conditional on the distributions being Gaussian, if the underlying distributions are equal variance, S is equal to 1. Yet, for different data sets, the mean magnitude of S in old-new recognition has consistently been found to be at least approximately equal to 0.8 (Ratcliff et al., 1992; Dougal & Rotello, 2007; Mickes et al., 2007). Thus, the equal-variance assumption of the measurement model of d´ is violated in recognition memory, where (conditional on Gaussian distributions) the underlying distributions are of unequal variance.

Our working hypothesis in this article, based on extant knowledge, is that the UVSD model is the recognition memory data-generating model. In contrast to SPSMs, the measure we have recently investigated (Levi et al., 2024), $d_a$, (Swets & Pickett, 1982), is congruent with the UVSD model. $d_a$ is derived from signal-detection theory. Computed in the way we propose, $d_a$ is a multi-point sensitivity measure – it relies on more than one operating point. Specifically, $d_a$ relies on at least three unique operating points (see below for further details). Like d´, the measurement model for $d_a$ assumes Gaussian lure and target distributions. Importantly, $d_a$ does not assume equal variances of these distributions.

The measure $d_a$ is based on the subtraction of Z-transformations of HR (denoted as H in the formula) and FAR (denoted FA). It 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), and can be shown to equal:

$$\boldsymbol{d_a = \sqrt{\frac{2}{1+\frac{SD_{lures}}{SD_{targets}}^{2}}}\left(z(H) - \frac{SD_{lures}}{SD_{targets}} \cdot z(FA)\right)} \qquad \textbf{(1)}$$

The use of SD ratio allows the measure to fit both equal and unequal Gaussian distributions models, but requires researchers to know the ratio of variances for their data. Note that if the lure and target distributions are equal variance, then $d_a$ equals d′. Importantly, *For the estimation of the SD ratio, we suggest using S.*

In recent simulations, we estimated S individually for each participant and computed $d_a$ accordingly. By using each participant's zROC slope, we predicted $d_a$ to be consistent with the UVSD model, presumably the data-generating model of recognition memory for each individual. To calculate $d_a$, we simulated data from six-point confidence ratings. To achieve this, we placed five equally spaced criteria corresponding to six confidence levels (Levi et al., 2021). The criteria were placed at different locations, depending on the simulation. We then computed operating points for the ensuing zROC curves and estimated S (see Footnote 4), separately for each participant using the mean value of S across the two conditions. It is this value of S that we used to calculate $d_a$ (Levi et al., 2024).

The individual estimation of S per participant was performed to account for documented variability in the magnitude of the SD ratio across participants. Recognition data reveal that participants may have different data-generating models, and that the reported SD ratio of 0.8 is only a mean estimate, with zROC slopes being highly variable (Macmillan et al., 2004). Indeed, different experimental conditions yield different mean zROC slopes, with values ranging as low as 0.5 and higher than 1 (Ratcliff et al., 1992). The use of each participant’s individual slope was therefore a potential remedy for the individual differences in the data-generating model.

To reiterate, the simulations revealed that $d_a$ measured sensitivity independent of bias for both equal and unequal Gaussian distributions (see Levi et al., 2024, Figure 5). Specifically, when comparing two iso-sensitive conditions with different levels of bias, $d_a$ showed significant results approximately 5% of the time, as expected. This pattern was observed across various manipulations of strength, criteria placements, sample sizes, and number of trials. Although some deviations from this generalization were found, they were inconsequential[5] (see Levi et al., 2024). Therefore, based on our computer-generated data, $d_a$ seems like a potential solution for the measurement crisis in recognition.

However, it is far from trivial to suggest that estimating S individually for each participant, from five (or fewer) obtained operating points, will be accurate enough to produce a valid sensitivity measure. As we previously mentioned (see "The essential role of empirical data"), simulations are always premised on some assumptions. Specifically, in our recent simulations (Levi et al., 2024), participants' data were simulated from the same lure and target distribution (same data-generating model) across the different iterations. Based on the successful use of S (to estimate the SD ratio) in simulations, we turned to empirical data to seal our conclusions regarding the validity of our proposed version of $d_a$.

### The essential role of empirical data

Proving that a measure is valid in computer simulations is a necessary but not sufficient criterion for establishing its true validity.  Simulations necessarily make assumptions (both overtly and covertly) regarding the data-generating model and task features, all of which might be wrong. For example, our simulations assumed an underlying Gaussian distribution of the mnemonic signal.  Under this assumption, $d_a$ was found to be bias-independent. The veracity of this assumption, however, is under debate (Yonelinas & Parks, 2007; Wixted & Mickes, 2010; Brady et al., 2022).  If the Gaussian assumption is incorrect, $d_a$ would necessarily be invalid. Moreover, our simulations assumed that each participant has an identical data-generating model,

[5] Deviations from 5% T1ER for $d_a$ were inconsequential because they were mostly revealed under conditions where many simulated participants were excluded, and where the S deviated from the simulated SD ratio. For more data and explanations, see Levi et al., 2024.

with the same form, means, SDs, and placement of criteria. It is again unclear how well $d_a$ would perform for real-world data where each participant likely has a different model.

Finally, even if the mnemonic signal is truly distributed in a Gaussian fashion for recognition tasks, simulations may not mimic certain aspects of the experimental task. For example, in the simulations, confidence ratings were sampled independently on every trial. However, a confidence carryover effect has previously been suggested in perceptual tasks (Treisman & Williams, 1984), and more importantly demonstrated in empirical recognition data (Kantner et al., 2019), with high and low confidence judgments more likely to be followed by high and low confidence judgments, respectively. Such sequential effects are a cause of concern, because by distorting the true levels of confidence afforded by the items at test, they bias S such that it may not be a valid estimate of the population lure-to-target SDs (Treisman & Faulkner, 1984; Van Zandt, 2000). This would result in $d_a$ being an invalid measure.

Our goal in this article was to use real-world recognition data to test the validity of $d_a$ as a measure of sensitivity. To this end, we presented each participant with two identical recognition memory conditions and manipulated bias at retrieval across the two conditions. Critically, any single study cannot provide an estimate of a T1ER, because for any single study, a finding is either significant or not. Therefore, we repeatedly ran the same experiments over and over (see Methods) to gauge the rate of type I errors. To the best of our knowledge, no study has been run multiple times to gauge statistical error rates, be they type I errors or type II errors. This methodology is novel in and of itself and is crucial for establishing the validity of any measure, and specifically, that of $d_a$.

We manipulated bias by using an implied base-rate manipulation, which has been successful in shifting participants' response bias (Rotello et al., 2006; Dube & Rotello, 2012). We informed participants that the rate of old items at test was higher / lower than 50% (see Methods). We did not change the objective rate of targets and lures between conditions, as such changes may increase differences in the estimation of sensitivity due to changes in binomial variability across operating points (Dube & Rotello, 2012).

We made two predictions. First, we predicted that the T1ERs of common SPSMs would be higher than 5% and that these rates would increase with sample sizes, as observed in the simulations (Rotello et al., 2008; Levi et al., 2024). Second, we predicted $d_a$ to be a valid measure for the task, therefore we expected it to have a T1ER of approximately 5% when comparing the iso-sensitive conditions with $d_a$ (as observed in simulations as well, Levi et al., 2024).

## **Methods**

### Participants

In all, 1485 participants completed an online experiment on the Prolific platform. Participants were native English speakers from the UK or Ireland. They received 5.15 pounds Sterling in exchange for their participation. To encourage participants to focus on the task, participants with outstanding performance, defined as being in the top 20% (N =192), were given an additional 50% bonus. In total, 960 participants were included in our analysis and 525 were excluded from analysis for one of three *a priori* criteria listed in the Results.

### Design

We used a one-way design, with implied base rate (30% ‘old’, 70%’old’) as a within-participant variable. The task was old-new recognition memory. Each participant went through two study-test cycles, corresponding to the two experimental conditions. We manipulated bias (liberal condition, an implied rate of more ‘old’ responses; conservative condition, an implied rate of fewer ‘old’ responses) between the two cycles. We counterbalanced the cycles – half underwent a cycle with a liberal test condition first, and half underwent the cycle with the conservative condition first. No manipulation was carried out at encoding. Indeed, a manipulation check using ROC analysis revealed that sensitivity was in fact equal in the two conditions (see below, Results).

Each cycle included 70 items (63 words, 7 non-words) in the study phase and 126 words in the test phase. The non-words were included at study as catch trials, from which we could subsequently exclude participants who presumably did not pay attention to the study list and therefore, did not catch these trials (see below, procedure). Of the 126 test words, half were ‘old’ (presented at study in the specific cycle) and half were ‘new’ (had not been previously presented in the experiment)—yielding an objective base rate of 50% ‘old’ words.

Despite the objective 50% base rate, a base rate of either 30% (Conservative condition) or 70% (Liberal condition) was implied at test via retrieval instructions. Implied base-rate manipulations have been shown to affect bias but not sensitivity (Rotello et al., 2006; Dube & Rotello, 2012; see Manipulation Check in the Results). As mentioned, the order of implied rates was counterbalanced across participants.

**Stimuli**

In all, 252 words were randomly sampled without replacement from the "Toronto Word Pool" (Friendly et al., 1982) to be used as experimental stimuli. An additional 20 words were sampled for the practice session. All words were four- to eight-letters long. The 252 words were randomly assigned into four lists of 63 words. The words from the four lists were counterbalanced such that across participants, the words appeared an equal number of times as targets and as lures and an equal number of times in the liberal and conservative conditions. Assignment of participants to a particular lure-target and conservative-liberal combination of the four 63-word lists was random. For the practice session, we chose 10 words to be targets and 10 words to be lures. The same practice targets and practice lures were used for all participants.

We created 16 non-words to be used as catch trials.  The non-words were strings of English letters that resembled legal English words, but did not have a meaning in English. We began by establishing the validity of our non-words through pilot tests; participants underwent a recognition task, with non-words presented at study along with legal English words. Participants were asked to press a key every time a non-word appeared. Using the validated non-words, we also ran Pilot tests to define a-priori exclusion criteria[6] (for details, see Results).

Of the 16 non-words, seven were allocated to each of the two study phases. The non-words lists were counterbalanced across participants such that the non-words from each list appeared an equal number of times in the first and second study-test cycles, and in the conservative and liberal conditions (see Procedure). The final two non-words were inserted into the practice list and were always used for the practice session.

**Procedure**

We ran seven pilot tests, including 20 to 25 participants each, to find experimental parameters that would yield equal sensitivity and different bias in the two conditions. For each pilot test, results were analyzed using ROC plots[7]. The design we ended up using in our experiments is depicted in Figure 2.

At the beginning of the experiment, participants were informed that they would be presented with lists of words that they were to commit to memory for a later test. They were told that items would be tested on a six-point confidence scale and that they should respond in a manner reflecting their true memory.

Participants were then told that the rates of 'old' words might change for different study-test cycles. They were informed that these rates would be given prior to each test phase and that they

[6] The majority of participants in the pilot studies succeeded in responding to non-words and not responding to legal words at similar rates. We set the a-priori for excluding participants according to these success rates.

[7] We used ROC curves to obtain rough estimates of the difference in sensitivity and bias between conditions. Note, however, that ROCs are plots and not numerical values that can be compared via a statistical test. Still, a numerical value can be obtained by measuring the area under the curve of the ROC. One option is to measure $A_z$, however, this measure has the same measurement model as $d_a$ (Hautus et al., 2022). The more popular option, $A_g$, is considered non-parametric and has no measurement model. Still, this measure has been shown to not be bias-independent (Levi et al., 2019).

were to adjust their performance to these rates (*"You will be told whether the percentage might be below / about / above 50% of 'old' words. As a result, you will be asked to try and adjust your 'old' responses to 30% / 50% / 70%, accordingly."*). Participants were only told the implied base rate at the beginning of the test phase, thereby minimizing the possibility that the bias manipulation would affect sensitivity. Participants were then presented with a practice session that included a study phase comprising 10 words and two non-words and a test phase comprising 20 words (10 old and 10 new words). For the test phase in the practice session, participants were always asked to adjust responses to 50% old responses. During the actual experiments, participants were only asked to adjust responses to 30% and 70% ‘old’ responses.

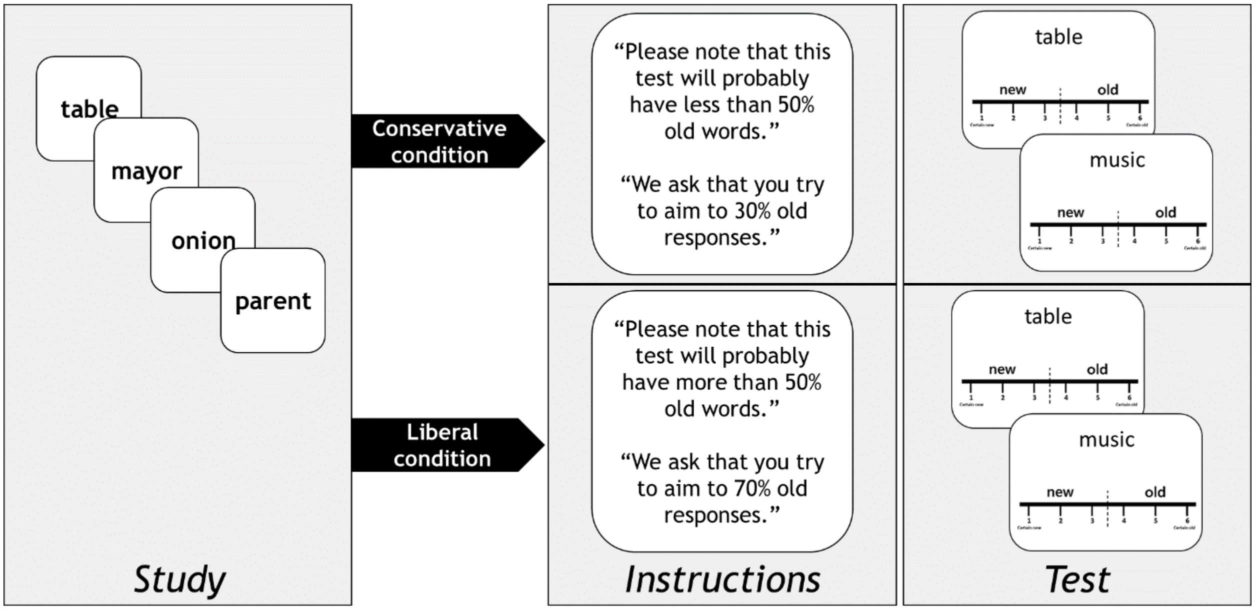
During the actual experiment, participants were presented with the two study-test cycles. Each study phase began with a reminder of the main goal (i.e., *" Try and remember the words as best as possible"*) and of the potential monetary reward. Participants then went through the study phase.  They were instructed to try memorizing the words.  They were asked to press the space bar on their keyboard if, and only if, a non-word appeared. Each item was presented for 1.3 seconds. Following each item, a fixation mark appeared in the center of their screen for one second.

A test phase immediately followed its corresponding study phase. At the beginning of each test phase, participants were informed of the forthcoming implied base rate (*"Please note that this test will probably have less than / more than 50% old words. We ask that you try to aim to 30% / 70% old responses."*). Then, the 63 words that were previously presented during the study were presented again, intermixed with 63 new words, in a random order. They were asked to respond using a 6-point confidence scale, separated into two sections marked new (positioned above ratings 1-3) and old (above ratings 4-6, see Figure 2, right panels).

To help participants monitor their ongoing ‘old’ response rates and comply with the bias instructions, feedback was given at nine variable intervals during the test, defined as the number of items tested since the beginning of that test phase or since the previous feedback was given. Feedback was given either after 11-, 14- or 17-item intervals, reminding participants of the desired response rate and indicating their actual rate over the most recent interval.

The second study-test cycle began immediately after the first, with participants given the opportunity to take a short break between the two cycles. At the beginning of the second cycle, participants were correctly told that memory for items from previous cycles would not be tested and that no items seen in previous blocks would be presented again.  Once the second cycle was completed, participants were thanked for their efforts and redirected back to the online platform.

Both the data and the materials will be made available through the OSF link – https://osf.io/su5q8, for the purpose of replication or any other analysis.



“Please note that this test will probably have less than 50% old words.”

“We ask that you try to aim to 30% old responses.”

“Please note that this test will probably have more than 50% old words.”

“We ask that you try to aim to 70% old responses.”

**Figure 2.** The design of our implied base-rate word recognition task. Left panel describes the study phase. The two right panels describe the test phases of each bias condition. Top-right panels describe the conservative condition instructions and test, and bottom-right panels describe the liberal condition instructions and test. The order of the two test conditions was counterbalanced across participants.

## Results

To obtain 960 participants[8] whose data we could analyze, we ran 1485 participants and excluded 525 based on the following three criteria, predefined through pilot studies, as well as the algorithms that we created to obtain reliable results in our Monte Carlo Simulations (Levi et al., 2024). First, 109 participants were excluded for poor performance on catch trials. Participants that missed more than three (out of seven) non-words in either study phase, or erroneously pressed a key more than three times when a legal word was presented, were excluded from the analysis. Second, 269 participants were excluded because they used fewer than three ratings—in either one of the two bias conditions[9]. Finally, 147 participants were excluded because their test performance were near chance—defined as a difference smaller than 0.1 between participants' HR and FAR[10] in either condition—suggesting they may have not paid attention at encoding or were not motivated to perform the task.

**Manipulation check**. Our implied base-rate manipulation was designed to induce a shift in participants' bias between conditions while maintaining a constant level of sensitivity (see Introduction). As a manipulation check, we plotted an average ROC curve, across participants, for each condition. We expected the two ROC curves (liberal ROC and conservative ROC) to be

[8] A pool size of 960 participants was chosen because it allowed us to evenly divide participants into subsets of experiments comprising several sample sizes (see below, for details).

[9] To reiterate, our earlier simulations (Levi et al., 2024) revealed that the estimation of the zROC slope for $d_a$, individually for each participant, requires a minimum of three unique operating points on the zROC, in each condition. To obtain three points, at least three confidence ratings from 2-6 in each of the two conditions were needed. For more details, see Levi et al. (2024).

[10] This criterion targets participants with ROCs that are very close to the major diagonal, indicating chance performance. Such low performance mostly influences power – these "chance-performance" participants add statistical noise because they increase the standard error. Therefore, including these participants would lead to *lower* Type-I error rates for $d_a$. Note that the outcome of excluding these participants works against our hypothesis that $d_a$ is associated with low rates of type I error. In an exploratory analysis, we tested the effect of including "chance-performance" participants in our sample. The difference in type I error rate turned out to be trivial.

overlapping reflecting equal sensitivity, but with the liberal condition yielding, on average, higher values for the operating points compared to the conservative condition. Examination of Figure 3 reveals this precise pattern.

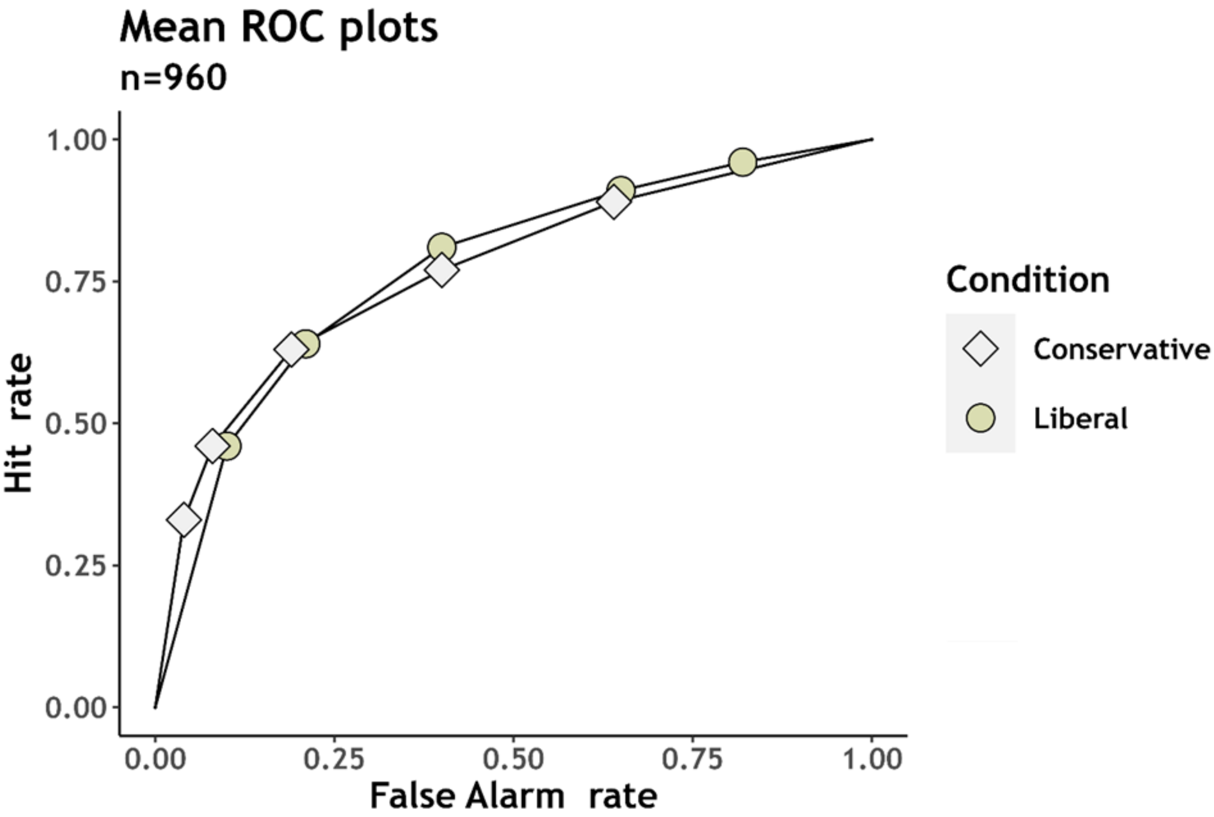


**Figure 3.** Mean ROC curves across the operating points of 960 participants in the liberal and conservative conditions. The circles represent the conservative conditions, and the diamonds represent the liberal condition. For each participant, the operating points computed in each condition were based on 126 trials. Equal sensitivity is reflected by the overlapping curves. Liberal bias is reflected by the operating points in the liberal condition spread closer to (1,1).

We also computed a zROC slope (S) for the 960 participants. We found a mean S of 0.799 (SE = 0.007). These values match previous reports of the magnitude of the slope in recognition memory data, as reviewed in the introduction (Ratcliff et al., 1992; Dougal & Rotello, 2007; Mickes et al., 2007).

Finally, we examined the rates of 'old' responses in the two bias conditions. To this end, we collapsed the data so that categories 1-3 and 4-6 would reflect 'new' and 'old' responses, respectively. When participants were explicitly asked to make 30% 'old' responses, their mean rate of 'old' responses was 40.6% (SE= 0.003), and when asked to make 70% 'old' responses, their mean rate of 'old' responses was 60% (SE = 0.003). These values seem to reflect a comprise between the true base rate, 50%, that participants possibly have the ability to approximate, and what they are told the base rate is (30% or 70%). Such a comprise is predicted by Bohil and Maddox's (2003) COBRA model, according to which a conflict between reward and accuracy will bring participants to compromise. Whereas the COBRA model was formulated for perceptual decisions, our data suggest the same compromise may be at play for memory.

**Primary analysis.** Our dependent variable was T1ERs. To compute T1ERs as a function of sample size, we generated permutations of our 960 participants into "experiments" with sample sizes of 24, 48, 96 and 192 (40, 20, 10 and 5 experiments in each permutation, respectively), by sampling without replacement. We repeated this process, using bootstrapping to generate 5000 random assignments of participants to different "experiments". For example, we ran 5000 iterations in which we randomly assigned our 960 participants into a set of 20 experiments of 48 participants. Then, for each set of 20 experiments, we calculated T1ER. We thus obtained 5000 estimates of T1ERs from 20 experiment of sample size 48. In the same manner, we obtained 5000 estimated of T1ERs for each of the remaining sample sizes. Overall, the bootstrapping method allowed us to get results regarding the T1ER as a function of different sample sizes without re-using any participants'' data in a particular set of experiments.

Our bootstrapping algorithm was designed to ensure that for each experiment, three constraints were kept. First, half of the participants underwent the conservative condition in the first study-test cycle, and half underwent the liberal condition in the first cycle. Second, the four 63-word lists (see Methods) were counterbalanced between the two study-test cycles.  Finally, each of the four lists were used as lures and targets an equal number of times across participants.

For each experiment, we tested several measures: Pr, A´, d´ and, most importantly, $d_a$. We computed these measures for each participant in each of the two bias conditions. We then tested the significance of the difference between the two conditions using a within-participant t test, separately for each measure in each experiment. For each measure, we computed the mean rate of incorrect significant results for each set of 40 (for experiments with N = 24), 20 (for N = 48), 10 (for N = 96) or 5 (for N = 192) experiments. Table 1 presents these T1ERs and Figure 4 depicts these results graphically.

| **N** | **$d_a$** | **d′** | **A′** | **H-FA** |
|---|---|---|---|---|
| 24 | 0.047 | 0.128 | 0.121 | 0.143 |
| 48 | 0.061 | 0.237 | 0.219 | 0.270 |
| 96 | 0.089 | 0.454 | 0.421 | 0.518 |
| 192 | 0.146 | 0.790 | 0.757 | 0.855 |

**Table 1**

Mean values of T1ER across 5,000 iterations. We manipulated the sample size in each set of experiments and measured T1ER for the single-point sensitivity measures (d´, A´, H-FA) and the multiple-point sensitivity measure of $d_a$.

Examination of Table 1 revealed several findings. First, for all three SPSMs, the T1ER when comparing two iso-sensitive conditions with different bias was higher than 5%. Moreover, as sample size grew, T1ER increased for these measures (see Figure 4). This pattern of results was comparable to the pattern of results found in earlier simulations (Rotello et al., 2008; Levi et al., 2024).

Parenthetically, we also used our bootstrapping algorithm to gauge T1ERs for HR. Though known to not be bias-independent, this measure still appears infrequently in the memory literature.  T1ERs for HR were 100% for all sample sizes.  We did not include HR in the table or figure because the 100% T1ERs is such a trivial result requiring only a minimal understanding of how choice is made under conditions of uncertainty: per definition, changes in bias yield differences in HR, large enough to be easily considered significant.

Second, and most important, when measuring sensitivity using $d_a$, T1ERs were approximately 5% (see Figure 4), at least for the smaller and mid-size sample sizes. As previously mentioned, for iso-sensitive conditions we expected T1ERs to be 5%, and we were able to reach the desired T1ER only when using $d_a$ to index sensitivity. Note that $d_a$ T1ERs approximating 5% were also found in our previous simulations (Levi et al., 2024), setting the current result as a successful empirical replication of previous simulation-based findings.

Finally, even though $d_a$ outperformed other sensitivity measures, consistent deviations from the desired 5% T1ER occurred for larger sample sizes. Specifically, for sample sizes of 96 and 192 participants, T1ER of $d_a$ was, on average, 8.9% and 14.6%, respectively (see Table 1). A similar pattern of results was also found in simulations that used large sample-sizes and small amount of trials (Levi et al., 2024).

In the Discussion, we discuss why the increase in sample size results in increasingly higher levels of T1ERs. Still, it is important to note that the rates are sufficiently low so as to be able to control in real experiments, by setting a slightly conservative alpha level (e.g., 2.5%; For further explanation, see Discussion).

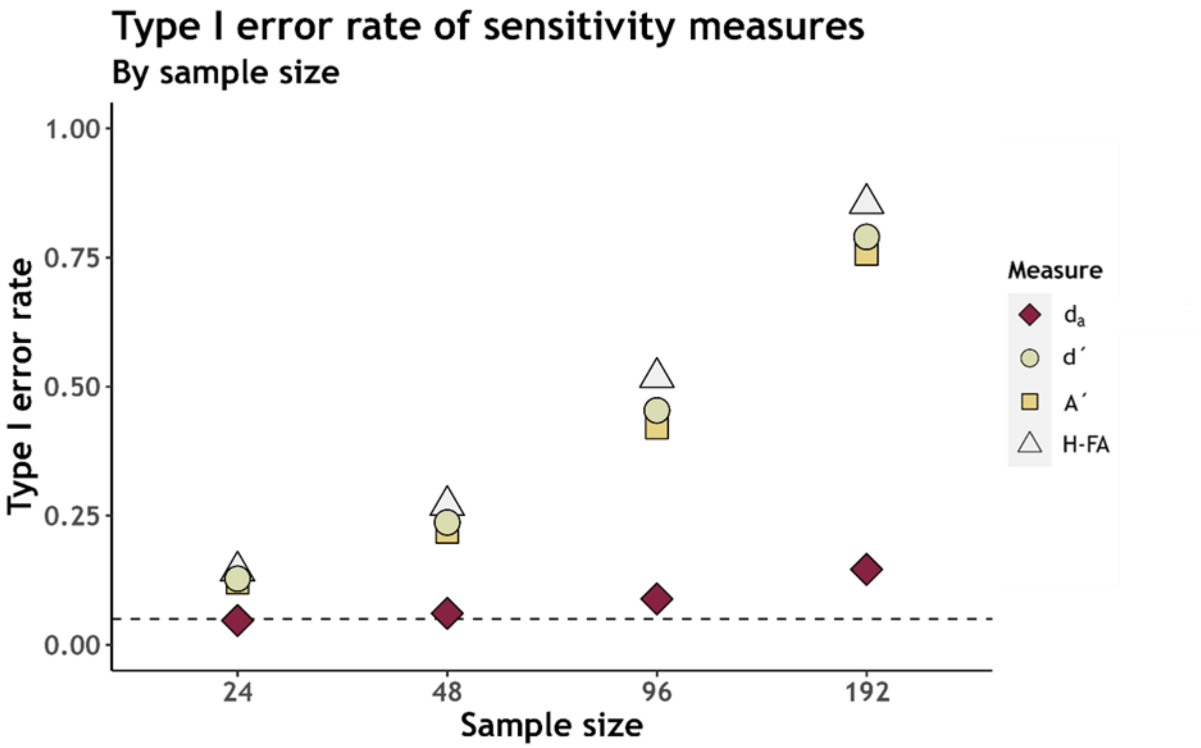


**Figure 4**
T1ER mean values as a function of sample size and sensitivity measures. Each T1ER value was computed across a set of experiments. We randomly created 5000 sets of experiments, resulting in 5000 values of T1ER we averaged for each sample size using each sensitivity measure. The pattern of T1ERs as a function of sample size and sensitivity measures closely resembles that found in our simulations (see above, Figure 1).

## Discussion

The current investigation explored the validity of $d_a$ for measuring sensitivity in recognition memory by means of empirical data. We compared two iso-sensitive conditions with different levels of bias and found approximately 5% TIERs across thousands of random sets of experiments with sample sizes ranging from 24 to 192. Thus, using individual estimations of S for each participant, $d_a$ seems to index sensitivity independent of bias. To date, this is the only measure shown to provide a valid index of sensitivity, not affected by changes in bias across conditions.

The conclusions of the current results converge with those of our Monte-Carlo simulations (Levi et al., 2024), establishing the idea that our version of $d_a$ is the best-known technique for measuring sensitivity in recognition memory tasks. We argue that $d_a$ should be adopted by recognition-memory researchers as the default measure to index sensitivity.

Importantly, when $d_a$ was the chosen measure—but not when the more common sensitivity measures were used—the rate of false discoveries curtailed at the 5% level. For the most common SPSMs, including Pr, A´ and d´ (Figure 4) T1ERs were consistently higher than 5%. Though the rates of Type 1 errors slightly varied between measures, the overall pattern of results was similar – T1ERs for SPSMs were considerably higher than 5%, across all sample sizes.

In reviewing our results, it is important to note that even if one SPSM seems to have better T1ERs than the others, most likely it does not reflect an actual advantage over the other SPSMs. Our simulations (Levi et al., 2024) showed that T1ERs of SPSMs differed as a function of the parameters, with no SPSM consistently showing lower or higher rates across all parameter values. In contrast to the simulations, here we only have one parameter with varying values (sample size), while all other parameters are either fixed (i.e., number of trials) or not under the experimenter control (e.g., distance between lure and target distributions; placement of the five response criteria associated with six-point confidence ratings). Therefore, slightly better T1ERs for a specific SPSM can be easily explained by a mere random difference attributed to the specific values of parameters (e.g., number of trials or distance between lure and target distributions).

Similar to the pattern of results found in the earlier simulations (Rotello et al., 2008; Levi et al., 2024), increases in sample size were associated with higher T1ERs for all the SPSMs. Interestingly, when sample size was large, Type-I errors in the simulations of SPSMs reached a rate of 100%. That is, all 10,000 simulated iso-sensitivity experiments were erroneously significant. However, in our empirical data, the T1ERs did not reach 100%. The highest rates for Pr, A´ and, d´ were 85.5%, 75.7% and 79%, respectively. The higher T1ERs in the simulation data are due to the fact that simulations were clean of any noise that exists in empirical settings. Specifically, in the simulations, identical observations were sampled in the two conditions[11]. Contrast this with the noise in empirical data, where two samples never have identical nominal observations, because of variance caused by the experimental procedure and because of the interaction between manipulation and participants. This noise in the empirical data reduces the probability of obtaining significant results, yielding lower levels of T1ER. Importantly, despite the higher noise levels in the empirical data, T1ERs of SPSMs were still alarmingly high.

A consistent finding in our simulations (Levi et al., 2024) was that when the measurement model was not aligned with the data-generating model, the T1ER was higher than 5%. The fact

[11] Theoretically, the simulations investigated T1ER under conditions that were optimal for finding no difference between conditions. The high rate of Type I errors was found **despite** the fact that the two conditions had identical observations and differed only in criterion placement. Alas, when such optimal conditions are combined with the discrepancy between the data-generating model and the measurement model (Levi et al., 2024), the lack of noise might intensify the T1ERs.

low TIERs were found when we measured sensitivity with $d_a$ suggests that the measurement model of $d_a$ is, in fact, aligned with the recognition memory data-generating model. To reiterate, the measurement model of $d_a$, which uses S to estimate the lure-to-target SD ratio, is consistent with an unequal-variance Gaussian model. Thus the success in indexing sensitivity with $d_a$ may provide converging evidence in support of the UVSD model as the data-generating model of recognition memory (for reviews, Wixted, 2020; Dube & Rotello, in press).

Support for the UVSD model comes against the backdrop of a heated debate regarding the nature of the data-generating model in recognition memory, and regarding what evidence can support one model over another. Using ROC and zROC predictions, previous studies have described results supporting the signal-detection assumptions of continuous, Gaussian underlying distributions (Swets, 1986b; Mickes et al., 2007; Wixted, 2007). In contrast, Rouder and colleagues (2010) showed how different data-generating models can fit the same data. In a reply, Wixted and Mickes (2010) argued that their data fit the Gaussian distributions model better than other reasonable models that have been proposed as the data-generating models in recognition memory including the models suggested by Rouder et al. as alternatives.

Note that the goal of the current project was neither to offer evidence regarding the data-generating model of the task, nor to confirm that the underlying distributions are Gaussian (Kellen et al., 2021). However, our results that reveal approximately 5% T1ER when $d_a$ is used seemingly provided support for the UVSD model. That is, the underlying lure and target distributions of the mnemonic signals seem to be Gaussian because $d_a$ would not have been a valid measure under any other distribution.

Note that at least two models provide a process interpretation of the underlying unequal variance Gaussian distributions.  The first is the UVSD model (Wixted, 2020), according to which recognition memory is mediated by a single mnemonic strength latent variable. The second model that is congruent with unequal-variance Gaussian underlying distributions is the dual-process signal-detection (DPSD) model (Yonelinas, 1994; for a review, see Yonelinas, 2002). According to the DPSD model, the unequal-variance Gaussian distributions are created by a combination of an equal variance familiarity process and a threshold recollection process. Our results do not offer evidence in favor or against either of these underlying mechanisms. They do, however, provide strong support for the hypothesis that the lure and target distributions are Gaussian in form and unequal in variance.

Our results suggest that the assumptions we made, overtly or covertly, while running the simulations (Levi et al., 2024) did not have any detrimental effect on yielding false conclusions regarding the validity of $d_a$. Specifically, we mentioned the confidence carry-over effect observed in recognition (see Introduction; Kantner et al., 2019), an effect documented for recognition tasks, that we did not incorporate in our simulations. In the introduction, we raised the possibility that in empirical data, when the carry-over effect occurs, the validity of $d_a$ might be diminished. However, the T1ERs obtained (approximately 5%, as expected) reveal that the effect was not sufficiently strong as to impair the validity of $d_a$.

An unintuitive pattern in our results was that for large-sample experiments, the T1ERs for $d_a$ were higher than 5% (8.9% and 14.6%, for N=96 and N=192, respectively). This pattern was also found in our simulation data (Levi et al., 2024). At first glance, this pattern may seem puzzling – sample size should only affect power, that is, should only affect performance when there is a true difference in sensitivity between the experimental conditions. Yet, we designed our experiments to have no difference in sensitivity between conditions. Stated differently, why would an increase in sample size affect the T1ER which is the rate of erroneously finding a difference between the conditions that have *equal* sensitivity?

The answer to this puzzle is that the measurement model used for each participant (UVSD with a **specific** lure-to-target ratio) does not, in fact, reflect a perfect fit for the data-generating model, but ‘only’ a very good fit (still, a considerably better fit than the SPSMs we tested). When estimating the lure-to-target SD ratio from the zROC slope, we use S as part of our measurement model assumptions (a larger S means we assume a wider target distribution, compared to the lure distribution)[12]. However, it is reasonable to assume that the S is close to the true SD ratio, but not exactly equal to it. It is further intensified by the fact that the slope is calculated from five data points and often even fewer than that (the impact of few operation points on T1ER was also found in previous simulations; Levi et al., 2024). This difference can lead to minor discrepancies between the measurement model and the data-generating model.

So how does sample size affect the rate of Type I errors? Typically, in small and medium sample sizes, the slight discrepancies between models will not lead to significant differences because there is not enough power for such a small "effect" to be detected. However, when the sample is larger, the increase in power becomes sufficiently large such that these slight discrepancies lead to significant differences between conditions that are tangential to the big issue (congruency between measurement models and data-generation models) but differences that nevertheless exist. Note that this is also the case for all SPSMs – the effect of a discrepancy between the data-generating model and the measurement model is more easily detected with larger sample sizes, leading to an increase in T1ERs.

[12] Our proposed application of $d_a$ requires the computation of S for each participant by collecting confidence ratings. Other versions of indexing sensitivity without collecting confidence ratings should be explored in future research, while adhering to the same validity standards shown here. As for the current research, the only proposed method for measuring sensitivity entails confidence ratings.

Because of the inborn discrepancy between a sample estimate and its corresponding parameter, it is difficult to envisage a measurement model that will invariably provide a perfect correspondence to the data-generating model. Despite this shortcoming, at an applied level, we argue that $d_a$ can and should be used as an index of sensitivity with the following adjustment. To ensure that the chosen alpha level truly reflects the T1ERs, the significance of the results should be tested with a corrected-alpha value for large-sample-size experiments. For example, based on the results reported in this article, if an alpha of 2.5% is adapted for a sample size of N = 100, the true rate of false discoveries will be approximately 5%.

Importantly, for large-sample experiments, researchers may have an additional instrument at their disposal for maintaining alpha at the 5% level. In contrast to increases in sample size, increases in the number of trials may reduce the T1ERs (to values close to 5%). As the number of trials increases, so should the precision of S as an estimate of the true lure-to-target SD ratio. As such, the discrepancy between the estimate of sensitivity and the true value should be attenuated, leading to reduced values of T1ERs even for large sample sizes. Indeed, in our simulation investigations (Levi et al., 2024), increases in number of trials for large sample experiments reduced T1ERs to values close to 5%. Future research should test whether the prediction of a negative correction between length of list and rate of type I errors is generalized to empirical data.

In contrast to $d_a$, we predict that such a negative correlation will not be observed for SPSMs. The measurement models of all SPSMs are not premised on the UVSD model. These alternate models, therefore, are necessarily not aligned with the true data-generation model. Critically, an increase in the number of trials cannot and will not align the models in a better way. To the contrary, a larger number of trials will produce a more precise estimate of the measure, perhaps serving to highlight the misalignment with the data-generating model. If anything, therefore, increases in the number of trials may increase, not decrease, the T1ERs when using SPSMs. An increased T1ER for a higher number of trials was also observed in our simulations when sensitivity was estimated by SPSMs (Levi et al., 2024).

To advocate for $d_a$ as a valid measure of sensitivity, it is critical to also demonstrate sufficient levels of power when used to compare conditions in recognition memory tasks that do differ in mnemonic strength. In this article, we only tested iso-sensitive conditions – both conditions of the experiment had the same level of sensitivity on average. From the perspective of memory research, these reflect null-hypothesis scenarios, wherein no difference in memory exists between the experimental conditions. Under these scenarios, the only errors that can occur are Type I errors. Still, it is possible the low levels of T1ERs are a function of few significant results, in that $d_a$ may have a low power in detecting differences between conditions even if these conditions are in fact different.

To date, there are no data regarding the power of $d_a$ to detect differences. Still, a good first answer to this question is the power of SPSMs to detect differences when the data-generating model is aligned with the measurement model. For example, in previous power simulations (Rotello et al., 2008), when equal-variance Gaussian distributions were simulated, d′ showed sufficient levels of power in detecting differences between aniso-sensitive conditions. Still, additional research is required to test the power of $d_a$ both in simulation studies and in empirical investigations.

Overall, our results further establish the claims to shun the use of d´, A´, H-FA, or other common SPSMs to measure sensitivity in recognition memory. To reiterate, simulation data are sufficient to refute the validity of a measure. However, researchers might question the assumptions made when running the simulations (e.g., the data-generating models chosen for the simulation) and thus not embrace the conclusions of the simulations. If these conclusions are rejected, researchers might continue to use invalid measures in their experiments and false discoveries will inevitably be published as true effects, risking our ability to progress in understanding human memory. Critically, when presenting our recognition-memory results, we did not need to advocate for one theoretical framework over another.  We simply showed which measures are valid and which lead to high rates of false discoveries, in actual experiments. Hopefully, the current results will help in restoring the production process of recognition research by shifting researchers' choice towards using $d_a$ as the only measure of sensitivity in recognition memory.

## References

Bohil, C.J., & Maddox, W.T. (2003). A test of the optimal classifier's independence assumption in perceptual categorization. *Perception & Psychophysics, 65*(3), 478-493.

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.

Danion, J. M., Rizzo, L., & Bruant, A. (1999). Functional mechanisms underlying impaired recognition memory and conscious awareness in patients with schizophrenia. *Archives of general psychiatry*, *56*(7), 639-644.

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

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 130-151.

Dube, 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.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*(4), 375–399.

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.

Hehman, E., Mania, E. W., & Gaertner, S. L. (2010). Where the division lies: Common ingroup identity moderates the cross-race facial-recognition effect. *Journal of Experimental Social Psychology*, *46*(2), 445-448.

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., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The Problem of Coordination and the Pursuit of Structural Constraints in Psychology. *Perspectives on Psychological Science*, *16*(4), 767–778.

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.

Levi, A., Rotello, C. M., & Goshen-Gottstein, Y. (2024). Stop using d′ and start using $d_a$: Part I. Simulation explorations of single- and multi-point recognition measures of sensitivity. *PsyArXiv*.

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., 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.

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.

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

Swets, J. A. (1986b). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181–198.

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.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of experimental psychology: Learning, memory, and cognition, 20(6)*, 1341-1354.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language*, *46(3)*, 441-517.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832.