Title

Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research

Journal

Psychonomic Bulletin & Review, 30(2)

ISSN

1069-9384

Authors

Brady, Timothy F
Robinson, Maria M
Williams, Jamal R
et al.

Publication Date

2023-04-01

DOI

10.3758/s13423-022-02179-w

Peer reviewed

HHS Public Access

Author manuscript

Psychon Bull Rev. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Psychon Bull Rev. 2023 April ; 30(2): 421–449. doi:10.3758/s13423-022-02179-w.

Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research

Timothy F. Brady[1], **Maria M. Robinson**[1], **Jamal R. Williams**[1], **John T. Wixted**[1]

[1]Department of Psychology, University of California, 9500 Gilman Dr. #0109, La Jolla, CA 92093, USA

Abstract

We argue that critical areas of memory research rely on problematic measurement practices and provide concrete suggestions to improve the situation. In particular, we highlight the prevalence of memory studies that use tasks (like the "old/new" task: "have you seen this item before? yes/no") where quantifying performance is deeply dependent on counterfactual reasoning that depends on the (unknowable) distribution of underlying memory signals. As a result of this difficulty, different literatures in memory research (e.g., visual working memory, eyewitness identification, picture memory, etc.) have settled on a variety of fundamentally different metrics to get performance measures from such tasks (e.g., A', corrected hit rate, percent correct, d', diagnosticity ratios, K values, etc.), even though these metrics make different, contradictory assumptions about the distribution of latent memory signals, and even though all of their assumptions are frequently incorrect. We suggest that in order for the psychology and neuroscience of memory to become a more cumulative, theory-driven science, more attention must be given to measurement issues. We make a concrete suggestion: The default memory task for those simply interested in performance should change from old/new ("did you see this item'?") to two-alternative forced-choice ("which of these two items did you see?"). In situations where old/new variants are preferred (e.g., eyewitness identification; theoretical investigations of the nature of memory signals), receiver operating characteristic (ROC) analysis should be performed rather than a binary old/new task.

✉Timothy F. Brady, tfbrady@ucsd.edu.

Author Contributions The initial draft was written by T.F.B. after discussions with M.M.R, J.R.W., and J.T.W. Major revisions of this draft were made by M.M.R., J.R.W., and J.T.W.

Conflicts of Interest None.

Code Availability Not applicable.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Open Practices Statement There are no experiments or data associated with the current article.

Keywords

Measurement; Best research practices in psychology; Theory; Computational modeling; Replication crisis; Recognition memory; Short-term memory; Visual working-memory; Receiver operating characteristics analysis; Signal detection theory

Motivation: Measuring memory is harder than you think

Imagine you wish to know which is better remembered: boats or cars. Or, you wish to know under which condition memory is better: while doing a simultaneous reading task, or while doing a simultaneous listening task. Or you wish to know who has a better memory: Tim or John. To ask these questions, you design an experiment. For example, you show both Tim and John 1,000 objects. Then, after a delay, you show them each of the objects again, along with an equal number of objects they did not see. During this memory test, you ask them to say which objects they remember – which they think are "old" – and which they do not remember – which they think are "new." How would data from such a task guide your inferences regarding whether Tim or John has the better memory?

This question turns out to be deceptively difficult to answer. In fact, as we highlight, an old/new recognition task, in which people are shown items one at a time and are asked to report only if they are old or new, *cannot* reliably tell us who has the better memory despite being among the world's most popular memory paradigms. We point to an example where all mainstream metrics in this and related tasks (e.g., percent correct; A'; diagnosticity ratio; K values; d') would lead researchers to conclude that John has the better memory, even though the data are based on a generative process where Tim has the better memory. This error arises simply because we cannot directly observe and know the true, latent memory strengths of Tim and John; because of this all metrics based solely on old/new performance must implicitly but unavoidably make strong assumptions about the distribution of these latent memory signals and processes, which are frequently incorrect. This is true even of a seemingly theory-free measure like overall "percent correct." The challenge of measuring recognition memory "well" is not simply theoretical: Data from old/new tasks that can lead to incorrect inference[1] are quite common (Glanzer et al., 1999; Ratcliff et al., 1992; Yonelinas & Parks, 2007); furthermore, real-life decisions are often made using old/new metrics like these, even when they subsequently turn out to be wrong when memory is studied with better, more principled measures (e.g., Wixted & Mickes, 2018).

The use of old/new metrics in recognition memory tasks has profound implications for both theory development and research-driven policy making (for related discussions of measurement issues, also see Eronen & Bringmann, 2021; Flake & Fried, 2020; Guest & Martin, 2021; Kellen et al., 2021; Luce & Krumhansl, 1988; Meehl, 1967; Oberauer & Lewandowsky, 2019; Rotello et al., 2015; Regenwetter & Robinson, 2017; Scheel et al., 2021). This type of "recognition memory test" is one of the most popular ways to test memory and is the focus of this article. Because recognition memory tasks are so prevalent

[1]For example, when receiver operating characteristics are asymmetric, as they often are (e.g., Yonelinas & Parks, 2007), all the metrics are straightforwardly incorrect.

across research domains, we use “memory” and “recognition memory” interchangeably throughout most of the article. Recognition memory tasks are not the only way to measure memory, however, and we discuss alternative methods in the General discussion. In this paper, we first illustrate and elaborate on different measurement approaches that a researcher might take to compare memory between people or across experimental conditions; we then explain why they are fundamentally flawed when estimated from these kinds of data (“did you see this item?”). Finally, we discuss ways to properly measure memory using such recognition memory paradigms, as well as how to test the critical assumptions underlying these measurements. We propose that forced-choice tasks (“which of these two items was previously seen?”) should be the default task in basic science memory research that simply seeks to understand which of two conditions or which set of people have the strongest memories, rather than old/new studies (“was this item seen?”). In situations where researchers need to use an old/new task, we suggest that receiver operating characteristic (ROC) analysis should always be used to measure recognition memory performance.

The need for counterfactual reasoning in measuring recognition memory

Imagine that in the above scenario, Tim correctly calls 900 of the 1,000 objects he saw “old,” whereas John only correctly calls 800 of the 1,000 objects “old.” Would you conclude Tim has a better memory? You might be tempted to say “yes.” It turns out that memory researchers do in fact sometimes draw such conclusions, arguing for differences between people, stimuli, and conditions with data based almost solely on “hit rate” (how many old items alone were recognized; e.g., Buchanan & Adolphs, 2002; Henderson & Hollingworth, 2003; Isola et al., 2011; Rouhani et al., 2020). However, only in very rare circumstances could this be a valid measure of memory. In general, we have absolutely no way of knowing, from the hit rate alone, whether Tim or John has a better memory, whether boats or cars are better remembered, or whether there is a difference in memory across experimental manipulations. This is because we do not know whether there might be a difference in response bias between different people or different stimuli. For instance, maybe Tim says “old” if he has any inkling the picture is familiar; whereas John says “old” only when he’s very sure it is old. Similarly, if the hit rate for cars exceeded that for boats, it would not necessarily mean that cars are more memorable than boats. Maybe people have an expectation that they would have a stronger memory for boats they saw than for cars, and they require a greater sense of familiarity to call boats “old” than to call cars “old.” These examples underscore the critical role of response bias in old/new recognition memory tasks; when it comes to measuring memory, it is a pernicious nuisance variable. Before we can attempt to measure memory, it is necessary to somehow take into account (and neutralize) such differences in response bias across people and across stimuli. To do this, we need to consider the “false alarm” rate: how often people incorrectly call items they did not see “old.”

Taking into account false alarms means we now have two measures – hits and false alarms – when we wish to have only one measure, which tells us “how strong” the memory was. One way to deal with this is to report hit and false alarm rates separately, not attempting to unify them into a single coherent measure of memory (e.g., Bainbridge et al., 2013; Bjork & Bjork, 2003; Castella et al., 2020; Chan & McDermott, 2007; De Brigard et al., 2017;

Gardiner & Java, 1991; Jiménez et al., 2020; Khader et al., 2007; Otero et al., 2011; Smith & Hunt, 2020; Soro et al., 2020; Yin et al., 2019). In many cases, this effectively results in inferences being made based on hit rates only, whereas false alarms are treated as a nuisance variable or as a measure of a completely distinct process, rather than a variable that can provide insight into which people have or which conditions lead to the strongest memories. In short, since no unified measure of memory strength is calculated, analyzing hits and false alarms separately – without a model that connects them – can lead to erroneous conclusions about variations in recognition memory.

Most researchers realize they need to integrate hit and false alarm rates into a unified measure of memory performance, and they attempt to do so. It turns out, however, that in their attempt to unify hit and false alarm rates from a binary[2] old/new paradigm researchers tacitly, and perhaps unknowingly, make strong, theory-based assumptions that are almost always incorrect (sometimes wildly so). As can be seen from Table 1, which summarizes mainstream measures across several memory disciplines, different papers and different literatures tend to choose different metrics, which rest on very different assumptions about underlying memory processes and architecture. This divergence limits our ability to integrate and accumulate knowledge across literatures and can lead to contradictory interpretations of the same data.

At its core, the issue with these metrics, and with the binary old/new task in general, is that we are faced with answering a *counterfactual*: "if Tim and John had the same false alarm rates, what would their hit rates be?" (Fig. 1). It is obvious from this framing that we cannot answer this question definitively from these data alone. That is, these data are consistent with either John or Tim having a better memory and to know who has the better memory we need to know how hit rates change as false alarms change. What may be less obvious is that this is the question that all of the most popular metrics (d', A', adjusted hit probability, etc.) are aiming to answer. Moreover, they all rest on different theories and, therefore, make different, and often contradictory "guesses" about the counterfactual scenario of how hits would vary with false alarms – all of which are likely incorrect to different degrees. Despite the fundamental limitations of these metrics, researchers may routinely pick a single metric without justifying their choice, or considering the difficult counterfactual scenario they are attempting to address by using such metrics.

As noted, the goal of this paper is to elucidate the issues with continuing these problematic measurement practices and to encourage the use of more appropriate paradigms and metrics. We also aim to demonstrate the value of incorporating previous advances in measurement into new work. The paper has three parts: First we explain what we mean by latent (and unobservable) memory strengths, and explain the strong theoretical assumptions the various metrics (d', A', K, etc.) make about latent memory strength. Next, we justify the critical importance of these measurement concerns for both proper memory research and for real-life, applied scenarios where measuring memory is critical (i.e., eyewitness identification). Finally, we explain how to measure memory more accurately, using ROC

[2]By "binary" we mean that participants are simply asked to indicate their memory on a 2-point scale of memory strength, "old" or "new," rather than a multi-point scale that indicates the strength of their belief that the item is old.

analysis or (whenever possible) a forced-choice procedure, and in conclusion we provide general recommendations for researchers interested in understanding memory strength.

All measures of old/new performance make strong and likely false assumptions about latent memory strengths

Understanding the assumptions that these different measures (*K*, overall percent correct, corrected hit rate, d', A', etc.) rest on about how hits change as false alarms change, as well as how these measures can lead researchers astray, requires that we consider the *full distribution* of memory signals for both genuinely old (previously seen) and genuinely new (unseen) items. We must consider these distributions because, when we ask who has the better memory, we are asking *who more reliably believes that old items are actually old and new items are actually new* – which is fundamentally a question about the entire distribution of underlying memory strengths associated with old and new items. Yet, such information is fundamentally unknowable from binary old/new data.

To illustrate this, let's go back to the case of Tim and John, with Tim having 900 hits but 500 false alarms, and John having 800 hits and only 150 false alarms. With these data alone, all of the measures summarized above agree with one another – they all indicate that John has a vastly superior memory (Fig. 2). Thus, *all* of these metrics would lead us to believe that in the counterfactual world where we force Tim and John to have the same false alarm rate as each other, John would have the higher hit rate.

Yet John *does not* have a better memory. In fact, these aggregate data were derived from hypothetical data where *Tim has the unequivocally better memory* (i.e., whenever their false alarm rates are matched, Tim always has the higher hit rate). The only reason Tim's superior memory is obscured is because Tim and John have different response criteria—a different propensity to call items "old" versus "new." This is particularly problematic because response bias is the variable that almost all of these measures are putatively designed to correct for!

Let's see how this is possible. The simplest way to do this is to make a separate table for Tim and John. We can then break down their memories by how strong the latent memory signal is for both genuinely old items and genuinely new items. That is, how much evidence there is in favor of each object being "old",[3] after their memory system integrates all the information it is using to make such a decision. Take Tim's memories, below. For genuinely old items, Tim has fairly strong memories, and most, but not all of the items he actually saw feel quite familiar to him, though some feel more familiar than others. By contrast, most of the genuinely new items feel unfamiliar to him, although they also vary in how unfamiliar they feel, such that a few new items may feel quite familiar (e.g., maybe a mug that is shown during the memory test was not in the study session, but happens to resemble Tim's officemate's actual coffee mug). The criterion Tim used to decide to call something "old" (a

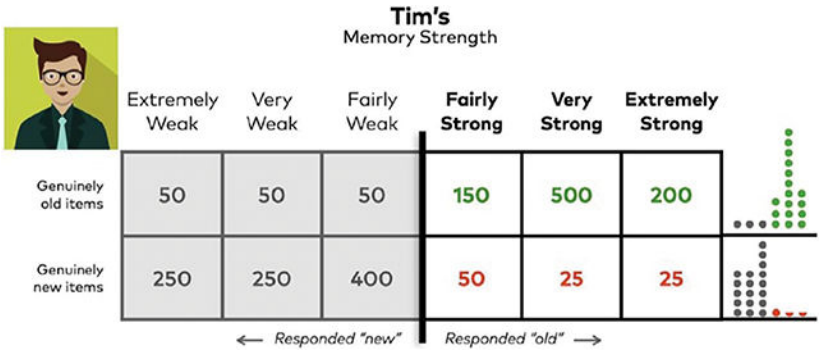
[3]For now, we will assume that memories do in fact vary in strength, that is, some items are remembered better than others, after integrating across all sources of information used to decide if the item is old or new.

liberal response bias) is shown by the dark line; to the right of that are things he called “old” and to the left are things he called “new.”

As shown, Tim has a very strong sense that many of the genuinely old items were “old,” but a more middle-of-the-road feeling about many of the genuinely new items. Based on his decision criterion, Tim took all of his fairly weak memory signals and higher, and called them “old,” which means he did so for 900 old objects, and 500 new objects. A consequence of Tim’s liberal decision criterion is that he has both a lot of hits (green) and a lot of false alarms (red). Now let’s compare this to John’s memory decisions.

John has a more conservative response bias than Tim and only called fairly strong items “old,” resulting in him having called 800 genuinely old objects “old,” and only 150 genuinely new objects “old.” As we noted above, this proportion of “old” responses to old and new items led us to believe that his memory was better than Tim’s. Furthermore, all the metrics based solely on the old/new data told us this was so (Fig. 1). But was it really?

Let’s imagine we convinced Tim to be as conservative in his responding as John, such that Tim only said “old” when his memory signals were “fairly strong” (and higher), like John did.



Now, Tim would have 850 old items he correctly called “old” but only 100 new items he incorrectly called “old.” Thus, with the same response criterion, Tim would now have more hits *and* fewer false alarms than John, who had 800 hits and 150 false alarms. So, despite the massive difference in false alarms and small difference in hits we observed when Tim used a more liberal response criterion, we now see that *Tim actually had a better memory all along*—he is now more accurate for both genuinely old and genuinely new items than John is. That is, Tim had more genuinely old items with strong signals, and more genuinely new items with weak signals than John did. Tim just chose a response criterion that was more liberal in identifying test items as “old” than John did. Importantly, all of the binary old/new metrics (A’, d’, etc., Fig. 2) agreed with one another that in the counterfactual world where we lowered Tim’s false alarm rate to more closely match John’s, Tim would have a lower hit rate than John, and they were all wrong! Ultimately, Tim had a higher hit rate when we actually made this change.

The crux of the issue is that we had no way of knowing that this was true because we had no access to either Tim’s or John’s internal memory states; instead, we only had information regarding which items each person decided to report as “old.” We were misled because, like in all old/new experiments, we did not have access to the “full table” (i.e., the full latent distribution of memory strengths), where we could see what would happen if we changed the response criterion for each person to match their false alarm rates. Instead. we only had access to which items Tim and John thought were “old” with whichever criterion they themselves chose to use. The reader might wonder if we had to generate unusual (nonrepresentative) numbers for this example, but the data used for Tim and John here are abstracted from a real long-term memory task: that is, they are strongly reminiscent of real data from real recognition tasks.

This is why long-term memory old/new or and working memory “change detection” tasks that rely on standard metrics cannot, in a theory-free way, tell us either who has the better memory or which items were better remembered. We can only answer the counterfactual question of who would have a higher hit rate with a particular false alarm rate if we know the entire internal distribution of memory signals for both genuinely old and genuinely new items, and take that into account in our calculations – and we cannot know this from only a single set of old/new responses. Indeed, that internal distribution of memory signals is fundamentally unknowable (for related points, see Kellen et al., in press; Rich et al., 2021).

At their core, this is what each of those theory-based measurement models (d', A', K, etc.) does: they each make strong theoretical assumptions about the nature of the distribution of memory signals that exist for both previously seen and not seen items, allowing them to "guess" what would happen in the counterfactual world of equal false alarm rates between conditions, people or stimuli. For example, "high-threshold" theories that justify measures like corrected hit probability (hits minus false alarms) or "K" values assume that there is no such thing as a weak memory: all memories are 100% strong or 100% absent. In such theories, you can never sort-of-think the couch you saw was red but not be completely sure, or be uncertain whether you ate a sandwich for lunch yesterday or not. According to standard threshold theories, you either remember something with 100% certainty or you do not remember at all, and any feeling of uncertainty in your memory is just noise that reflects nothing about your memory states. Such metrics thus assume that the latent distribution of memories has nothing in the middle columns of Tim's and John's table: all memory signals are either maximally diagnostic or completely uninformative.

Even just calculating overall "percent correct," which, as noted in Table 1, *feels* theory neutral, makes this same strong assumption, that is, percent correct is only coherent if memory is all-or-none in this way. If there is such a thing as gradations in memory strength, a measure like "percent correct" is an incoherent measure of latent memory. This is because as soon as we assume that people can vary in the strength of their memory signals (i.e., memory is not all-or-none) we also assume that they must set a decision criterion for responding whether an item is old or not old. In such cases, threshold-based measures like percent correct would make incorrect assumptions about the critical counterfactual scenario of which would have a better hit rate if two conditions had equal false alarm rates.

By contrast, calculations based on signal detection theory make the more well-validated assumption (Wixted, 2020) that items vary in memory strength (e.g., that you might sometimes feel like something you have not seen is somewhat familiar, and items you have seen might elicit either strong or weak memory signals). However, metrics based on signal detection almost always assume that memory strength for both genuinely old and genuinely new items are normally distributed. Furthermore, the most prevalent signal detection measure d', makes the stronger and less plausible assumption that both the old and new items have the same variation in strength, and differ only in their mean. This means that d' only provides a pure measure of memory strength independent of response criteria if all items are encoded exactly equally well, with no variation between items (Dougal & Rotello, 2007; for detailed discussions regarding how encoding variability and multiple continuous memory processes relate to parametric assumptions of signal detection models, see Jang et al., 2012; Wixted, 2007; Wixted & Mickes, 2010). This strong assumption also automatically fails if memory strength is modulated in a qualitatively different way for targets vs. foils, for example if multiple sources of evidence are integrated to arrive at "old" item's strength but not "new" item's strengths (as envisioned by some dual-process models of memory; e.g., Yonelinas, 2002).

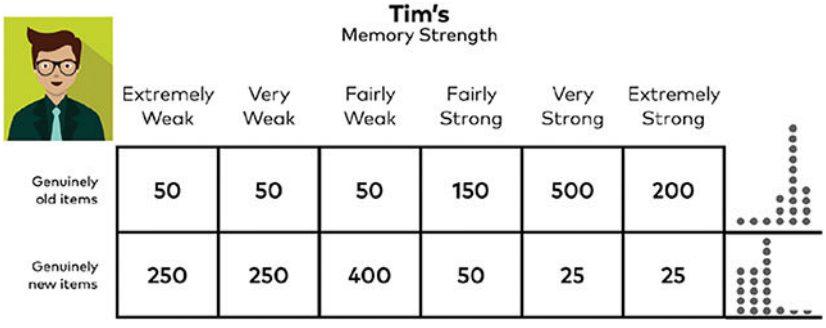
Similarly strong and immutable assumptions about the nature of the latent memory strength distribution underlie all of the other measures, like A' and diagnosticity ratio. Intuition notwithstanding, none of these assumptions are very tenable. Different situations almost

certainly lead to different memory strength distributions, and it is unlikely they follow any clean rule most of the time. For instance, there is little evidence to suggest that memory processes exhibit absolute discreteness (with no true variation in memory strength, "precision" or genuine variation in "confidence"), or that memory signals follow a perfect normal distribution with zero variability in how well items are encoded and only a single source of memory information (e.g., item memory alone). Therefore, relying on these theoretical assumptions to extract information about a person's entire distribution of memory signals – and thus about what would happen in a counterfactual scenario where we forced two people to have equal false alarm rates – from a *single point* (e.g., just a single set of hit rate and false alarm rates from an old/new task) is deeply problematic. As we explain in the next section, this issue is not simply problematic from a theoretical standpoint, and is not circumscribed to hypothetical examples, but has substantive real-world implications. To understand this more fully, it is useful to think first about how we would measure people's full memory strength distributions, and how these old/new metrics relate to such measurements. The core principle behind this analysis is that if we wish to know who would have a higher hit rate at the same false alarm rate, we should use a measure that reveals or at least depends upon the entire "table" of memory strengths.

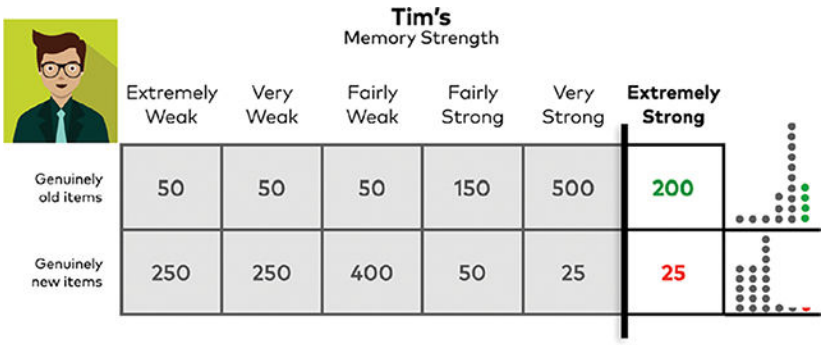
Assessing the full latent memory strength distribution: Idealized receiver operating characteristic (ROC) analysis

If you wanted to measure who had the better memory, based on the full underlying distribution of a subject's memory strength, how would you do it? We will go in depth on how to do this empirically later (e.g., how to get such measurements), and what assumptions are required by various methods of doing so. For now, let's assume that we can somehow directly "read out" people's memory strength, with no noise in the read-out process, and think about what we'd do with that data to determine who had the better memory.

ROC analysis is the technique that directly answers the question of who can more reliably distinguish targets from lures – i.e., who has a higher hit rate across the full range of false alarm rates. To conduct an ROC analysis, let's again look at Tim's latent memory strength value for genuinely old and genuinely new items, as shown below. There are six levels of memory strength in this example, ranging from 1 (Extremely Weak) to 6 (Extremely Strong).



To compute an ROC, we use these memory strength bins to measure how participants' performance would change for different response criteria and compute a *series* of hit and false alarm rates. For example, let's pretend Tim responded "old" only when he reported he had very high memory strength and otherwise he said "new," i.e., Tim calls items "old" if he has a latent memory strength of level 6 (Extremely Strong), and otherwise calls them "new."



The hit rate is the total number of hits (200) divided by the total number of genuinely old items (1,000), and the false alarm rate is the total number of false alarms (25) divided by the total number of genuinely new items (1,000). This gives us a "hit rate" of 200/1,000 = .20, and a "false alarm rate" of 25/1,000 = .025. Next, assume instead that Tim calls items "old" if they have a latent memory strength of level 5 (Very Strong) *or more*.

Now, there would be 500 + 200 = 700 hits and 25 + 25 = 50 false alarms. This gives us a hit rate of 700/1,000 = .70 and a false alarm rate of 50/1,000 = .05.

We continue to cumulate the proportion of hits and false alarm rates in this way until we get to the last possible criterion, where we call anything with latent memory strength of level 2 (Very Weak) or above "old." This represents the sort of responses we'd get if Tim were incredibly liberal, since he called almost everything "old," as shown below.

This gives us a hit rate of 950/1,000 = .95 and false alarm rate of 750/1,000 = .75. We can now plot these points and connect them to make a curve, plotting the hit rate against the false alarm rate for each of the five criterion values, and do the same for John as shown in Fig. 3.

When we plot out performance in this way, which reflects the full distribution of each person's memory signals, we can see that, in fact, Tim has a clearly better memory than John. There are multiple ways to quantitatively capture that fact, but in this case you can tell this just by looking. That is because for every possible false alarm rate (*x*-axis), Tim has a higher hit rate (*y*-axis) than John – Tim's curve is further up and to the left, closer to perfect performance. This means that if we fixed the false alarm rate at any value, Tim would always correctly identify more old items as old than John. That is exactly how you determine whose memory is best.

ROC curves, then, insofar as we can solve the problem of how to empirically get access to people's latent memory strength, address the core question of which conditions or participants or stimuli would have higher hit rates at the same false alarm rate. In other words, this approach provides a way of measuring memory performance without relying on relatively unconstrained, theory-based extrapolation, as is required if you collect only binary old/new data. ROC curves are in many ways just a quantitative change from a binary/old new task, of course, since here we observe 6 levels of memory strength (5 points on the ROC) as compared to two levels of strength in the binary task (a single point). However, this quantitative change makes a large difference in understanding the latent memory signals.

How does an ROC analysis compare to common metrics of memory?

Why were the measures that we used in the old/new task, such as d', corrected hit probability or A', so wrong about Tim and John? They were wrong because these measures rest on strong theoretical assumptions about the distribution of underlying memory signals, which is the very reason they can be used to extrapolate the full distribution of memory signals from a single point (Macmillan & Creelman, 1990). What this means is that when applying these measures, researchers take a single point on this ROC curve – just one pair of hits and false alarms, from a single old/new task – and infer what they think the full ROC would look like, in order to perform the counterfactual reasoning highlighted in Fig. 1. If the extrapolation of the distribution of latent memory strengths is incorrect, as it nearly always is when it is made with a single point, then so too will be the separation of memory performance from response criteria shifts. For example, Fig. 4 shows the actual ROCs from Tim and John and the implied ROCs based on the d', the corrected hit-rate, and the A' values we obtained from the old/new task.

Each of the ROC curves shows how Tim's and John's hits and false alarm rate *should* change across different response criteria, given the d', K, or A' values we calculated from the old/new task, if these metrics were correct in the inferences they make about the latent memory strength distributions. That is, every point along each curve has the exact same d' or K or A' value as Tim and John's actual data, and so, in theory, should reflect the exact same memory strength distribution. However, as we see in this example, all of the assumptions these models make are quite wrong in the case of John and Tim. Thus, they all make incorrect predictions about how Tim and John's performance would change for different levels of response bias. As a consequence, they all incorrectly lead us to believe that John had the better memory, when, as we saw, Tim has better memory than John – a higher hit rate for every possible false alarm rate.

In short, if the very particular assumptions of these measures turn out to be false, these measures confound response bias with performance, and do not in any way serve as a correction for "guessing' or "response bias." This makes it clear why it is so problematic that different papers, and different literatures, use different methods without clear justifications: they cannot all be right about what the true latent memory distributions (and thus, true ROCs) look like. In fact, as explored in the next section, even though some are almost certainly more accurate than others (i.e., ROCs almost always look more like those postulated by d' than by the diagnosticity ratio), it is unlikely *any of them* are "correct" in the general case when they perform this counterfactual reasoning from a single set of hit and false alarm rates.

How important are these concerns?

Are such concerns solely theoretical, rather than practical? It is certainly possible that they have little practical relevance even if they are theoretically important. If ROC curves are always identical to those predicted by d', for example, then researchers should simply use that metric rather than worrying about measuring latent memory strengths. Furthermore, even if ROC curves do vary in shape, if all of the metrics always converge anyway in real-world applications, then it will not matter that they make different assumptions. Unfortunately, neither of these is the case.

There are at least two pieces of evidence that there is a genuine crisis in research practice rather than these measurement practices being a solely theoretical concern. First, think about how many papers use K values vs. d' vs. A' vs percent correct vs. hits minus false alarms, for example, the small subset of such papers cited in Table 1. These measurements make different assumptions about how response criterion shifts affect hits vs. false alarms, and thus about the counterfactual matching of false alarm rates, and they cannot all be right at the same time (Fig. 4). That means results that hold for "K" values may not hold in d' or A' or vice versa: they simply can't hold if response criterion differs across conditions or people, since the models fundamentally disagree on how performance for people or conditions with different false alarm rates should be judged – and in real data, no two people or conditions will reliably have the *exact* same false alarm rate. This means many memory studies using old/new paradigms must confound response criterion differences with memory performance differences to at least some degree.

In addition, it is empirically true that the way hit rates change when we change false alarm rates varies quite a bit across studies (see Fig. 5) – and insofar as there is such a thing as a "modal" shape for an ROC in memory research, it is one that matches *none* of these metrics since it is curvilinear but asymmetric (see Wixted, 2007; Yonelinas, 2002). Such asymmetric curves – like we showed with the example of Tim and John – can easily lead to important situations where more accurate measures of latent memory strength "flip" the conclusions from binary old/new measures. Once researchers in a particular field have started looking at ROCs, they have sometimes found existing measures got things exactly backwards (e.g., Wixted & Mickes, 2018). We detail these concerns in the next section.

ROCs really do vary in shape quite a lot

Do ROC curves vary in shape, and/or mismatch the shapes assumed by the metrics, or is one metric (e.g., d', A', or correct hit probability) nearly always correct about the shape of ROCs? Unfortunately, estimates of ROCs do indeed vary quite a bit, ranging from approximately linear to extremely curvilinear, and nearly all have varying degrees of asymmetry around the y = 1 − x (upper left to lower right) line. For example, Fig. 5 shows a sample of real ROC shapes from actual memory papers.

None of these ROC curves perfectly match any of the implicit assumptions made by d', A' or any of the other metrics (compare them to Fig. 4). While nearly all are at least somewhat curvilinear (reflecting that memory is not, in fact, all-or-none, but has gradations of strength), they do not all follow the same shape, with some much more curvilinear (e.g., purple) than others (e.g., light green). Thus, empirically it is true that given a single hit rate and false alarm rate, our inferences about memory are, at best, uncertain, insofar as ROC curves do not all follow the exact same shape.

Even more unfortunate for the metrics used in old/new tasks is that, insofar as there is a single canonical shape of an ROC, it is one that matches none of the old/new metrics. While the majority of curves are curvilinear, in line with d' and signal detection theory more broadly, memory ROCs tend to be asymmetric. Most of the ROCs in Fig. 5, in fact, are at least somewhat asymmetric with respect to the y = 1 − x line. This is in contrast to nearly all of the metrics, which, except for the diagnosticity ratio, all assume symmetry along this axis. Though there are different views about how to parameterize such curvilinear and asymmetric curves (e.g., DeCarlo, 2010; Wixted, 2007; Yonelinas, 2002; Yonelinas et al., 1996), there is generally strong agreement, at least in studies of long-term memory, that empirical ROC curves are neither linear (as required for threshold models, like adjusted hit probability or percent correct), nor curvilinear and symmetric (as required by d' and A'), and this is likely true of working memory in many circumstances as well (Robinson, Benjamin, & Irwin, 2020a; c.f. Williams et al., 2021). Instead, as in these examples, they are generally curvilinear and asymmetric with respect to the y = 1 − x line (being steeper on the left and shallower on the right). Such asymmetries are seen as practically inevitable in common theories of memory: Signal detection theory, even with just a single underlying memory representation as the basis of the signal, suggests that memory-based ROC curves should be to some extent asymmetric in this way, since such an asymmetry arises naturally if there is any variance in how well items are encoded,

which there presumably is (e.g., Wixted, 2007; Jang et al., 2012; however, note that there is ongoing work examining how direct manipulations of encoding strength affect estimates of this kind of asymmetry, e.g., Spanton & Berry, 2020); and dual process models of memory strength, common in recognition memory, propose this necessarily arises whenever multiple sources of information are contributing to memory responses, as in the case of recollection and familiarity (e.g., according to both the conceptions of Yonelinas, 2002, and Wixted, 2007; Wixted & Mickes, 2010).

Thus, while d' may prove the most useful measure in many cases – given the likelihood that memories do vary in strength and that ROCs do tend to be curvilinear (e.g., given that if forced to pick a metric from Fig. 4, d' most closely matches the ROCs in Fig. 5) – it is almost certainly "incorrect" as well. "Old' items almost certainly vary more in strength, either because of variation in how well they are encoded or the contribution of more sources of memory information to their ultimate strength, creating asymmetries that are not properly accounted for by d' but instead require something like an "unequal variance signal detection model' (e.g., where the strength of items that were actually seen is more variable than the strength of items that were not seen), or another conceptualization of the asymmetric ROC curve (e.g., Yonelinas, 2002). Unequal variance signal detection models and alternative conceptions of asymmetrical distributions cannot be fit from a single pair of hit and false alarm rates – they require the full ROC curve. This is because such models can lead to complicated scenarios where the relationship between hit rates and false alarm rates can vary across levels of response bias (for instance, if Tim's distribution of memory signals for old items is higher in variance than John's, Tim could have a higher hit rate at low false alarm rates but a lower hit rate at high false alarm rates). Given these scenarios, it is imperative that researchers attempt to measure the full distribution of memory signals with as high resolution as possible.

Taken together, real-world data strongly suggest that these measurement issues are not merely theoretical: real ROC curves (1) vary in shape, and (2) insofar as they have a canonical shape, it does not match any of the shapes assumed by metrics that can be computed from old/new tasks. Overall, then, we suggest that all of these metrics incorrectly estimate discriminability in memory, and they do so by making incorrect "counterfactual" predictions about how hit rate would change as false alarm rate changes.

This has major implications for memory research and for society

Are such measurement issues simply a concern for memory researchers, or do they have broader implications? Clearly, the fact that empirical ROCs differ from the theoretical ROCs that are implied by these measures does not mean that *all* model-based approximations will necessarily lead to the wrong answer. All computational modeling efforts require making simplifying assumptions (e.g., Fum et al., 2007) and, in some contexts, these assumptions may lead researchers to the correct answer most of the time (as when fitting reasonable models, like the equal variance signal detection model to estimate d' from full ROCs rather than a single point; see below). However, our claim is that making these simplifying assumptions while extrapolating from a *single point* is *especially problematic*, and leads to errors that have major implications for nearly all memory research and the application of

this research to society. In the specific case of extrapolating from old/new performance to memory strength, findings have often been used to support widespread policy changes that can have serious, life changing ramifications in the here and now, and measurement issues have proven critical to such issues (for related points, see Rotello et al., 2015).

A critical example comes from the domain of eyewitness identification, which provides an extremely important and salient warning to other memory researchers. There is much interest in how eyewitness identification is influenced by contextual factors – such as the way in which faces in a lineup are presented – because this may affect an eyewitness's ability to identify the perpetrator in a lineup, if the perpetrator is present. Thus, many variations on lineup presentation formats have been compared to see which results in the best performance, with one critical example being sequential vs. simultaneous presentation of the to-be-judged faces. A lineup contains the face of one suspect (innocent or guilty) and five or more physically similar fillers. In sequential lineups, the faces are presented one at a time, whereas in simultaneous lineups all faces are presented at once. A lineup is a kind of old/new recognition task in that the witness is asked to state whether the perpetrator is in the lineup (old) or not (new). It is more complicated than a standard single-item old/new task because more than one item is presented and, if the decision is "old," the witness must additionally specify which face is that of the perpetrator. Despite these differences, the considerations discussed above (e.g., witnesses can be liberal or conservative) still apply.

For decades, numerous studies purportedly showed that diagnostic accuracy improves when faces are presented sequentially rather than simultaneously (e.g., Lindsay & Wells, 1985; Steblay et al., 2011) – and this evidence had a major influence on policy, with many police departments changing their line-up procedures as a result (Police Executive Research Forum, 2013). However, these studies used the "diagnosticity ratio" to measure people's memory for perpetrators, based on the intuition that taking the ratio of hits and false alarms should "correct" for response bias. As can be seen, however, by comparing Fig. 4 and Fig. 5, the diagnosticity ratio is perhaps the worst choice for extrapolating how hit rates would change as false alarm rates change: its predictions about how hit rates will change are not just slightly wrong, but incredibly so (Rotello et al., 2015). Thus, different combinations of hit and false alarm rates, even with the same underlying discriminability and same underlying latent memory strengths, yield wildly different values of the diagnosticity ratio, and differences in diagnosticity ratios are virtually uninformative about which conditions result in the best memory performance. In essence, diagnosticity ratios are largely measures of response bias, rather than memory strength. Unfortunately, this means that when memory researchers used proper measurement techniques in eyewitness identification, reasonable metrics that consider the full distribution of latent strengths (by using ROC analysis) revealed that performance was, if anything, superior for simultaneous, not sequential lineups (Mickes et al., 2012; Mickes & Wixted, in press). This not only subverted mainstream literature on eyewitness identification (for a discussion, see Rotello et al., 2015), but also had major policy implications, showing that what scientists had been telling police departments for decades was due to a failure of proper measurement (National Research Council, 2014). More precisely, it reflected a failure to realize that old/new performance necessarily requires making a *model-based* counterfactual assumption about what hit rates would be as false alarm rates change.

These measurement issues also have large-scale implications for theories of memory. In visual working memory, "*K* values" – an all-or-none, threshold-based measure – are ubiquitous, because of the claim that they purportedly represent the "number of discrete items represented in memory" (which is an intuitive proposition; e.g., Alvarez & Cavanagh, 2004; Brady & Alvarez, 2015; Cowan, 2001; Fukuda, Vogel, et al., 2010; Irwin, 2014; Pailian et al., 2020). *K* values are used, oddly enough, even by researchers who strongly argue that memory is not all-or-none for each item, but that items differ in "precision" (e.g., Awh et al., 2007), even though as we have seen, *K* only makes sense as a way to counterfactually predict hit rates for different false alarm rates if memories are all-or-none, not if they differ in strength. As you would expect, then, from the rich literature showing working memories do in fact vary in strength (i.e., in both "precision" and "confidence," which covary – Rademaker et al., 2012; plausibly because both measure the same underlying memory strength – Williams et al., 2022), *K* values are not a good match to the actual shape of ROCs in most working memory situations, with working memory ROCs being generally curvilinear (Robinson, Benjamin, & Irwin, 2020; Robinson et al., 2022; Williams et al., 2021; also Fig. 5). Curvilinear ROCs are consistent with the fact that signal detection models are able to successfully account for data from a variety of visual working memory paradigms (Williams et al., 2022). Unfortunately, unless ROC curves in working memory are perfectly linear – and despite some claims from small samples (Rouder et al., 2008), the evidence strongly suggests they are not, even when using measures that do not depend on confidence ratings (e.g., Robinson et al., 2022; Williams et al., 2021) – this means that nearly all conclusions based on *K* values are suspect, as they do not properly discount response criteria differences, and thus measure a combination of response bias and memory strength.

The fact that *K* metrics rest on untenable assumptions is arguably more problematic in the visual working memory literature than in many other literatures because differences between measures like *K* and the (actual) curvilinear empirical ROCs are far more prominent when criteria are very conservative (i.e., when false alarm rates are very low) than in the middle of ROCs (see Fig. 4, and Williams et al., 2021). Data from recognition memory "change detection" tasks, which are frequently used to estimate *K*, seem to lead to extremely conservative responding in many situations. For example, reanalyzing data from 3,849 people completing a change detection task with 4 items (Balaban et al., 2019), shows that 73.4% had false alarm rates below 0.1. This is the exact area in ROC space where "*K*" values and empirically curvilinear ROCs most strongly diverge, and thus the area where *K* values are most likely to be picking up on differences in response bias rather than genuine differences in memory strength.

As explained by Williams et al. (2021), the area of research within visual working memory where such mismeasurement may have the most profound implications is individual differences work. Much prior work suggests that individual differences in *K* values from simple change detection tasks are large and reliable (Vogel & Awh, 2008), and have been found to correlate with fluid intelligence (e.g., Fukuda et al., 2010). This is surprising because such tasks, unlike many working memory tasks that engage broad executive components, are mostly seen as measures of simple storage capacity, which is usually not thought to be associated with intelligence in the same way as more executive control-based measures (Conway et al., 2002; Engle, 2002, 2018; Engle & Kane, 2004). However, since *K*

does not well describe the shape of ROCs in this domain, these conclusions could indicate that it is mainly individual differences in response bias that correlates with intelligence, rather than (or in addition to) visual working memory "capacity" *per se*. In fact, it is known that response bias, and propensity to adapt this bias to a task, tends to vary across and be quite stable within individuals (Aminoff et al., 2012; Kantner & Lindsay, 2012; Miller & Kantner, 2020). Although it may not be that response bias explains all of the covariance between working memory capacity and intelligence,[4] it remains a viable possibility that many individual differences in simple storage visual working memory tasks, as measured by K, are in fact stable differences in response criteria between participants, potentially undermining measured relationships to intelligence and other critical conclusions about the nature of cognitive architecture.

Similar examples could be given in many other literatures, beyond eyewitness memory and visual working memory. These are simply two examples to demonstrate that the measurement issues we have raised have significant real-world impacts as well as significant impacts on theory development.

What do we do about this? How do we measure memory more accurately?

We hope it is clear that there is no way to understand memory performance from old/new tasks alone without strong and often inaccurate assumptions about latent memory distributions, and that we should all do something different if we wish to understand memory. But what if you already have binary old/new data? How should you analyze it?

At a minimum, if using old/new data, researchers should report the hits and false alarm rates clearly, and should report whether their results are robust across different measures. Of course, as demonstrated above, this may still be inadequate because measures such as d', corrected hits rate, and A' may agree with one another but still yield incorrect inferences about the true underlying memory signal distributions when memory distributions are asymmetric, as they often are (see Figs. 4 and 5). Yet with only old/new data, we are stuck with simply using the best assumptions possible. As seen in Fig. 5, most ROCs are curvilinear, and signal detection theory is a strong default framework (see below). Thus, we generally believe that with old/new data alone the default metric should be d' (see also Mickes et al., 2014). This should only be done when hits and false alarms correspond directly, however: No measure of memory strength can be derived if hits and false alarms are not one-to-one (e.g., if there is one global false alarm rate and many separate hit rates, as in many continuous recognition tasks, Brady et al., 2008; or tasks where there are different kinds of "new' stimuli but only one kind of "old' stimulus mixed together, Brady & Alvarez, 2015). We revisit the question of optimal theory-based measures in later.

However, our strongest recommendation is to not collect such data (binary old/new or change detection) in the first place. If you have the freedom to choose your method for a

[4]For now, we will assume that memories do in fact vary in strength, that is, some items are remembered better than others, after integrating across all sources of information used to decide if the item is old or new (see Section 5). (Unsworth, et al., 2014; though we note that such covariance techniques can still confound variations in memory strength with response bias; for a discussion of alternative linking modeling approaches see Turner, et al., 2019).

new study, how should you properly test whether one stimulus class is better remembered than another, or if one person remembers information better than another? In general, there are two ways: The first way is to try to assess the full distribution of memory signals for both previously seen and previously unseen items in the way you probe memory, rather than relying on assumptions about these distributions, using ROC analysis. The second (and easier) way is to have people choose between multiple potential answers (e.g., pick which of two items is the "old' item; a.k.a., forced-choice tasks, which we describe below). Contrary to many people's intuition, forced-choice turns out to be significantly more theory-neutral than old/new tasks, because it forces participants to make use of their full latent strength distribution for previously seen and previously unseen items across trials. In the limit of perfect measurement, these two techniques – ROC analysis and forced-choice – are equivalent, in the sense that the area under the ROC curve is the same as forced-choice performance (Green, 2020; Green & Swets, 1966; 1988). However, they have different pros and cons in practice.

Note that in theory there are other possibilities than forced-choice and ROC for accurate measurement. For example, researchers could attempt to use adaptive procedures that equate false alarms across individuals or experimental conditions via instructional and/or feedback manipulations, which would eliminate the need for comparing people with different false alarm rates. However, forced-choice and ROC analysis are the most common techniques, and so we discuss these two possibilities in detail next.

Correct measurement: Forced-choice

The simplest possibility for assessing memory with minimal theoretical baggage (and the one we have preferred in most of the first-author's work; e.g., Brady et al., 2008; Brady et al., 2016) is using forced-choice tasks like 2-AFC. In a 2-AFC task (or any multiple alternative choice, *m*-AFC task, in the general case), participants do not decide whether a single stimulus is old or new, but instead on each test trial must pick which of two items is the old one (where one old and one new item are always present). Such tasks provide more theory-neutral measures of memory because regardless of the shape of the ROC, or distribution of memory signals, order is always preserved: the better-remembered stimulus will always yield better memory performance.

To understand why forced-choice but not old/new provides an accurate index of memory, consider what would happen if we took Tim and John's distribution of memory strengths, and gave them a forced-choice test instead of old/new. On each trial, we take a random old item and a random new item and pair them, and they pick the item that evokes the stronger memory signal (Fig. 6). In this scenario, Tim gets 87.8% correct, and John only 78.9% correct.[5] With no ROCs or model of memory signal strengths, we have finally recovered the true fact that Tim has a better memory than John! Intuitively, we can see that forced-choice captures the entire underlying memory strength distribution – whereas old/new does not – because in forced-choice, the old and new items on each trial are experimenter-controlled random samples from *anywhere* in the entire memory strength distribution. Forced-choice

[5]To compute these, we simply simulated this exact process 10,000 times given the memory strengths in Table 1.

thus provides an unbiased estimate of how many genuinely old items have higher memory strength relative to genuinely new items. In contrast, in old/new, we only get a window into whether old or new items generate memory signals that are above or below a line (decision threshold) chosen by the participant, obscuring most of the information about the distribution of items' strengths.

The relative accuracy of forced-choice holds true in terms of all the measures available to quantify forced-choice performance. For instance, rather than using percent correct, we might use the equivalent of *K*, or corrected hit probability, for 2-AFC: $R=(2*PC-1)$ (e.g., Brady et al., 2008). This would suggest, in a world of all-or-none memories, that Tim "remembered" 75.5% of the items and John 57.8% (if memory were all-or-none). Similarly, d' for 2-AFC is $(\Phi(\text{hits})-\Phi(\text{false alarms}))/\sqrt{2}$ (e.g., Macmillan & Creelman, 2004; Makovski et al., 2010), which suggests, in a world of normally distributed, equally well encoded memories, that Tim (who had d' of 1.65) had a stronger memory than John (who had a d' 1.14), etc.

Naturally, there is still room for non-linearities when comparing different metrics of performance in forced choice tasks. For instance, corrected hit probability values computed from 2-AFC performance will not be a linear function of d' values computed from 2-AFC performance. Therefore, if researchers are interested in testing theory-specific hypotheses regarding *how much* performance should change across different experimental conditions, their choice of theory will necessarily play a role. Importantly, however, these measures will always be strictly increasing functions of one another, meaning that if corrected hit probability is larger in condition A than B, d' will also be larger in condition A than B. Thus, in 2-AFC tasks, researchers' conclusions cannot "flip" depending on which measure they use, unlike in old/new tasks. Thus, the choice of measurements and theory still matters if researchers choose to interpret these measurements in a linear way – for instance, because there may be a non-cross-over interaction with one measure, but not another measure (Loftus, 1978; Wagenmakers et al., 2012). However, 2-AFC or other *m*-AFC tasks at least have the property that regardless of what measure is employed they should not confound response bias and memory accuracy and lead to qualitative differences in researchers' conclusions regarding ordinal differences in memory across people or experimental conditions.

People sometimes object to 2-AFC as a measure of memory based solely on a mistaken intuition that when faced with a choice of which of two boats you have seen, you are not *really* measuring memory, because rather than remembering one of the boats, people might instead be very sure that they did not see the other boat. Under this account, your ability to reject the new lure rather than remember the old item may "inflate" your performance. Ideas such as these have been cited in papers attempting to argue against 2-AFC (e.g., Cunningham et al., 2015). But this objection does not hold up to scrutiny: As we have repeatedly seen in this manuscript, even in old/new tasks, we must consider both hits *and* false alarms to measure memory. Therefore, the exact same issue arises in old/new tasks. That is, people might have a "high" ROC not because they remembered more boats, but because they were just very, very sure they had not seen the new boats! In fact, such decision rules have been identified in both the long-term (e.g., Rotello et al., 2000) and

visual working memory literatures (Cowan et al., 2013). In many ways, such concerns get to the heart of why measuring memory can be so counterintuitively complex: we cannot ever measure memory for things we have seen, at least not in the intuitive way. We can only measure the *difference* between memory for items that were seen and items that were not seen.

In fact, because 2-AFC makes the comparison explicit, it may guide our thinking and result in more interpretable data. This holds because in 2-AFC, more attention is drawn to explicitly thinking through how the unseen items (lures) compare to the seen items (targets) they are paired with, and how best to match such comparisons to ensure fair comparisons across conditions (e.g., when comparing different stimuli, it is critical that the foils be somehow matched for difficulty or similarity between stimuli, Brady & Stoermer, 2020). By contrast, in old/new tasks, subjects can employ a wide range of strategies to inform their decisions about whether the presented item is old or new, and such strategies can impact performance in old/new tasks in complex ways (e.g., Robinson, Wixted, & Brady, VSS 2020), which researchers might not anticipate.

Notably, variants of forced-choice, in particular “continuous report’ (see Fig. 7), are extremely popular in visual working memory research. In such tasks, an item from a continuous feature space (e.g., a color wheel) is shown, and then at test, participants must choose what color it was from the entire feature space. In many ways, these tasks share the same benefits as 2-AFC for performance, with no direct need for ROCs, since they can be conceived of as effectively 360-AFC tasks. Unlike a normal 2-AFC task where one item is seen and the other, the foil, is unseen, however, continuous report includes “foils” that are extremely similar, and even perceptually confusable with, the “old” item. This has led to many complex models taking into account not just percent correct in choosing the exactly right color, but also the distribution of responses to different foils (e.g., errors to nearby, similar items being treated as “precision errors” vs. errors to far-apart, dissimilar items being treated as “guess” errors; Zhang & Luck, 2008). For example, mixture models (e.g., Bays et al., 2009; Zhang & Luck, 2008), variable precision models (e.g., van den Berg et al., 2012) and neurally inspired models (e.g., population coding; Taylor & Bays, 2020) of visual working memory all attempt to characterize what the distribution of people’s responses to foil items reflects about visual working memory processes. Recent work by Williams et al. (2022) has suggested, however, that such tasks – when reconceived as 360-AFC tasks – are ultimately the same as 2-AFC, with response distributions simply appearing complicated because some of the foils are quite similar to the target and some are not. Schurgin et al. showed that once the perceptual confusability of the stimuli and their similarity to the remembered item are taken into account, continuous color report and 2-AFC appear to estimate the same, single underlying measure of memory strength.

Although most would probably agree that the problem of response bias is less pronounced in forced-choice tasks compared to old/new tasks, in practice, even a forced-choice test does not *necessarily* live up to its potential. For example, in a 2-AFC recognition memory task for words, Jou et al. (2016) found that participants tended to show a left-side bias, sometimes effectively making an old/new decision for the word on the left (ignoring the word on the right altogether). Likewise, Starns et al. (2017) used eye-tracking to monitor people’s

attention to tested items presented in 2-AFC tasks, and found that sometimes participants respond without even looking at all of the alternatives. Thus, even 2-AFC can be affected by response biases that, when present, reduce accuracy below what it otherwise would be. This does not appear to generally be a major issue given that other studies using 2-AFC have reported negligible response biases (Kroll et al., 2002; Smith & Duncan, 2004; Westerberg & Marsolek, 2003). Still, we agree with Jou et al. (2016) that it makes sense to take steps to avoid possible response biases even when using force-choice tasks. For example, in 2-AFC, the target should appear on the left versus the right (or the top vs. the bottom) 50% of the time. Similarly, in more complex forced-choice tasks, such as continuous report, the response wheel could be randomly rotated on a trial-by-trial basis. In addition, before the test, participants could be informed that the target will appear equally often in the available spatial locations. Finally, it would make sense to examine the data after the fact for any evidence of a response bias instead of simply assuming that no such bias was present (e.g., in 2-AFC, checking to see if approximately half the responses were made to the left and half to the right), and only considering the task a valid measure of memory if such response biases are minimal.

Correct measurement: How to construct and evaluate empirical ROC curves

Rather than using forced-choice, another possibility for assessing memory with minimal theoretical baggage is to try to assess the full distribution of memory signals for both previously seen and previously unseen items (rather than relying on assumptions about these distributions as in old/new) using receiver operating characteristic (ROC) analysis. This is the method that has generally been preferred by the senior author and his lab (e.g., Wixted & Mickes, 2018). ROC analysis provides two potential benefits. First, the area under the ROC curve is a theoretically neutral measure of memory performance in the same way as forced-choice performance is (e.g., Wixted & Mickes, 2018). Second, for those wishing to build theories of memory performance, rather than simply assess it, ROCs provide more useful data than forced-choice. For example, given an ROC, you can estimate theory-driven metrics like d' far more accurately than from a single set of hit and false alarm rates. You can also estimate measures of memory strength using other models, which are based on more sophisticated theories about how memory-based decisions are made (like d_a, e.g., Mickes et al., 2007; Goshen-Gottstein, 2019). Of course, ROCs also allow probing memory for a single item at a time (e.g., Is this item old or new? How sure are you?) – which, in real-world situations, is the more realistic way memory is probed (e.g., eyewitness memory line-ups can be “rejected,” and are thus inherently old/new, not forced-choice).

Computing ROCs also comes with practical benefits as compared to a binary old/new task. In computing d' in a binary old/new task, sometimes participants will be a ceiling or floor. You can use a correction if you have hit or false alarm rates at ceiling or floor (Macmillan & Creelman, 2004). However, if you have to use this correction too often (e.g., more than 15% of the time), then it is a sign that you did not have a sufficient number of trials in your experiment. This issue does not arise almost ever when using ROCs to compute d' or other similar measures, another reason to prefer ROC analysis.

As we saw above, if we could directly "read-out" participants' memory strength, ROC curves provide a straightforward way to analyze them. However, there are (at least) three issues that arise when actually trying to work with ROCs rather than discussing their theoretical properties: the first relates to how to best construct ROC curves in psychological studies (how does one "read-out" memory strength?); the second is that proper aggregation across trials and subjects can be more difficult than expected; and the third is the question of how to reduce an ROC to a single measure of memory performance. We take these on next.

Methods for constructing ROCs

The intuition behind ROC analysis is straightforward: we want to somehow get access to additional information about latent memory strength, beyond what is available in old/new. However, it is important to note that it is ROC analysis per se, not the mere addition of more information about memory strength on its own, that ensures we are isolating latent memory strength from response bias. This is because the ROC technique directly addresses the counterfactual question at issue in memory research: what the hit rate would be if we changed the false alarm rate.

By contrast, many researchers have the intuitive sense that collecting more than just a binary old/new judgment may be important, and thus end up collecting some additional data about latent memory strength, but this does not always result in more clarity about latent memory signals. For example, a well-known method of collecting richer data, the "behavioral pattern separation" paradigm, in which participants must classify individual test items as old, similar or new, rather than just old/new (Borota et al., 2014; Stark et al., 2013; Toner et al., 2009; Yassa & Stark, 2011) does not solve the major counterfactual problem that is at the core of old/new memory measurement. "Similar" responses may add to our knowledge of latent memory strength, but they provide information about a different dimension of memory compared to the old/new task. The relevant dimension of memory used to make a decision in a recognition memory task is determined by the question posed to the participant ("Is this item old or new?" "Is this item similar to a target from the list or not?" "Was this item presented in Source A or Source B?," Wixted & Mickes, 2010). As described in detail below, a separate ROC can be constructed for each question by sweeping a decision criterion across the corresponding dimension using, for example, confidence ratings (e.g., on a 1–6 scale, 1 = "Sure Not Similar" and 6 = "Sure Similar"). Thus, because typical "behavioral pattern separation' paradigms mix dimensions (the old-new dimension and the similar-not-similar dimension), they fail to separate the memory signal strength along a given dimension from response criteria placed along that same dimension (Loiotile & Courtney, 2015). Therefore, they do not directly address the question of how hits relate to false alarms along any one dimension of memory, as an ROC analysis does. A similar concern has been raised about the popular "remember/know" distinction, which is that it conflates memory strength and confidence in an old/new decision based on one dimension of memory involving multiple components (such as a combination of item and contextual information, or a combination of recollection and familiarity) with another Remember-Know dimension of memory defined only by knowledge of contextual details (Wixted & Mickes, 2010). Thus, it is critical not just to collect more data about latent memory strength or item "precision' (i.e., Awh et al., 2007). Instead, to correctly separate memory strength from

response bias, the data must be collected and analyzed according to the principles of ROC analysis.

How does one construct an ROC? The two main ways of constructing ROCs are by asking participants to report their memory strength (e.g., reporting how confident they are that an item is old), or by using external manipulations to try to shift their response criterion (e.g., manipulating how often items tend to be old vs. new, or rewarding participants differentially for being more or less conservative). Both are ways to attempt to learn more about the latent distribution of memory strengths along a dimension established by the question posed to the participant (e.g., "Is this item old or new?"). Of course, both methods provide imperfect measures of these latent distributions, but they reveal far more than a simple old/new test.

Using confidence reports is as straightforward as asking participants to report their memory decision on an, e.g., 6-point scale (sure it was old <-> sure it was new) rather than a 2-point scale (old <-> new). Thus, confidence reports are simple to collect with naive participants, require fewer observations than direct manipulations of response bias (Wickens, 2001), and are straightforward to turn into empirical ROCs by simply treating confidence as indices of memory strength and performing the analyses seen above. To the latter point, confidence tends to strongly track properly computed performance metrics, as would be expected if it tracked the strength of the underlying memory signal (e.g., Mickes et al., 2007; Mickes et al., 2011). However, some researchers have argued that confidence judgments may yield a distorted measure of latent memory strength, and while this is of course possible in theory, it is important to ask about the plausibility of the role of such distortions in empirical data. For instance, Malmberg (2002) speculated people may sometimes report that they are unsure that the old item is old, even if they are extremely sure the old item is old; or, conversely, people may sometimes report that they are extremely sure an old item is old, even if they are not sure an old item is old. Clearly, these types of "noisy" response policies could distort measures of the underlying memory distribution. However, evidence generally supports the idea that confidence data are quite meaningful; attempts to empirically test whether such noisy responses, rather than noisy latent signals, could underlie confidence have generally provided strong evidence against the view that this is a major factor (e.g., Williams et al., 2021; Delay & Wixted, 2021). For example, working memory data show that people have direct access to their own memory strength and use their own assessments of how strong their memories are nearly optimally (e.g., Fougnie et al., 2012), and people never report anything other than the highest and lowest confidence when memory is strong (e.g., set size 1 working memory) but frequently do so when memory is weak (e.g., set size 6 working memory), arguing against straightforward "confidence is simply noisy" accounts (e.g., Williams et al., 2021). Note that the use of "confidence" in such reports does not necessarily suggest that people have a form of "meta-" memory per se. At least through the lens of signal detection theory, confidence reports in such studies are conceived of as simply and directly reflecting the strength of the very signal elicited by the probe stimulus (e.g., using a 6-point scale instead of a 2-point scale), which people use to make the old/new decision in the first place (for theories regarding how to measure meta-memory judgments within the signal detection theory framework, see Galvin et al., 2003). In this view, there is only a quantitative difference, not a qualitative difference, between asking for responses on the 2-point scale used in many studies (old/new), which are not seen as subjective and noisy,

and responses on a 6-point scale (sure old<->sure new), which sometimes are seen in such a way.

Independent of confidence, another way to measure the distribution of latent memory strengths for genuinely old and genuinely new items is to convince participants to change their own response criteria in different blocks of trials. Manipulating base rates or reward is a common way to do this, but it can be difficult to implement in practice because it requires that subjects are sensitive to these manipulations and can change their own criteria accordingly (Cox & Dobbins, 2011). When base rate manipulations are employed, researchers change – across experimental blocks – the proportion of trials on which an old item is shown, thus changing people's expectations about the probability that an item is old or new on a given trial. Similarly, in studies where rewards are used to manipulate bias, researchers vary the payoff structure (e.g., paying participants more if they correctly identify an old item as being old than if they correctly identify a new item as being new) across blocks to bias participants towards making a specific response (e.g., old vs. new). Thus, in both types of studies, on different trials we either encourage participants to respond "old" only when they are very sure it is old, or we encourage participants to respond "old" even if they have just an inkling that is "old." From these blocked manipulations we can infer the entire distribution of memory strengths for genuinely old and new items. As mentioned, these manipulations are only useful if participants are sensitive to them, and only if they actually move response bias enough to give you diagnostic, high-resolution data about the shape of the ROC (e.g., Robinson et al., 2022). Manipulations of response bias, of course, also rest on the potentially problematic assumption of selective influence (Van Zandt, 2000): that is, the idea that these manipulations selectively affect response bias without affecting people's accuracy on the memory task, and that people maintain a fixed criterion throughout an entire block of such trials. Finally, because manipulating response bias requires training participants, such tasks naturally require a larger number of trials than ROC studies with confidence.

Overall, however, despite potential limitations of both methods, there is evidence that they may both yield similar results (e.g., Williams et al., 2021 show this for the case of visual working memory) – although this outcome hinges on the diagnosticity of the data (e.g., Dube, Rotello, & Heit, 2011). For instance, if researchers collect data using a base rate manipulation, but participants are insensitive to this manipulation and, in the extreme case, all points cluster around a single value, the ROC function will be no more diagnostic than data collected from an old/new task. Therefore, at a minimum, researchers should always plot and check that points in the empirical ROC function do not cluster in this way. Important in the current context is that, in principle, these methods may both provide a useful window into the latent memory strength distributions, which is what must be assessed to determine who has better memory or which condition led to better memory performance, and both allow for ROC analysis. Given the fact that confidence judgments can be easily administered to naive subjects and require fewer observations – coupled with compelling evidence that confidence tracks memory strength well – we strongly promote the use of confidence-based measures, which can be straightforwardly analyzed as described above.

Finally, we overview a few alternative methods for constructing ROCs, which may be useful for researchers who seek to understand the processes that people use when reading out their memory strengths in ROC tasks. In particular, it is possible to use reaction time and neural measures as a window into the latent memory strength distribution, in addition to or instead of confidence and/or bias manipulation. More precisely, some researchers use recognition memory tasks that emphasize both accuracy and speed of responding, and include reaction times in their analysis of ROCs (e.g., Ratcliff & Starns, 2009). This approach couples ROC analysis with diffusion modeling and permits researchers to quantify how evidence accumulates towards each decision criterion. The second approach involves using neural data to construct ROCs (e.g., Weidemann & Kahana, 2016; 2019). For instance, Weidemann and Kahana (2019) used EEG and multivariate classifiers to quantify how neural evidence for recognition-based memory decisions accumulates over time. These researchers found that ROCs obtained from temporal fluctuations in neural data covaried with ROCs obtained from confidence data. Both the use of reaction times and neural data may provide a richer characterization of the architecture and dynamics of recognition memory, and help link them to other (e.g., neural) processes. However, if researchers' goal is to simply compare levels of memory sensitivity across experimental conditions and individuals, we promote the standard method for collecting ROCs by measuring hits and false alarm rates at the level of behavioral responses with no time pressure.

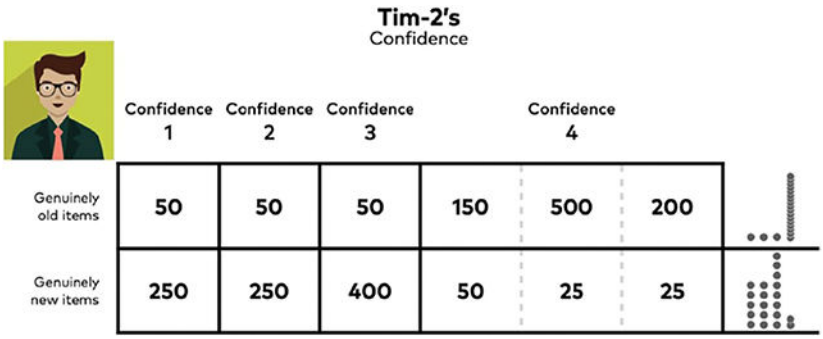
Aggregation

Working with ROCs requires thinking in a non-linear way because as we have seen, empirical ROCs tend to be curvilinear. One consequence of this nonlinearity is that it can lead to aggregation artifacts. For instance, if people differ in their confidence criteria (e.g., how they map memory strength to confidence), then the average of points taken from two participants can yield a lower ROC curve and a correspondingly lower estimate of A' or d' than you would obtain from plotting out the ROC curves separately. For illustration, imagine that you administer a 4-point confidence scale to people who we will call Tim-1 and Tim-2. Note that Tim-1 and Tim-2 each have the same underlying distribution of memory signals (i.e., the data in each table is the same), however, they map their memory states to the confidence scale differently. Tim-1 uses confidence level 1 a lot:

Tim-1's Confidence

	Confidence 1			Confidence 2	Confidence 3	Confidence 4
Genuinely old items	50	50	50	150	500	200
Genuinely new items	250	250	400	50	25	25

By contrast, Tim-2 uses confidence level 4 a lot:



Now if we take their data and plot two separate ROCs – shown below – we see that, indeed, they overlap as expected, and both make up part of the overall Tim-ROC we saw above (Fig. 4). However, if we average the two, computing a hit rate and false alarm rate for confidence level 1, another for confidence level 2, etc., this group average ROC is actually lower than both the component ROCs as shown in Fig. 8.

Unfortunately, this means the average ROC is actually lower than the true performance; this is true regardless of the metric you use to measure performance from the ROC. Formally, this result follows from Jensen's inequality for concave functions (like ROCs), which entails that a randomly selected individual's performance should exceed that of the average (e.g., Kuczma, 2009). This is by no means a unique problem for ROC analysis: similar aggregation issues arise when calculating other measures, like A' or d', that are non-linear transforms of hits and false alarms, and in many other contexts as well (e.g., Estes, 1956; Estes & Maddox, 2005). However, such heterogeneity can result in an average score that is unrepresentative of the data in terms of performance, and should be taken into account for ROCs, just as it should be taken into account in other contexts, like when using A' or d' on binary old/new data.

In the case of data where all of the variability is at the level of individual subjects, one way to take this into account is to compute subject-specific ROCs. Indeed, the individual-subject approach is frequently recommended for many kinds of quantitative analyses of data (e.g., Cohen et al., 2008; Estes, 1956; Estes & Maddox, 2005). However, response criterion variability can also potentially differ between conditions, and even within conditions, if response criteria shift or vary over time more in one condition than another. In most cases, this seems unlikely, but it is an issue that can potentially arise in ROC analysis and may be a reason to prefer forced-choice instead, if modeling latent memory signals *per se* is not of interest. Such averaging artifacts can be dealt with to some extent by using more sophisticated modeling approaches, such as hierarchical Bayesian measurement models (Rouder et al., 2017). However, such approaches are still imperfect, as there can be heterogeneity caused by factors we are unaware of or not explicitly manipulating, even when matching items and participants (e.g., drift over the course of the experiment).

Importantly, ROC analysis done at the level of a single participant clearly provides more information about the relevant distribution of memory signals than a single hit and false alarm rate – and as noted, aggregation issues arise just as much when calculating non-linear measures like d' to single hit and false alarm rates, and, therefore, are not unique to ROCs.

However, despite being a problem common to many measures, they are important to keep in mind when using ROCs. We also emphasize that, while aggregation may result in an underestimation of the true level of performance in each of two conditions being compared, it is unlikely to reverse the conclusion about which condition yields higher discriminability (whereas using the wrong measure, such as the diagnosticity ratio, can easily do so).

Getting single measures from ROC curves

How do you quantify memory performance with a single number when using ROC analysis? As noted, area under the curve (AUC) is a theoretically neutral measure of an ROC, effectively equivalent to forced-choice performance (Green, 2020), that can be used without subscribing to a particular model of memory (e.g., signal detection). However, there are potential concerns that researchers should be aware of when using AUC (Wixted & Mickes, 2018). For example, in practice, researchers may be forced to make some parametric assumptions in order to extrapolate the ROC curve, if ROCs in different conditions cover different ranges. Of course, researchers can assess whether their results are robust across different parametric assumptions. They may also opt to calculate the partial area under the curve, instead of the full area under the curve, to avoid extrapolation. In some conditions, partial area may be the most relevant measure of interest, without extrapolation. For example, in the study of eyewitness memory, researchers may need to calculate the partial area under the curve because participants are instructed to identify an old object (guilty suspect) amongst a set of fillers or to reject the lineup (in this sense, some eyewitness tasks combine both elements of recognition and m-AFC memory tasks). In such tasks, researchers must work with truncated ROCs, however, the height of the ROC at high false alarm rates is of little interest, because practically speaking, only low false alarm rate procedures would ever be used (Wixted & Mickes, 2018).

Finally, we note that ROCs allow scientists to not just measure memory but try to understand the basis of memory decisions. Thus, many researchers will choose to embrace a specific theoretical framework, such as signal detection theory (e.g., Wixted, 2020), in understanding ROC data and converting ROC data to a single measure of memory strength. For example, memory researchers can measure memory by fitting models to a participant's empirical ROC curve and estimating parameters or calculate metrics that reflect a participant's sensitivity, such as d_a, the equivalent measure of d' for unequal variance signal detection models (which can only be measured with a full ROC). This approach is undoubtedly better than relying on a single point to extrapolate the full ROC function, and even fitting measures like d' or K is done more accurately with a full ROC. For instance, when we fit an equal variance signal detection model to Tim's and John's full empirical ROCs, rather than just a single point, to recover d', we correctly recover a higher d' for Tim than John (1.70 and 1.07, respectively). Thus, even if all we wish is to calculate the same metrics we use in old/new (i.e., K, d', A'), ROC analysis provides a major improvement on old/new precisely because it allows researchers to assess performance across levels of response bias. It is, of course, still critical for researchers who adopt a specific theory to justify their choice of model (e.g., via model comparison), just as they should when analyzing old/new data, and to be cognizant of how their modeling approach can affect their measurement of memory. For example, the MATLAB ROC toolbox (Koen et al., 2017) could be used for fitting

theoretically informed models to ROCs and doing model comparison, or simply to visualize that the measure of interest actually capture the shape of the data adequately.

How to measure memory more accurately: Overall recommendations

The best measurement approach will ultimately depend on the researcher's goal; however, we make two major recommendations for improving measurement practices. First, forced-choice tasks provide an unequivocally better way to measure memory than old/new tasks. Forced-choice has the same benefits as ROC analysis for those solely interested in which conditions or participants have the best overall memory performance, and it is simpler to implement and analyze data from such tasks. Thus, for those researchers who are interested in purely assessing memory performance and how it varies across conditions, stimuli or individuals, we strongly recommend forced-choice tasks like 2-AFC, where both an old and new item are presented on each trial in a counterbalanced spatial location and participants must indicate which is old. In fact, if there is one overarching recommendation emerging from our inquiry into the measurement of memory, it is this: use 2-AFC whenever possible.

ROC analysis is also unequivocally a better method than old/new tasks for measuring memory performance, and it has unique advantages despite also being more complex than 2-AFC. For example, ROC analysis provides a way of assessing the entire distribution of a subject's memory signals (often useful in testing theories of where the underlying signals come from), and thus is a considerably more accurate method for isolating memory discriminability from response bias than metrics based on old/new data alone. It provides major additional clarity on what old/new can only reason about as a counterfactual: which participants and conditions have the highest hit rate when false alarm rates are matched, and ultimately which have the strongest memories.

Although both methods have some assumptions and potential pitfalls, they rely on relatively non-overlapping assumptions. Thus, hyper-scrupulous researchers who want to be 100% certain an effect is due to a difference in memory strength per se (e.g., because it is being used to make policy recommendations) could ensure they obtain the same results with both ROC analysis and forced-choice tasks, which would allay any worries that the experimental results are artifacts of the methods employed.

What theory-based measures are best for default analysis?

Throughout this paper, we have argued that more careful measurement is needed, regardless of what theory of memory people subscribe to. That said, many situations call for theoretically informed measures – for example, when faced with old/new data with no confidence, or when wishing to interpret the magnitude of a difference in performance in forced-choice. What should be the preferred theory of memory in such cases? We believe the evidence shows that when all else is equal (i.e., where researchers do not have strong evidence for an alternative view being more appropriate), the default for memory research should be signal detection-based measures (like d_a and d').

Why signal detection theory? By default, researchers in many subfields of memory tend to intuitively compute measures of memory that treat memory as extremely discrete (i.e.,

threshold models, which assume memories are simply *present* or *absent*), simply because we are likely all used to thinking in a discrete, all-or-none way (Wixted, 2020). However, in the case of memory, there is significant evidence that favors the idea that memories vary in strength – and that this is true for both items you genuinely have seen before and ones you genuinely have not (e.g., Kellen et al., 2021; Wixted, 2007). Such variation in the strength evoked by genuinely old and genuinely new items is the core claim of signal detection theory.

Consider Fig. 9, which shows two previously seen cartoon people and two never-before-seen "real" people. If you had to say which cartoon person you were most confident you had seen before, signal detection theory says this is clearly a question you can answer: that is, having seen the image of "Tim" many times in the paper, but the image of "John" only once, your memory for the Tim image is *stronger* than your memory for the John image, and so you should be both more confident in your assessment that the Tim image is old and more likely to correctly identify it as old. By contrast, this same intuition is difficult to instantiate in threshold theories, where Tim and John must both either be in or out of memory, but memories cannot vary in strength.

What about the opposite – the two images you have never seen before? Both of the images on the left side of Fig. 9 were generated from StyleGAN (Karras et al., 2020), a deep network that can "imagine" images of people who do not exist. Neither is an image of an actual person, and so we are quite sure you have never seen either before. But if forced to choose, which would you say feels more familiar, evoking a stronger memory signal? We suspect most people would say the right person evokes a stronger memory signal – and they would be quite confident in that assessment, and more likely to falsely claim they had previously seen this image before (even though this is impossible). In this case, this happens because this image happens to be a doppelganger of Barack Obama (Suchow & Peterson, 2019). However, more generally, any unseen items may feel more or less familiar to particular observers due to many factors, including simply because there is noise in both the perceptual and memory system.

The core intuition that memories vary in strength for both genuinely old and genuinely new item, as well as a huge variety of empirical evidence, such as the general structure of ROCs and the neural instantiation of memories (for a review, see Wixted, 2020) has led to signal detection theory being a dominant framework for theorizing about memory decisions. Thus, when forced to choose a default model for memory measurement, signal detection-based measures should almost certainly be preferred over those based on threshold views or that do not derive from a theory directly. In particular, unequal variance signal detection models – where the distribution of memory strengths for "previously seen" items varies more than the distribution for unseen items – provide a good account of ROC curves and many memory scenarios (Wixted & Mickes, 2020), and can be used even if the origin of the memory signals used in the decision is not unitary (e.g., if people rely on both "familiarity' and "recollection' to derive a strength for a memory: Wixted, 2010). Thus, measures like d_a, which instantiate discriminability in unequal variance signal detection models and can be computed from ROCs, are likely to be broadly appropriate as theoretically informed measures of memory strength. And in forced-choice tasks, which are

necessarily “equal variance” because of the way they are designed, d' is likely to be more appropriate than metrics based on threshold-based views. In old/new tasks, where d_a cannot be computed, d' is likely better than the alternatives, even if it is almost certainly imperfect. In such scenarios, researchers may use hit and false alarm rates and calculate “possible” values of d_a using principled assumptions regarding how variance may vary as a function of experimental manipulations and/or individuals (e.g., one could ask if the findings are robust if the seen-items distribution has 1.2 times the standard deviation of the unseen-items distribution). If results are robust across these assumptions, then researchers may report so, whereas otherwise, we suggest that they collect data with confidence.

Of course those interested in the full distribution of latent memory signals and the underlying memory representations that give rise to them may prefer other theoretically informed models (e.g., dual process models of recognition memory, Parks & Yonelinas, 2009; Yonelinas, 1994, 2002; or mixture models of working memory; Adam et al., 2017; Zhang & Luck, 2008) that reject some or all aspects of signal detection theory (for related work, see also, e.g., Kellen & Klauer, 2015; Province & Rouder, 2012; Rouder et al., 2008). They may also prefer to model memory as more than simply strength and instead think about the actual features stored and how they are used (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). We suggest signal detection theory measures only as a reasonable default that is unlikely to lead you entirely wrong if you are interested in measuring memory performance (i.e., most ROCs, even if they have a different underlying basis, are well approximated by unequal variance signal detection models), not as the only way to conceive of memory.

Finally, it is important to note that, while signal detection theory postulates that memory signals are distributed along a single dimension (of memory strength), this does not entail that signal detection theory is solely compatible with the view that memory is a unidimensional construct. That is, signal detection theory describes how people use information to make judgments in memory tasks. In particular, it postulates that information is combined into a unitary decision variable when making memory judgments along a unitary axis (e.g., when asked to report confidence in an item being old or new). Such a view is fully compatible with the assumption that memory itself is multidimensional and that multiple sources of information from different channels are integrated and used to make such memory decisions (for extended discussion on this topic, see Wixted & Mickes, 2010).

General discussion: Remaining issues

In this section we discuss a few remaining issues that pertain to best practices in measurement. First, researchers may still wonder if it is important to choose between measurements if, in practice, even suboptimal measures may sometimes lead researchers to the “right” conclusion. For instance, not infrequently, various metrics based on old/new tasks and ROC analysis will converge. Does this mean that choosing between them is a moot point? We have attempted to show that the answer is a definitive “no”: choosing between these metrics in a principled way is both theoretically and practically imperative. Throughout this paper we used hypothetical and real-world examples that demonstrate how different measures can lead to drastically different conclusions, as well as how choosing

between them can have potentially large theoretical and practical implications. For instance, different measures can lead to qualitatively different conclusions regarding how to measure eyewitness memory optimally, as well as how individual differences in visual working memory capacity relate to other indices of cognitive function. It is true that measures based on different models will often be aligned with one another. However, this fact does not justify an unprincipled use of metrics within any research domain. Even if in some scenarios a suboptimal metric leads to the "right" conclusion, this does not warrant the sweeping generalization that it will do so across all variations in experimental procedures and analyses.

A second related concern is whether the alternative measurement approaches we promote, such as AUC or the use of forced-choice designs, are truly superior to other commonly used old/new metrics, such as those based on threshold models (K is visual-working memory experiments). For instance, as we discussed, the interpretation of ROC data involves accepting some auxiliary assumptions regarding how people map memory states to confidence judgments. Importantly, our claim is not that these alternative measures are assumption free; collecting and interpreting any measure in psychology involves making auxiliary as well as simplifying assumptions, which may be incorrect to some degree. This latter point is not specific to the measures we consider – it is an inherent problem of quantifying hypothetical, unobservable constructs (e.g., Kellen et al., 2021). Our goal is to identify and evaluate the hidden assumptions behind mainstream metrics in memory research and provide a usable guide for researchers to improve on routine measurement practices.

Third, we underscore that the intended scope of our article is limited in several ways. First, we do not discuss these measurement issues in the context of other research domains, such as perception and decision-making, where researchers may also seek to separate sensitivity from response bias. It is certainly true that the measurement issues we raise apply to any attempt to separate sensitivity from response bias. We focus on memory research because, as reviewed, these measurement issues arise in the study of a wide range of topics including research on different memory systems, relationship between individual differences in memory and other indices of cognitive function, and a range of memory phenomena. Therefore, in our view, these issues are extremely prominent in the recognition memory domain, as evidenced by the number of reviewed papers that use different metrics for quantifying memory performance in such tasks.

Likewise, we point out that our paper focuses on a single type of memory task, that is, recognition memory tasks, change detection tasks and eyewitness line-ups. As noted, we focus on recognition memory tasks because they are extremely prevalent. It is also the default talk across several research domains in working memory and long-term memory, which permits us to bridge these measurement issues in a comprehensive overview. Nevertheless, we acknowledge that separate measurement issues arise in the study of recall memory. For instance, in free recall tasks researchers must postulate theoretical assumptions to determine how to quantify "semantic relatedness' and assess clustering of related items (e.g., Shuell, 1975; Stricker et al., 2002). Likewise, researchers must determine whether and how to analyze false memory intrusions (as reviewed in Cleary, 2018). These measurement

questions fall outside of the scope of separating sensitivity from response bias, however, they are also important for understanding memory processes that go beyond the study of memory discrimination (the focus of the current article), such as dynamics of retrieval.

Finally, we note that our article serves as a kind of tutorial for the issues that matter when measuring recognition memory, but we do not attempt to provide guidance on how to interpret measures of memory once these are obtained. Notably, recent work by Starns et al. (2019) suggests that researchers can vary substantially in how they interpret the same results from recognition memory studies, a phenomenon coined the *inference crisis*. This divergence may be due in part because researchers vary substantially in their auxiliary assumptions, their rankings of which auxiliary assumptions are less or more plausible (Strevens, 2020), their incentives and understanding of methodology, and theory. These authors propose a blinded-inference procedure as one way of dealing with the inference crisis. In this procedure researchers are asked to make inferences about experimental manipulations rather than (assumed) latent variables, as well as to formally communicate their degree of certainty in their inferences. We endorse this as a promising approach towards improving theorizing in the recognition memory domain as well as social sciences more broadly.

Conclusion

We have argued that problematic memory measurement is common in recognition memory research, including both working memory and long-term memory. People in many subfields regularly use tasks and metrics that are known to be poor measures of underlying memory strength – and do so seemingly without careful, theoretically informed consideration of their decision. We have focused in particular on the difficulties of understanding memory using “old/new” tasks, but also pointed to the difficulties raised by other tasks. Overall, we demonstrate that despite a large literature on how to properly measure memory performance, spanning decades, it remains common to measure memory incorrectly some, or even most, of the time. This has profound implications for both theory building and policy making because these “mainstream” but problematic measurement approaches can lead to qualitatively incorrect conclusions regarding how experimental manipulations affect memory and how memory varies as a function of individual differences like intelligence. We explained how measuring memory accurately requires a comparison between items that have genuinely been seen and ones that have not, and making such comparisons accurately requires knowledge of the full underlying, latent distribution of memory signals. This means simply asking someone whether they recognize something – as in “old/new” or “change detection” tasks – cannot be used to accurately measure memory, despite the fact that these are some of the most prominent tasks in the memory literature.

In our view, this is something of a “crisis” of measurement: Even though psychology is designed to be a cumulative science, memory researchers routinely employ measures that either do not measure the latent variables they set out to study, or do so only under certain, unusual conditions, which they do not check for. Although similar points have been repeatedly raised about memory research in the past (e.g., Snodgrass & Corwin, 1988; Rotello, Heit, & Dube, 2015; Wixted & Mickes, 2018) the continued prevalence

of the old/new task in memory research indicates that scientists are generally unaware of this fundamental problem in measurement or do not take it seriously. The cumulative result of this research practice is that entire literatures report measures that likely fail to adequately capture the actual strength of memory (e.g., diagnosticity ratio, Mickes et al., 2014; K values: Williams et al., 2021). The general problem of developing rigorous measurement tools for theory building in the social sciences has been pointed out many times before (Eronen & Bringmann, 2021; Flake & Fried, 2020; Guest & Martin, 2021; Kellen, Davis-Stober, Dunn, & Kalish, 2021; Luce & Krumhansl, 1988; Meehl, 1967; Oberauer & Lewandowsky, 2019; Rotello et al., 2015; Regenwetter & Robinson, 2017; Scheel et al., 2021), yet it remains a pernicious issue across all domains of memory research.

Thus, we suggest that in order for the psychology and neuroscience of memory to become a cumulative, theory-driven science, much more attention should be given to measurement issues. For everyday memory research interested in overall memory performance, we make a particularly concrete suggestion: that the default memory task should change from old/new ("is this item 'old' or 'new'?") to two-alternative forced-choice ("which of these two items is old?"). We also provided pointers for how to implement ROC analysis where appropriate, such as when the distribution of latent memory signals or the underlying nature of memory representations that gives rise to such memory signals is of interest. Finally, we suggest that signal detection theory is a useful default theory for memory research, and should be preferred unless there is a theoretically informed reason to suppose a different way of analysis is more appropriate in a given situation.

Funding

TFB was supported by NSF BCS-1653457 and NSF BCS-1829434. M.M.R. was supported by an NIH F32 NRSA postdoctoral fellowship (award number: 1F32MH127823-01). J.R.W. was supported by an NSF GRFP.

Data Availability

Not applicable.

References

Adam KC, Vogel EK, & Awh E (2017). Clear evidence for item limits in visual working memory. Cognitive Psychology, 97, 79–97. [PubMed: 28734172]

Alvarez GA, & Cavanagh P (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological Science, 15, 106–111. [PubMed: 14738517]

Alvarez GA, & Cavanagh P (2008). Visual short-term memory operates more efficiently on boundary features than it does on the surface features. Perception & Psychophysics, 70, 346–364. [PubMed: 18372755]

Aly M, & Turk-Browne NB (2018). Flexible weighting of diverse inputs makers hippocampal function malleable. Neuroscience Letters, 680, 13–22. [PubMed: 28587901]

Aminoff EM, Clewett D, Freeman S, Frithsen A, Tipper C, Johnson A, Grafton ST, & Miller MB (2012). Individual differences in shifting decision criterion: A recognition memory study. Memory and Cognition, 40, 1016–1030. [PubMed: 22555888]

Awh E, Barton B, & Vogel EK (2007). Visual working memory represents a fixed number of items, regardless of complexity. Psychological Science, 18, 622–628. [PubMed: 17614871]

Bainbridge WA, Isola P, & Oliva A (2013). The intrinsic memorability of face images. Journal of Experimental Psychology: General, 142, 1323–1334. [PubMed: 24246059]

Balaban H, Fukuda K, & Luria R (2019). What can half a million change detection trials tell us about visual working memory? Cognition, 191, 103984. [PubMed: 31234117]

Bays PM, Catalo RFG, & Hussain M (2009). The precision of visual working memory is set by allocation of shared resource. Journal of Vision, 9, 1–11. [PubMed: 19761316]

Benjamin A, & S., & Bjork RA (2000). On the relationship between recognition speed and accuracy for words rehearsed via rote versus elaborative rehearsal. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 638–648. [PubMed: 10855422]

Bjork EL, & Bjork RA (2003). Intentional forgetting can increase, not decrease, residual influences of to-be-forgotten information. Journal of Experimental Psychology: Learning, Memory, and Cognition, 524–531. [PubMed: 12924855]

Blackwell HR (1953). Psychological thresholds: Experimental studies of methods of measurement. University of Michigan, engineering research institute bulletin, 36. Ann Arbor: University of Michigan Press

Borota D, Murray E, Keceli G, Cang A, Watabe JM, Ly M, Toscano JP, & Yassa MA (2014). Post-study caffeine administration enhances memory consolidation in humans. Nature Neuroscience, 17, 201–203. [PubMed: 24413697]

Bower GH, & Holyoak K (1973). Encoding and recognition memory for naturalistic sounds. Journal of Experimental Psychology, 101, 360–366. [PubMed: 4753861]

Bowman CR, & Dagmar Z (2020). Training set coherence and set size effects on concept generalization and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 46, 1442–1464. [PubMed: 32105147]

Brady TF, & Alvarez GA (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. Journal of Experimental Psychology: Learning, Memory and Cognition, 41, 921–929. [PubMed: 25419824]

Brady TF, Alvarez G, & Störmer V (2019). The role of meaning in visual memory: Face-selective brain activity predicts memory for ambiguous face stimuli. Journal of Neuroscience, 39, 1100–1108. [PubMed: 30541914]

Brady TF, Konkle T, Alvarez GA, & Oliva A (2008). Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences, USA, 105, 14325–14329.

Brady TF, Shafer-Skelton A, & Alvarez GA (2017). Global ensemble texture representations are critical to rapid scene perception. Journal of Experimental Psychology: Human Perception and Performance, 43, 1160–1176. [PubMed: 28263635]

Brady TF, Störmer V, & Alvarez GA (2016). Working memory is not fixed capacity: More active storage capacity for real-world objects than simple stimuli. Proceedings of the National Academy of Sciences, 113, 7459–7464.

Brady T, & Störmer VS (2020). Comparing memory capacity across stimuli requires maximally dissimilar foils: Using deep convolutional neural networks to understand visual working memory capacity for real-world objects. PsyArxiv.

Buchanan TW, & Adolphs R (2002). The role of the human amygdala in emotional modulation of long-term declarative memory. In Moore SC & Oaksford M (Eds.), Advances in consciousness research, Vol. 44. Emotional cognition: From brain to behaviour (p. 9–34). John Benjamins Publishing Company.

Cappell KA, Gmeindl L, & Reuter-Lorenz PA (2010). Age differences in prefrontal recruitment during verbal working memory maintenance depend on memory load. Cortex, 46, 462–473. [PubMed: 20097332]

Castella J, Pina R, Baques J, & Allen RJ (2020). Differential effects of working memory load on priming and recognition of real images. Memory and Cognition, 48, 1460–1471. [PubMed: 32601843]

Chan JCK, & McDermott KB (2007). The testing effect in recognition memory: A dual process account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(2), 431–437. [PubMed: 17352622]

Chubala CM, Guitard D, Neath I, Saint-Aubin J, & Surprenant AM (2020). Visual similarity effects in immediate serial recall and (sometimes) in immediate serial recognition. Memory & Cognition, 48, 411–425. [PubMed: 31701325]

Chunharas C, Rademaker RL, Sprague TC, Brady TF, & Serences J (2019). Separating memoranda in depth increases visual working memory performance. Journal of Vision, 19. 10.1167/19.1.4

Clark SE, & Wells GL (2008). On the diagnosticity of multiple-witness identifications. Law and Human Behavior, 32, 406–422. [PubMed: 18095147]

Cleary AM (2018). Dependent measures in memory research: From free recall to recognition. In Handbook of research methods in human memory (pp. 19–35). Routledge.

Cohen AL, Sanborn AN, & Shiffrin RM (2008). Model evaluation using grouped or individual data. Psychonomic Bulletin & Review, 15, 692–712. [PubMed: 18792497]

Conway AR, Cowan N, Bunting MF, Therriault DJ, & Minkoff SR (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. Intelligence, 30(2), 163–183.

Cortese MJ, McCarty DP, & Schock J (2015). A mega recognition memory study of 2897 disyllabic words. Quarterly Journal of Experimental Psychology, 68, 1489–1501.

Cowan N (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24, 87–114. [PubMed: 11515286]

Cowan N, Blume CL, & Saults JS (2013). Attention to attributes and objects in working memory. Journal of Experimental Psychology: Learning, Memory and Cognition, 39, 731–747. [PubMed: 22905929]

Cox JC, & Dobbins IG (2011). The striking similarities between standard, distractor-free, and target-free recognition. Memory & Cognition, 39, 925–940. [PubMed: 21476108]

Cunningham CA, Yassa MA, & Egeth HE (2015). Massive memory revisited: Limitations on storage capacity for object details in visual long-term memory. Learning and Memory, 22, 563–566. [PubMed: 26472646]

De Brigard FD, Brady TF, Ruzic L, & Schacter DL (2017). Tracking the emergency of memories: A category-learning paradigm to explore schema-driven recognition. Memory and Cognition, 45, 105–120. [PubMed: 27496024]

DeCarlo L (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. Journal of Mathematical Psychology, 54, 304–313.

Delay CG & Wixted JT (2021). Discrete-state vs. continuous models of the confidence-accuracy relationship in recognition memory. Psychonomic Bulletin & Review, 28, 556–564. [PubMed: 33111256]

Diana RA, Peterson MJ, & Reder LM (2004). The role of spurious feature familiarity in recognition memory. Psychonomic Bulletin & Review, 11, 150–156. [PubMed: 15117001]

Donkin C, Tran SC, & Nosofsky RM (2014). Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. Attention, Perception & Psychophysics, 76, 2103–2116.

Dougal S, & Rotello CM (2007). “remembering” emotional words is based on response bias, not recollection. Psychonomic Bulletin & Review, 14, 423–429. [PubMed: 17874582]

Dube C, Rotello CM, & Heit E (2011). The belief bias effect is aptly named: A reply to Klauer and Kellen. Psychological Review, 118(1), 155–163. [PubMed: 21244191]

Endress AD, & Potter MC (2014). Large capacity temporary visual memory. Journal of Experimental Psychology: General, 143, 548–565. [PubMed: 23937181]

Engle RW (2002). Working memory capacity as executive attention. Current Directions in Psychological Science, 11, 19–23.

Engle RW (2018). Working memory and executive attention: A revisit. Perspectives on Psychological Science, 13, 190–193. [PubMed: 29592654]

Engle RW, & Kane MJ (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In Ross BH (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 44, pp. 145–199). Elsevier Science.

Eriksson J, Vogel EK, Lansner A, Bergström F, & Nyberg L (2015). Neurocognitive architecture of working memory. Neuron, 88, 33–46. [PubMed: 26447571]

Eronen MI, & Bringmann LF (2021). The theory crisis in psychology: How to move forward. Perspectives on Psychological Science.

Estes WK (1956). The problem of inference from curves based on group data. Psychological Bulletin, 53, 134–140. [PubMed: 13297917]

Estes WK, & Maddox WT (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. Psychonomic Bulletin & Review, 12, 403–408.

Fisher AV, & Sloutsky VM (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. Child Development, 76, 583–597. [PubMed: 15892780]

Flake JK, & Fried EI (2020). Measurement Schmeasurement: Questionable measurement practices and how to avoid them. PsyArXiv.

Fougnie D, Suchow JW, & Alvarez GA (2012). Variability in the quality of visual working memory. Nature Communications, 3, 1229.

Fukuda K, Awh E, & Vogel EK (2010a). Discrete capacity limits in visual working memory. Current Opinion in Neurobiology, 20, 177–182. [PubMed: 20362427]

Fukuda K, Woodman GF, & Vogel EK (2015). Individual differences in visual working memory capacity: Contributions of attentional control to storage. Mechanisms of sensory working memory: Attention and perfomance XXV, 105.

Fukuda K, Kang MS, & Woodman GF (2016a). Distinct neural mechanisms for spatially lateralized and spatially global visual working memory representations. Journal of Neurophysiology, 116, 1715–1727. [PubMed: 27440249]

Fukuda K, & Vogel EK (2019). Visual short-term memory capacity predicts the “bandwidth” of visual long-term memory encoding. Memory & Cognition, 47, 1481–1497. [PubMed: 31236821]

Fukuda K, Vogel E, Mayr U, & Awh E (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. Psychonomic Bulletin & Review, 17(5), 673–679. [PubMed: 21037165]

Fum D, Del Missier F, & Stocco A (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words [editorial]. Cognitive Systems Research, 8, 135–142.

Galvin SJ, Podd JV, & Whitmore J (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychonomics Bulletin and Review, 10, 843–876.

Gao Y, & Theeuwes J (2020). Learning to suppress a distractor is not affected by working memory load. Psychonomic Bulletin & Review, 27, 96–104. [PubMed: 31797259]

Gardiner JM, & Java RI (1991). Forgetting in recognition memory with and without recollective experience. Memory and Cognition, 19, 617–623. [PubMed: 1758306]

Gardiner JM, Kaminska Z, Dixon M, & Java RI (1996). Repetition of previously novel melodies sometimes increases both remember and know responses in recognition memory. Psychonomic Bulletin and Review, 3, 366–371. [PubMed: 24213939]

Geiselman RE, & Bjork RA (1980). Primary versus secondary rehearsal in imagined voices: Differential effects on recognition. Cognitive Psychology, 12, 188–205. [PubMed: 7371376]

Glanzer M, Kim K, Hilford A, & Adams JK (1999). Slope of the receiver-operating characteristic in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25(2), 500.

Goshen-Gottstein Levy & Rotello(2019). Talk presented at the 60th annual meeting of the Psychonomic society, Montreal

Greene B, & Soto. (2010). Interplay between affect and arousal in recognition memory. PLoS One.

Green DM, & Swets JA (1966). Signal detection theory and psychophysics (Vol. 1, pp. 1969–2012). New York: Wiley.

Green DM (2020). A homily on signal detection theory. The Journal of the Acoustical Society of America, 148, 222. [PubMed: 32752757]

Green DM, & Swets JA (1988). Signal detection theory and psychophysics (reprint ed.). Peninsula Publishing.

Guest O, & Martin AE (2021). How computational modeling can force theory building in psychological science. Perspectives on Psychological Science.

Hakim N, Adam KCS, Gunseli E, Awh E, & Vogel EK (2019). Dissecting the neural focus of attention reveals distinct processes for spatial attention and object-based storage in visual working memory. Psychological Science, 30, 526–540. [PubMed: 30817220]

Harthsorne JK, & Makovski T (2019). The effect of working memory maintenance on long-term memory. Memory and Cognition, 47, 749–763. [PubMed: 31073790]

He K, Li J, Wu F, Wan X, Gao Z, & Shen M (2020). Object-based attention in retaining binding in working memory: Influence of activation states of working memory. Memory & Cognition. 10.3758/s13421-020-01038-0

Henderson JM, & Hollingworth A (2003). Eye movements and visual memory: Detecting changes to saccade targets in scenes. Perception & Psychophysics, 65, 58–71. [PubMed: 12699309]

Hudon C, Belleville S, & Gauthier S (2009). The assessment of recognition memory using the remember/know procedure in amnestic mild cognitive impairment and probable Alzheimer's disease. Brain and Cognition, 70, 171–179. [PubMed: 19250730]

Irwin DE (2014). Short-term memory across eye blinks. Memory, 22, 898–906. [PubMed: 24147932]

Isola P, Xiao J, Torralba A, & Oliva A (2011). What makes an image memorable? IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 145–152.

Jacoby LL, Shimizu Y, Daniels KA, & Rhodes MG (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. Psychonomic Bulletin & Review, 12, 852–857. [PubMed: 16524001]

Jang Y, Mickes L, & Wixted JT (2012). Three tests and three corrections: Comment on Koen and Yonelinas (2010). Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 513–523. [PubMed: 22390323]

Jiang YV, Remington RW, Asaad A, Lee HJ, & Mikkalson TC (2016). Remembering faces and scenes: The mixed-category advantage in visual working memory. Journal of Experimental Psychology: Human Perception and Performance, 42, 1399–1411. [PubMed: 27123683]

Jiménez L, Méndez C, Agra O, & Ortiz-Tudela J (2020). Increasing control improves further control, but it does not enhance memory for the targets in a face–word stroop task. Memory & Cognition, 48, 994–1006. [PubMed: 32144648]

Johnson MK, Mitchell KJ, Raye CL, & Greene EJ (2004). An age-related deficit in prefrontal cortical function associated with refreshing information. Psychological Science, 15, 127–132. [PubMed: 14738520]

Johnson MK, Reeder JA, Raye CL, & Mitchell KJ (2002). Second thoughts versus second looks: An age-related deficit in reflectively refreshing just-activated information. Psychological Science, 13, 64–67. [PubMed: 11892780]

Johnson JS, Spencer JP, Luck SJ, & Schöner G (2009). A dynamic neural field model of visual working memory and change detection. Psychological Science, 20, 568–577. [PubMed: 19368698]

Jou J, Flores S, Cortes HM, & Leka BG (2016). The effects of weak versus strong relational judgments on response bias in two-alternative-forced-choice recognition: Is the test criterion-free? Acta Psychologica, 167, 30–44. [PubMed: 27104925]

Juola JF, Caballero-Sanz A, Munoz-Garcia AR, Botella J, & Suero M (2019). Familiarity, recollection, and receiver-operating characteristic (ROC) curves in recognition memory. Memory and Cognition, 47, 855–876. [PubMed: 30949925]

Kantner J, & Lindsay DS (2012). Response bias in recognition memory as a cognitive trait. Memory & Cognition, 40, 1163–1177. [PubMed: 22872581]

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, & Aila T (2020). Analyzing and improving the imagine quality of stylegan. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8110–8119.

Khader P, Ranganath C, Seemuller A, & Rosler F (2007). Working memory maintenance contributes to long-term memory formation: Evidence from slow event-related brain potentials. Cognitive, Affective, & Behavioral Neuroscience, 7, 212–224.

Kellen D, Davis-Stober CP, Dunn JC, & Kalish MJ (in press). The problem of coordination and the pursuit of structural constraints in psychology Perspectives on Psychological Science.

Kellen D, & Klauer KC (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. Psychological Review, 122(3), 542. [PubMed: 26120910]

Kellen D, Winiger S, Dunn JC, & Singmann H (2021). Testing the foundations of signal detection theory in recognition memory. Psychological Review, 128(6), 1022–1050. [PubMed: 34110843]

Koen JD, Barrett FS, Harlow IM, & Yonelinas AP (2017). The ROC Toolbox: A toolbox for analyzing receiver-operating characteristics derived from confidence ratings. Behavior research methods, 49(4), 1399–1406. [PubMed: 27573007]

Kroll NEA, Yonelinas AP, Dobbins IG, & Frederick CM (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. Journal of Experimental Psychology: General, 131(2), 241–254. [PubMed: 12049242]

Kuczma M (2009). Inequalities (pp. 197–226). Birkhäuser Basel.

Lamont AC, Stewart-Williams S, & Podd J (2005). Face recognition and aging: Effects of target age and memory load. Memory & Cognition, 33, 1017–1024. [PubMed: 16496722]

Lee HJ, & Cho YS (2019). Memory facilitation for emotional faces: Visual working memory trade-offs resulting from attention preference for emotional facial expressions. Memory and Cognition, 47, 1231–1243. [PubMed: 30977105]

Lind SE, & Bowler DM (2009). Recognition memory, self-other source memory, and theory-of-mind in children with autism spectrum disorder. Journal of Autism and Developmental Disorders, 39, 1231–1239. [PubMed: 19353262]

Lindsay RC, & Wells GL (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. Journal of Applied Psychology, 70, 556–564.

Loiotile RE, & Courtney SM (2015). A signal detection theory analysis of behavioral pattern separation paradigms. Learning and Memory, 22, 364–369. [PubMed: 26179230]

Loftus GR (1978). On interpretation of interactions. Memory & Cognition, 6, 312–319.

Luce RD, & Krumhansl CL (1988). Measurement, scaling, and psychophysics. In Atkinson RC, Herrnstein RJ, Lindzey G, & Luce RD (Eds.), Stevens' handbook of experimental psychology: Perception and motivation (pp. 3–74). Learning and cognition.

Luck SJ, & Vogel EK (1997). The capacity of visual working memory for features and conjunctions. Nature, 390, 279–280. [PubMed: 9384378]

Luria T, & Vogel EK (2011). Shape and color conjunction stimuli are represented as bound objects in visual working memory. Neuropsychologia, 49, 1632–1639. [PubMed: 21145333]

Macmillan NA, & Creelman CD (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. Psychological Bulletin, 107, 401–413.

Macmillan NA, & Creelman CD (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. Psychonomic Bulletin & Review, 3, 164–170. [PubMed: 24213864]

Malmberg KJ (2002). On the form of the ROCs constructed from confidence ratings. Journal of Experimental Psychology: Learning, Memory and Cognition, 28, 380–387. [PubMed: 11911394]

MacLin OH, & MacLin MK (2004). The effect of criminality on face attractiveness, typicality, memorability and recognition. North American Journal of Psychology, 6, 145–154.

Macmillan NA, & Creelman CD (2004). Detection theory: A user's guide. Psychology press.

Makovski T, Watson LM, Koutstaal W, & Jiang YV (2010). Method matters: Systematic effects of testing procedure on visual working memory sensitivity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1466–1479. [PubMed: 20854011]

Maxcey-Richard AM, & Hollingworth A (2013). The strategic retention of task-relevant objects in visual working memory. Journal of Experimental Psychology: Learning, Memory and Cognition, 39, 760–772. [PubMed: 22845068]

McClelland JL, & Chappell M (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. Psychological Review, 105, 724–760. [PubMed: 9830377]

Meehl PE (1967). Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103–115.

Mickes L, Flowe HD, & Wixted JT (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. Journal of Experimental Psychology: Applied, 18, 361–376. [PubMed: 23294282]

Mickes L, Hwe V, Wais PE, & Wixted JT (2011). Strong memories are hard to scale. Journal of Experimental Psychology: General, 140, 239–257. [PubMed: 21417544]

Mickes L, & Wixted JT (in press). Eyewitness memory. In Kahana MJ & Wagner AD (Eds.), Oxford handbook of human memory. Oxford University Press.

Mickes L, Moreland MB, Clark SE & Wixted JT (2014). Missing the information needed to perform ROC analysis? Then compute d', not the diagnosticity ratio. Journal of Applied Research in Memory and Cognition, 3, 58–62.

Mickes L, Wixted JT, & Wais PE (2007). A direct test of the unequal-variance signal-detection model of recognition memory. Psychonomic Bulletin & Review, 14, 858–865. [PubMed: 18087950]

Miller MB, & Kantner J (2020). Not all people are cut out for strategic criterion shifting. Current Directions in Psychological Science, 29, 9–15.

Monti JM, Cooke GE, Watson PD, Voss MW, Kramer AF, & Cohen NJ (2015). Relating hippocampus to relational memory processing across domains and delays. Journal of Cognitive Neuroscience, 27, 234–245. [PubMed: 25203273]

National Research Council. (2014). Identifying the culprit: Assessing eyewitness identification. The National Academies Press.

Oberauer K, & Lewandowsky S (2019). Addressing the theory crisis in psychology. Psychonomic Bulletin & Review, 26, 1596–1618. [PubMed: 31515732]

Otero SC, Weekes BS, & Hutton SB (2011). Pupil size changes during recognition memory. Psychophysiology, 48, 1346–1353. [PubMed: 21575007]

Parks CM, & Yonelinas AP (2009). Evidence for a memory threshold in second-choice recognition memory responses. Proceedings of the National Academy of Sciences, 106(28), 11515–11519.

Pashler H (1988). Familiarity and visual change detection. Perception & Psychophysics, 44, 369–378. [PubMed: 3226885]

Pastore RE, Crawley EJ, Berens MS, & Skelly MA (2003). "nonparametric" A′ and other modern misconceptions about signal detection theory. Psychonomic Bulletin & Review, 10, 556–569. [PubMed: 14620349]

Pailian H, Simons DJ, Wetherhold J, & Halberda J (2020). Using the flicker task to estimate visual working memory storage capacity. Attention, Perception and Psychophysics, 82, 1271–1289.

Parra MA, Della Sala S, Logie RH, & Morcom AM (2014). Neural correlates of shape-color binding in visual working memory. Neuropsychologia, 52, 27–36. [PubMed: 24120612]

Pessoa L, Gutierrez E, Bandettini P, & Ungerleider L (2002). Neural correlates of visual working memory: fMRI amplitude predicts task performance. Neuron, 35, 975–987. [PubMed: 12372290]

Police Executive Research. (2013). Forum https://www.policeforum.org/

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

Poon LW, & Fozard JL (1980). Age and word frequency effects in continuous recognition memory. Journal of Gerontology, 35, 77–86. [PubMed: 7350224]

Potter MC, Staub A, Raud J, & O'Connor DH (2002). Recognition memory for briefly presented pictures: The time course of rapid forgetting. Journal of Experimental Psychology: Human Perception and Performance, 28, 1163–1175. [PubMed: 12421062]

Postle BR, Druzgal TJ, & D'Esposito M (2003). Seeking the neural substrates of visual working memory storage. Cortex, 39, 927–946. [PubMed: 14584560]

Province JM, & Rouder JN (2012). Evidence for discrete-state processing in recognition memory. Proceedings of the National Academy of Sciences, 109(36), 14357–14362.

Rademaker RL, Tredway C, & Tong F (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. Journal of Vision, 12, 1–13.

Rahnev D, Desender K, Lee ALF, Adler WT, Aguilar-Lleyda D, Akdoğan B, Arbuzova P, Atlas LY, Balci F, Bang JW, Bègue I, Birney DP, Brady TF, Calder-Travis J, Chetverikov A, Clark TK, Davranche K, Denison RN, Dildine TC, & Zylberberg A (2020). The confidence database. Nature Human Behaviour, 4, 317–325.

Rajaram S, & Pereira-Pasarin LP (2007). Collaboration can improve individual recognition memory: Evidence from immediate and delayed tests. Psychonomic Bulletin & Review, 14, 95–100. [PubMed: 17546737]

Ratcliff R, Sheu CF, & Gronlund SD (1992). Testing global memory models using ROC curves. Psychological Review, 99(3), 518. [PubMed: 1502275]

Ratcliff R, & Starns JJ (2009). Modeling confidence and response time in recognition memory. Psychological review, 116(1), 59. [PubMed: 19159148]

Regenwetter M, & Robinson MM (2017). The construct-behavior gap in behavioral decision research: A challenge beyond replicability. Psychological Review, 124, 533–550. [PubMed: 28504522]

Reppa I, Williams KE, Greville WJ, & Saunders J (2020). The relative contribution of shape and colour to object memory. Memory and Cognition, 48, 1504–1521. [PubMed: 32542477]

Ricker TJ, Sandry J, Vergauwe E, & Cowan N (2020). Do familiar memory items decay? Journal of Experimental Psychology: Learning, Memory, and Cognition, 46, 60–76. [PubMed: 31107048]

Rich P, de Haan R, Wareham T, & van Rooji I (2021). How hard is cognitive science? PsyArXiv.

Robinson MM, Williams J, Brady TF (2022). What does it take to falsify a psychological theory? A case study on recognition models of visual working-memory. PsyArxiv.

Robinson MM, Benjamin AS, & Irwin DE (2020a). Is there a K in capacity? Evaluating the discrete-slot model of visual short-term memory. Cognitive Psychology.

Rotello CM, Macmillan NA, & Van Tassel G (2000). Recall-to-reject in recognition: Evidence from ROC curves. Journal of Memory and Language, 43, 67–88.

Rotello CM, Heit E, & Dubé C (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. Psychonomic Bulletin & Review, 22, 944–954. [PubMed: 25384892]

Rouder JN, Morey RD, Cowan N, Zwilling CE, Morey CC, & Pratte MC (2008). An assessment of fixed-capacity models of visual working memory. Proceedings of the National Academy of Sciences, 105, 5975–5979.

Rouder JN, Morey RD, & Pratte MS (2017). Bayesian hierarchical models of cognition. In Batchelder WH, Colonius H, Dzhafarov EN, & Myung J (Eds.), Cambridge handbooks in psychology. New handbook of mathematical psychology: Foundations and methodology (pp. 504–551). Cambridge University Press.

Rouhani N, Norman KA, Niv Y, & Bornstein AM (2020). Reward prediction errors create event boundaries in memory. Cognition, 203.

Sahakyan L, Waldum ER, Benjamin AS, & Bickett SP (2009). Where is the forgetting with list-method directed forgetting in recognition? Memory & Cognition, 37(4), 464–476. [PubMed: 19460953]

Scheel AM, Tiokhin L, Isager PM, & Lakens D (2021). Why hypothesis testers should spend less time testing hypotheses. Perspectives on Psychological Science.

Schurgin MW, & Brady TF (2019). When "capacity" changes with set size: Ensemble representations support the detection of across-category changes in visual working memory. Journal of Vision, 19, 1–12.

Scotti PS, Janakiefski L, & Maxcey AM (2020). Recognition-induced forgetting to schematically related pictures. Psychonomics Bulletin & Review.

Shiffrin RM, & Steyvers M (1997). A model for recognition memory: REM—Retrieving effectively from memory. Psychonomic Bulletin & Review, 4(2), 145–166. [PubMed: 21331823]

Shipstead Z, Lindsey DRB, Marshall RL, & Engle RW (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. Journal of Memory and Language, 72, 116–141.

Shoval R, Luria R, & Makovski T (2020). Bridging the gap between visual temporary memory and working memory: The role of stimuli distinctiveness. Journal of Experimental Psychology: Learning, Memory, and Cognition, 46, 1258–1269. [PubMed: 31647285]

Shuell TJ (1975). On sense and nonsense in measuring organization in free recall: Oops, pardon me, my assumptions are showing. Psychological Bulletin, 82, 720–724.

Sligte IG, Scholte HS, & Lamme VA (2008). Are there multiple visual short-term memory stores? PLoS One.

Sloutsky VM, & Fisher AV (2004). Induction and categorization in Young children: A similarity-based model. Journal of Experimental Psychology: General, 133, 166–188. [PubMed: 15149249]

Smith DG, & Duncan MJJ (2004). Testing theories of recognition memory by predicting performance across paradigms. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(3), 615–625. [PubMed: 15099130]

Smith RE, & Hunt RR (2020). When do pictures reduce false memory? Memory and Cognition, 48, 623–644. [PubMed: 31808050]

Snodgrass JG, & Corwin J (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. Journal of Experimental Psychology: General, 117, 34–50. [PubMed: 2966230]

Soro JC, Ferreira MB, Carneiro P, & Moreira S (2020). Memory illusions and category malleability: False recognition for goal derived reorganizations of common categories. Memory and Cognition, 48, 885–902. [PubMed: 32383150]

Spanton RW, & Berry CJ (2020). The unequal variance signal detection model of recognition memory: Investigating the encoding variability hypothesis. Quarterly Journal of Experimental Psychology.

Stark SM, Yassa MA, Lacy JW, & Stark CEL (2013). A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment. Neuropsychologia, 51, 2442–2449. [PubMed: 23313292]

Starns JJ, Chen T, & Staub A (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. Journal of Memory and Language, 93, 55–66.

Starns JJ, Cataldo AM, Rotello CM, Annis J, Aschenbrenner A, Bröder A, … & Wilson J (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. Advances in Methods and Practices in Psychological Science, 2(4), 335–349.

Steblay NK, Dysart JE, & Wells GL (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. Psychology, Public Policy, and Law, 17, 99–139.

Strevens M (2020). The knowledge machine: How irrationality created modern science. Liveright Publishing.

Stricker JL, Brown GG, Wixted JT, Baldo JV, & Delis D (2002). New semantic and serial clustering indices for the California verbal learning test 2: Background, rationale, and formulae. Journal of the International Neuropsychological Society, 8, 425–435. [PubMed: 11939700]

Suchow JW, & Peterson J (2019). 1,000 doppelgangers. https://suchow.io/1k-doppelgangers/. Accessed Jul 2021.

Swick D, & Knight RT (1999). Contributions of prefrontal cortex to recognition memory: Electrophysiological and behavioral evidence. Neuropsychology, 13, 155–170. [PubMed: 10353368]

Tas AC, Luck SJ, & Hollingworth A (2016). The relationship between visual attention and visual working memory encoding: A dissociation between covert and overt orienting. Journal of Experimental Psychology: Human Perception and Performance, 42, 1121–1138. [PubMed: 26854532]

Taylor R, & Bays PM (2020). Theory of neural coding predicts an upper bound on estimates of memory variability. Psychological Review, 127(5), 700–718. [PubMed: 32191074]

Toh YN, Sisk CA, & Jiang YV (2020). Effects of changing object identity on location working memory. Attention, Perception and Psychophysics, 82, 2862–2875.

Toner CK, Pirogovsky E, Kirwan CB, & Gilbert PE (2009). Visual object pattern separation deficits in nondemented older adults. Learning & Memory, 16, 338–342. [PubMed: 19403797]

Tulving E, & Thomson DM (1971). Retrieval processes in recognition memory: Effects of associative context. Journal of Experimental Psychology, 87, 116–124.

Turner BM, Forstmann BU, & Steyvers M (2019). Joint models of neural and behavioral data. Springer.

Unsworth N, Fukuda K, Awh E, & Vogel EK (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. Cognitive Psychology, 71, 1–26. [PubMed: 24531497]

Unsworth N, Fukuda K, Awh E, & Vogel EK (2015). Working memory delay activity predicts individual differences in cognitive abilities. Journal of Cognitive Neuroscience, 27, 853–865. [PubMed: 25436671]

van den Berg R, Shin H, Chou WC, George R, & Ma WJ (2012). Variability in encoding precision accounts for visual short-term memory limitations. Proceedings of the National Academy of Sciences of the United States of America, 29, 8780–8785.

Van Zandt T (2000). ROC curves and confidence judgments in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 582–600. [PubMed: 10855419]

Vogel E, & Awh E (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. Current Directions in Psychological Science, 17, 171–176.

Vogel EK, & Machizawa MG (2004). Neural activity predicts individual differences in visual working memory capacity. Nature, 428, 748–751. [PubMed: 15085132]

Wagenmakers E-J, Krypotos A-M, Criss AH, & Iverson G (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. Memory & Cognition, 40, 145–160. [PubMed: 22069144]

Wagner U, Kashyap N, Diekelmann S, & Born J (2007). The impact of post-learning sleep vs. wakefulness on recognition memory for faces with different facial expressions. Neurobiology of Learning and Memory, 87, 679–687. [PubMed: 17336554]

Weidemann CT, & Kahana MJ (2016). Assessing recognition memory using confidence ratings and response times. Royal Society open science, 3(4), 150670. [PubMed: 27152209]

Weidemann CT, & Kahana MJ (2019). Dynamics of brain activity reveal a unitary recognition signal. Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(3), 440. [PubMed: 30024265]

Wells GL, & Lindsay RC (1980). On estimating the diagnosticity of eyewitness nonidentifications. Psychological Bulletin, 8, 776–784.

Wells GL, Small M, Penrod S, Malpass RS, Fulero SM, & Brimacombe CAE (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. Law and Human Behavior, 22, 603–647.

Westerberg CE, & Marsolek CJ (2003). Sensitivity reductions in false recognition: A measure of false memories with stronger theoretical implications. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(5), 747–759. [PubMed: 14516210]

Wickens TD (2001). Elementary signal detection theory. Oxford University Press.

Williams J, Robinson M, Schurgin M, Wixted J, & Brady TF (2022). You can't "count" how many items people remember in working memory: The importance of signal detection-based measures for understanding change detection performance. PsyArxiv.

Wixted JT (2007). Dual-process theory and signal-detection theory of recognition memory. Psychological Review, 114, 152–176. [PubMed: 17227185]

Wixted JT (2020). The forgotten history of signal detection theory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 10.1037/xlm0000732

Wixted JT, & Mickes L (2010). A continuous dual-process model of remember/know judgments. Psychological Review, 117, 1025–1054. [PubMed: 20836613]

Wixted JT, & Mickes L (2018). Theoretical vs. empirical discriminability: The application of ROC methods to eyewitness identification. Cognitive Research: Principles and Implications.

Woodman GF, & Vogel EK (2008). Selective storage and maintenance of an object's features in visual working memory. Psychonomic Bulletin and Review, 15, 223–229. [PubMed: 18605507]

Yan X, Young AW, & Andrews TJ (2017). The automaticity of face perception is influenced by familiarity. Attention, Perception, & Psychophysics, 79, 2202–2211.

Yassa MA, & Stark CEL (2011). Pattern separation in the hippocampus. Trends in Neurosciences, 34, 515–525. [PubMed: 21788086]

Yin S, O'Neill K, Brady TF, & De Brigard F (2019). The effect for category learning on recognition memory: A signal detection theory analysis. Proceedings of the Cognitive Science Society.

Yonelinas AP (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(6), 1341. [PubMed: 7983467]

Yonelinas AP (2002). The nature of recollection and familiarity: A review of 30 years of research. Journal of Memory and Language, 46, 441–517.

Yonelinas AP, Dobbins I, Szymanski MD, Dhaliwal HS, & King L (1996). Signal detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. Consciousness & Cognition, 5, 418–441. [PubMed: 9063609]

Yonelinas AP, & Parks CM (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. Psychological Bulletin, 133(5), 800. [PubMed: 17723031]

Zhang W, & Luck SJ (2008). Discrete fixed-resolution representations in visual working memory. Nature, 453, 233–235. [PubMed: 18385672]

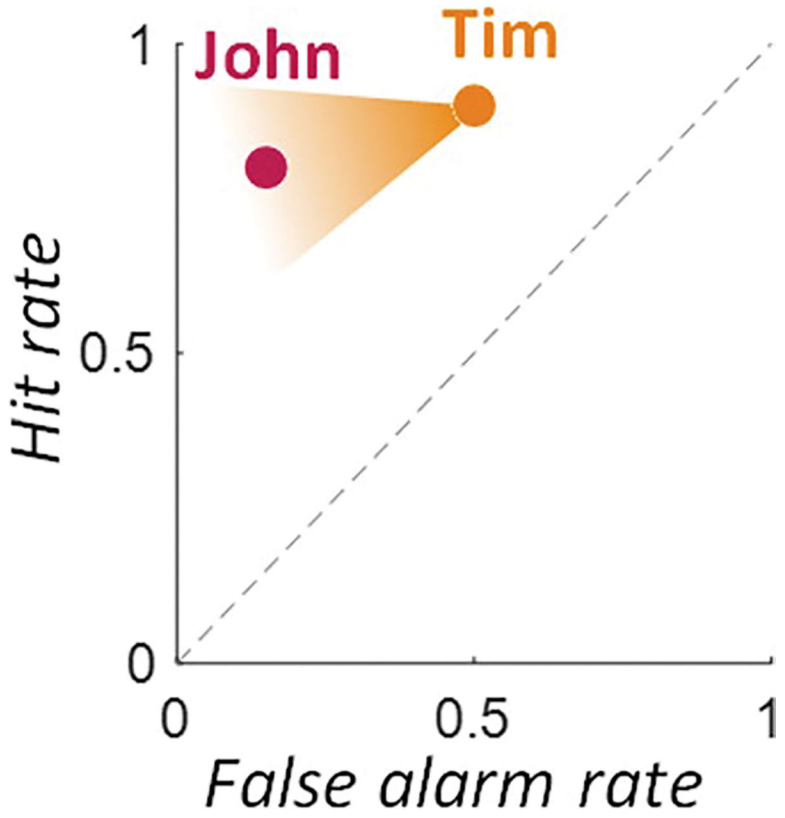


Fig. 1.

Tim had 900 hits and 500 false alarms (90% hit rate and 50% false alarm rate) and John had 800 hits and only 150 false alarms (80% hit rate, 15% false alarm rate). But the question of who had the better memory is actually "who would have more hits if they each had the same number of false alarms"? The core difficulty of old/new tasks – and measuring memory in general – is this problem of counterfactual reasoning. To know who had the better memory, we need to know who would have the better hit rate if they had the same false alarm rate, and many answers are possible and even plausible (full range of possibilities denoted by orange triangle). All of the metrics people use to combine hits and false alarms (A', d', adjusted hit probability, etc.) are different ways of answering this question, each coming up with different answers to the question "If we somehow forced Tim to have fewer false alarms, to match John, would his hit rate end up higher or lower than John's?" – and each ending up at different parts of the gradient of possible answers

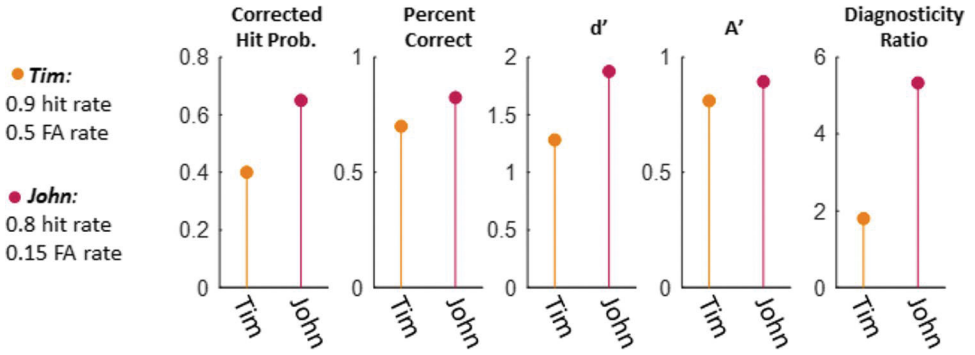


Fig. 2.

All common measures based solely on a single set of hits and false alarms (e.g., an old/new task) incorrectly believe John has a better memory than Tim

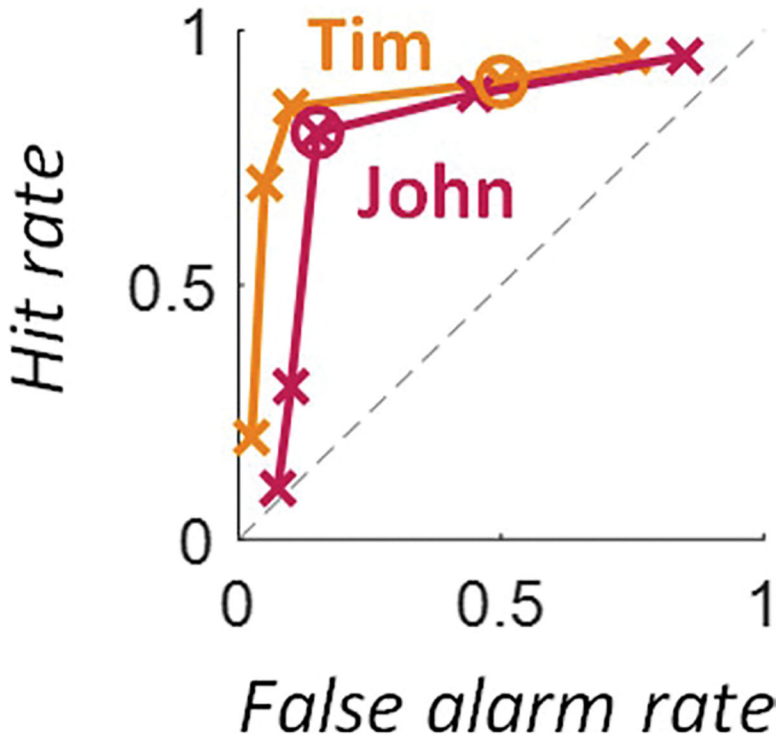


Fig. 3.

Receiver operating characteristic (ROC) curves for Tim and John. Rather than just a single hit or false alarm rate per person that we derive in an old/new paradigm (i.e., the circles, which show the binary old/new points we have been dealing with so far), if we collect a measure of latent memory strength, this allows us to plot an entire curve, capturing the entire distribution of memory strengths. From such a curve we can now directly read out who has the better memory (i.e., the higher hit rate at a given false alarm rate), at least given the interpolation of the curve (since the two participants never had the exact same false alarm rates). These data make clear that in fact Tim has a better memory than John, in direct contrast to all of the model-based counterfactual predictions (d', A', adjusted hit probability, etc.), since Tim's curve is reliably above John's

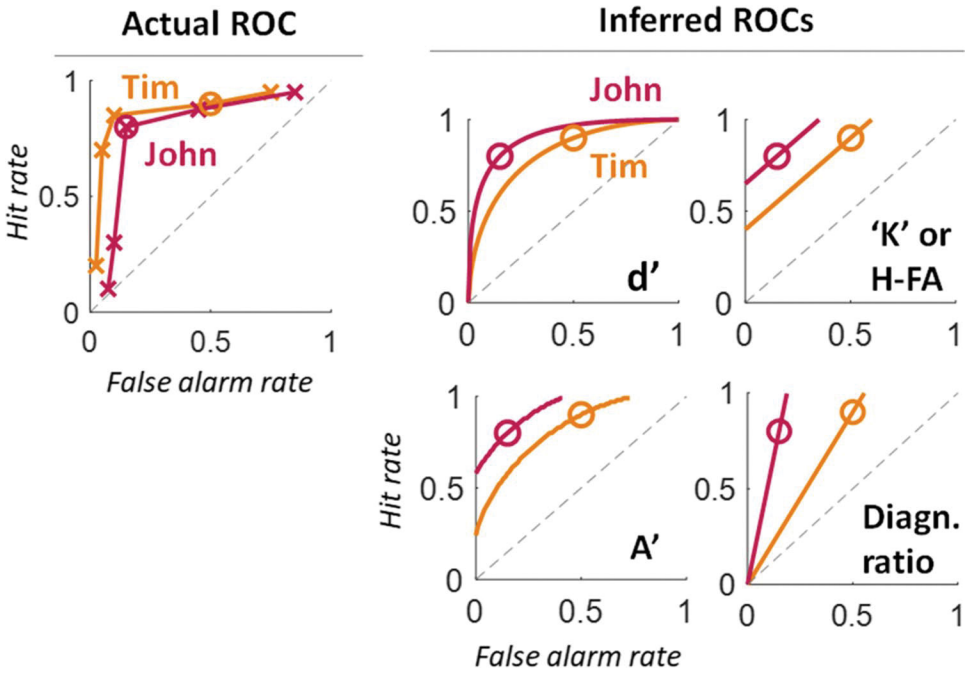


Fig. 4.

Left: The receiver operating characteristic (ROC) curve from the real latent memory strengths reveals that Tim has a superior memory to John. **Right**: The inferred ROCs from d'; *K*/hits minus false alarms; A'; and the diagnosticity ratio, as fit solely from the original, binary old/new data, all result in incorrectly higher ROCs for John than Tim (percent correct gives the same implied ROC as *K* or hits minus false alarms). An inferred ROC means that every pair of hit rates and false alarm rates along the d' curves would give the same d' for Tim and John as from their actual old/new response data; every pair of hit rates and false alarm rates on the A' curve would give the same A'; and likewise for the other measures

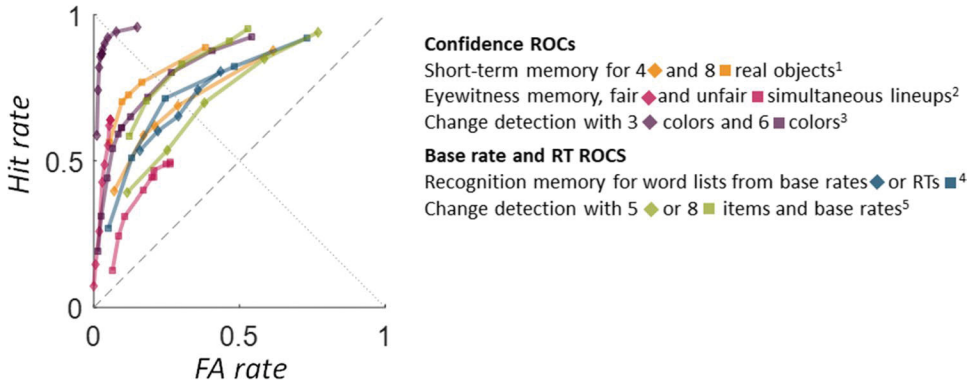


Fig. 5.

Example group-averaged receiver operating characteristic (ROC) curves from working memory and long-term memory tasks, using confidence, base rates and reaction times. The ROCs are generally curvilinear, though to varying degrees (e.g., green lines[5] are more linear), and generally asymmetric with respect to the dashed (y=1-x) line, though not always (e.g., change detection data[3] are quite symmetric). [1]Robinson et al. (2020a), Experiment 3 (both sessions); [2]Mickes et al. (2012), Experiment 1a and 2; [3]Unpublished data from Williams et al. (2022), available via Rahnev et al. (2020) database; [4]Juola et al. (2019); [5]Donkin et al. (2014), Experiment 2

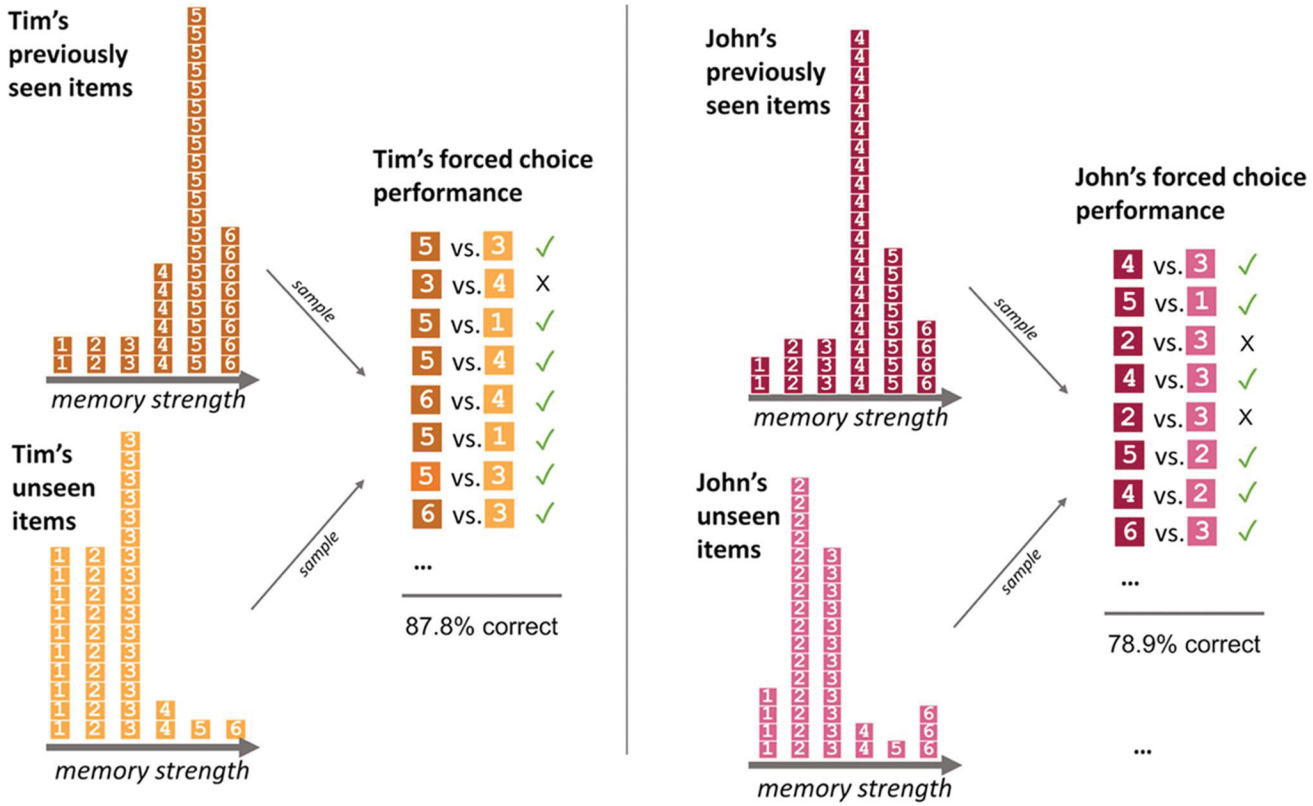


Fig. 6.

Forced-choice visualized. Each square represents 25 items from Table 1. Because on each trial (right side), a random item from the previously seen and previously unseen item distributions are paired, forced-choice performance necessarily depends on the entire distribution of both, unlike old/new performance. It thus provides a theory-neutral measure of the proportion of previously seen items that have stronger memories than the previously unseen items, providing an accurate index of memory

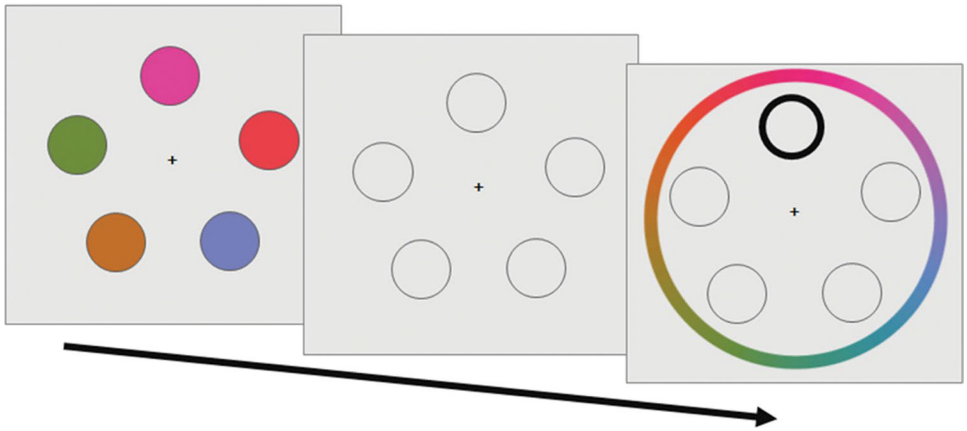


Fig. 7.

Sample continuous report tasks. In continuous report (i.e., the method of adjustment; continuous reproduction), participants must select which color (or shape, or orientation, …) was present in the probed location. This is effectively an m-AFC task in that participants are given many options, one old and many new, and asked to identify which is the old one, but is often analyzed in more complex ways (for a discussion, see Williams et al., 2022)

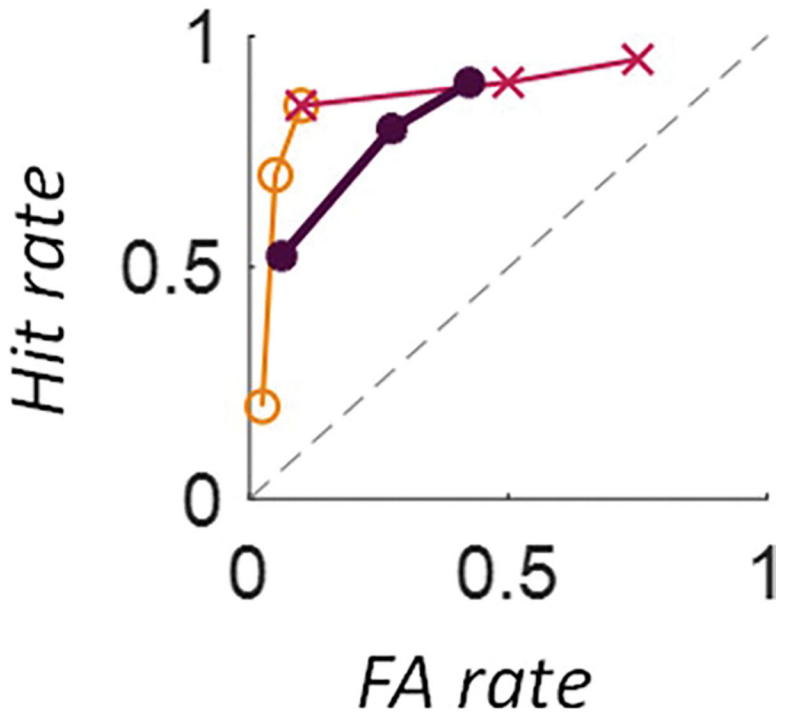


Fig. 8.

Tim-1 and Tim-2 have the same exact underlying distribution of memory strengths, and thus the exact same receiver operating characteristic curve (ROC) (i.e., orange and pink make up two halves of the same ROC). But they have very different response criteria: one being very conservative (in orange), and one very liberal (in pink). Unfortunately, averaging their ROC points – the average hit rate and false alarm rate across both subjects for confidence ≥1, ≥2, etc. – results in a lower ROC than either of their actual ROCs (the black curve). This is a general difficulty with ROCs, that the average is not necessarily representative of the performance of individuals, and in particular, conditions with more heterogeneous response criteria will appear to have lower ROCs in the average

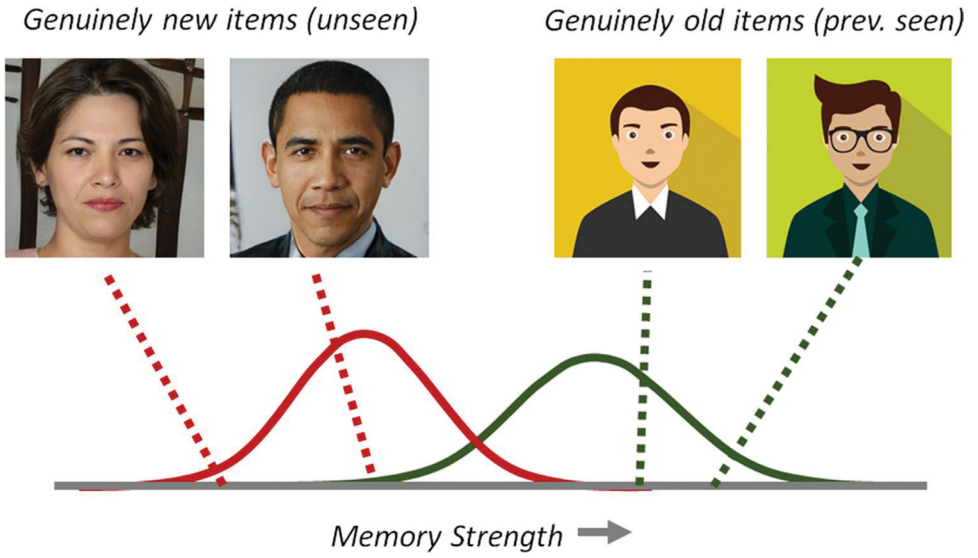


Fig. 9.

Signal detection simply instantiates the idea that memories vary continuously in strength: both genuinely seen items and genuinely unseen items can feel more or less familiar

Table 1

Summary of mainstream metrics in recognition memory research

Measure	Literature used in most often	Assumptions
K N(H - FA), where N is the number of items shown[1]	Visual working memory[2]	Memory is all-or-none, and we can therefore count how many items are remembered or not remembered from the set of items shown
Diagnosticity ratio H / FA	Eyewitness identification[3]	No memory theory underlies this measure. Instead, it rewards hits and punishes false alarms, and is valid only if every false alarm is somehow accompanied by N hits.
Corrected hit rate[4] H – FA	Long-term recognition memory[5]	Memory is all-or-none and we can therefore "correct for guessing' by subtracting false alarm rate (a measure of "guessing')
Percent correct mean(H & CR)	Broadly used[6]	Despite seeming atheoretical, this metric's validity also depends on memory being all-or-none
d′[7] Φ(H) - Φ(FA)	Broadly used[8]	Memory is continuous and distribution of memory signals for both old and new items follows equivalent normal distributions
A′[9]	Broadly used[10]	Rests on theoretical assumptions about how memory signals vary that are largely untenable when made explicit[11]

[1] From Cowan, 2001; see also Pashler, 1988

[2] For example, Alvarez & Cavanagh, 2004; Alvarez & Cavanagh, 2008; Brady & Alvarez, 2015; Chunharas et al., 2019; Endress & Potter, 2014; Eriksson et al., 2015; Fukuda & Vogel, 2019; Fukuda et al., 2010; Fukuda, Woodman, & Vogel, 2015; Fukuda et al., 2016a; Hakim et al., 2019; Irwin, 2014; Pailian et al., 2020; Schurgin & Brady, 2019; Shipstead et al., 2014; Sligte et al., 2008; Unsworth et al., 2014; Unsworth et al., 2015; Vogel & Machizawa, 2004; Woodman & Vogel, 2008

[3] For example, Clark & Wells, 2008; Wells & Lindsay, 1980; Wells et al., 1998

[4] Also known as the standard correction for guessing (Blackwell, 1953) or corrected hit probability.

[5] For example, Bower & Holyoak, 1973; Bowman & Dagmar, 2020; Cortese et al., 2015; Gardiner et al., 1996; Geiselman & Bjork, 1980; He et al., 2020; Jacoby et al., 2005; Johnson et al., 2004; Johnson et al., 2002; Potter et al., 2002; Scotti et al., 2020; Swick & Knight, 1999; Tulving & Thomson, 1971

[6] For example, Cappell et al., 2010; Fukuda et al., 2010a; Gao & Theeuwes, 2020; Harthsorne & Makovski, 2019; Luck & Vogel, 1997; Luria & Vogel, 2011; Maxcey-Richard & Hollingworth, 2013; Parra et al., 2014; Pessoa et al., 2002; Postle et al., 2003; Potter et al., 2002; Ricker et al., 2020; Shoval et al., 2020; Sloutsky & Fisher, 2004; Tas et al., 2016; Wagner et al., 2007; Yan et al., 2017

[7] Macmillan & Creelman, 2004

[8] For example, Benjamin, & S.,, & Bjork, R. A., 2000; Brady et al., 2019; Brady et al., 2017; Chubala et al., 2020; Diana et al., 2004; Greene, Bahri,, & Soto, 2010; Jiang et al., 2016; Johnson et al., 2009; Lamont et al., 2005; Lee & Cho, 2019; Monti et al., 2015; Rajaram & Pereira-Pasarin, 2007; Sahakyan et al., 2009; Schurgin & Brady, 2019; Toh et al., 2020

[9] Pollack & Norman, 1964

[10] For example, Aly & Turk-Browne, 2018; Fisher & Sloutsky, 2005; Hudon, Belleville, & Gauthier, 2009; Lind & Bowler, 2009; MacLin & MacLin, 2004; Poon & Fozard, 1980; Potter, Staub, Raud, & O'Connor, 2002; Reppa, Williams, Greville, & Saunders, 2020

[11] See: Macmillan & Creelman, 1996; Pastore, Crawley, Berens & Skelly, 2003; Wixted, 2020