Testing the Foundations of Signal Detection Theory in Recognition Memory

David Kellen

Syracuse University

Samuel Winiger

University of Zurich

John C. Dunn

University of Western Australia

Edith Cowan University

Henrik Singmann

University College London (UCL)

University of Warwick

Author Note

David Kellen, Samuel Winiger, and Henrik Singmann were supported by SNSF grant 100014_165591. John C. Dunn was supported by ARC grant DP130101535. The authors thank Michael J. Kahana, Rani Moran, Klaus Oberauer, John T. Wixted, Clintin P. Davis-Stober, and two anonymous reviewers for valuable comments. Portions of this work were presented at the 59th Annual Meeting of the Psychonomic Society, the 2019 Context and Episodic Memory Symposium (CEMS), and at the 52nd Annual Meeting of the Society for Mathematical Psychology.

Data, R scripts, and other supplemental materials can be found at the Open Science Framework: https://osf.io/zw9yr/

Correspondence: davekellen@gmail.com (David Kellen)

Abstract

Signal Detection Theory (SDT) plays a central role in the characterization of human judgments in a wide range of domains, most prominently in recognition memory. But despite its success, many of its fundamental properties are often misunderstood, especially when its comes to its testability. The present work examines five main properties that are characteristic of existing SDT models of recognition memory: i) random-scale representation, ii) latent-variable independence, iii) likelihood-ratio monotonicity, iv) ROC function asymmetry, and v) non-threshold representation. In each case, we establish testable consequences and test them against data collected in the appropriately-designed recognition memory experiment. We also discuss the connection between yes-no, forced-choice, and ranking judgments. This connection introduces additional behavioral constraints and yields an alternative method of reconstructing yes-no ROC functions. Overall, the reported results provide a strong empirical foundation for SDT modeling in recognition memory.

Keywords: signal detection theory, ROCs, recognition memory, area theorem, axiom testing

In the construction of scientific knowledge, it is useful to distinguish theoretical accounts at different levels. At one level, we consider a *theory* as usually understood, while at another level, we consider an instantiation of the theory as a specific *model*. Characterized in this way, a theory consists of a set of assumptions that builds a given *picture* of the domain of interest while a model contains additional assumptions in order to make contact with the world and to serve specific goals such as testing or parameter estimation (e.g., Bailer-Jones, 2009; Frigg & Hartmann, 2018; Kellen, 2019; Morgan & Morrison, 1999; van Fraassen, 1980). These additional auxiliary assumptions help connect the theory to data but are not part of the theory as such. For example, as we discuss in detail below, Signal Detection Theory pictures the evidence on which a decision is based as a latent variable (such as memory strength) having some distribution on a continuum. To turn this into a testable model, the auxiliary assumption that the distribution has a specified form (usually Gaussian) is added. Accordingly, a theory may be viewed as the intersection of the set or family of models that are consistent with it (e.g., Suppes 2002; van Fraassen 1980). The distinction between the core assumptions of a theory and the auxiliary assumptions of a specific model derived from said theory complicates theory testing: Any attempt to test a *theory* requires a demonstration that the conclusions being drawn do not hinge on the auxiliary assumptions of the *model* that actually comes into contact with the data. Rather, these conclusions should hold across all relevant models. Otherwise, one could always attribute the failure of a model to its auxiliary assumptions rather than to the theoretical picture at its core (Duhem, 1954).

Unfortunately, the distinction between theory and model is often overlooked. One example concerns the functional relationship between a latent variable and the dependent variable that measures it. While at a theoretical level this relationship may take any form, it is frequently assumed to be linear, as presupposed in ANOVA-type models (Dunn & Kalish, 2018; Garcia-Marques, Garcia-Marques, & Brauer, 2014; Kellen, Davis-Stober, Dunn, & Kalish, in press; Loftus, 1978; Wagenmakers, Krypotos, Criss, & Iverson, 2012). Based on this assumption, statistically-significant interactions

are frequently interpreted as supporting more complex theories (e.g., Ashby & Valentin, 2017) although such interactions may disappear under alternative models that do not assume linearity (for a recent overview, see Stephens, Matzke, & Hayes, 2019). In some domains such as syllogistic reasoning, the dismissal of linearity has led to drastic reinterpretations of long-standing empirical results (Rotello, Heit, & Dubé, 2015). In the context of response-time modeling, Jones and Dzhafarov (2014) showed that the ability of diffusion and ballistic accumulator models to successfully describe people's responses hinges on auxiliary distributional assumptions that are not an essential part of the underlying theory.

If a particular model fails to fit the data, it is often possible to attribute this to its auxiliary assumptions rather than to the core assumptions of the underlying theory. This possibility is concerning as it may serve to protect the theory from falsification. A similar issue arises when comparing models associated with competing theories. The fact that one model outperforms another does not mean that a similar result would be obtained under alternative models. For these reasons, it is desirable to identify *critical tests* of properties that can be found across large sets or families of models (e.g., Bamber, 1979; Birnbaum, 2008; Dunn & Anderson, 2018; Dunn & Kalish, 2018; Dunn & Rao, 2019; Falmagne, 1985; Karabatsos, 2005; Krantz, Luce, Suppes, & Tversky, 1971; Luce, 2010; Regenwetter, Dana, & Davis-Stober, 2011; Steingrimsson, 2016; Suppes, Krantz, Luce, & Tversky, 1989; Townsend & Nozawa, 1995). Such tests are examples of what Platt (1964) has famously referred to as *strong inference*.[1]

Relative to traditional model-fit comparisons (e.g., Pitt & Myung, 2002; Roberts & Pashler, 2000), critical tests offer two main advantages: First, they establish a transparent relation between theory and data – something that is not provided by penalized-fit statistics, regardless of their sophistication (Birnbaum, 2011a; Kellen, 2019). Second, the test results have farther reaching implications. For instance, the critical-test results reported by Birnbaum (2008) do not merely reject some parametric

[1] It is worth noting that the same concerns and desire for greater generality are found in some of the model-comparison approaches that have been recently proposed (e.g., *multiverse analyses*, see Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; *factorial model comparisons*, see van den Berg, Awh, & Ma, 2014).

implementation of Prospect Theory – they reject the *entire family* of Prospect-Theory models formalized by Tversky and Kahneman (1992). Similarly, Regenwetter et al.'s (2011) failure to reject the assumption of transitivity in people's preferences demonstrates the viability of a large family of models without having to enumerate the unique aspects of each of its members.

The aim of the present paper is to apply a critical-test approach to *families of Signal Detection Theory* (SDT) *models*, focusing on their application to the domain of recognition memory. SDT is arguably one of the most successful theoretical frameworks in psychology today (for overviews, see Green & Swets, 1966; Kellen & Klauer, 2018; Macmillan & Creelman, 2005; Wickens, 2002) and has played a central role in the modeling of recognition-memory judgments (see Egan, 1958; Wixted, 2007; Rotello, 2018; Yonelinas & Parks, 2007). The popularity of SDT can be attributed to two factors: its *empirical success* and its *theoretical inclusiveness*. First, SDT has been shown to characterize judgments across a wide variety of psychological domains beyond memory, such as perception and reasoning (e.g., Green & Swets, 1966; Macmillan & Creelman, 2005; Rotello, 2018; Trippas et al., 2018). Second, its core assumption that judgments are based on an evaluation of latent-strength values sampled from continuous distributions plays well with popular theoretical accounts of learning, forgetting, and generalization, among others (Lockhart & Murdock, 1970). Nevertheless, SDT has been traditionally applied in ways that rely heavily on auxiliary assumptions, which are often simply taken for granted by researchers. This reliance is problematic as it tends to promote misconceptions regarding the distinction between SDT and alternative models. For example, *Threshold Models* are often contrasted with *Gaussian SDT models*, under the premise that they provide theoretical accounts that are fundamentally distinct (e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012). But a closer look shows that both models can be cast as members of large families of SDT models (see Falmagne, 1985, Chap. 10; Kellen & Klauer, 2018; Macmillan & Creelman, 2005; Malejka & Bröder, 2019; Rouder, Province, Swagman, & Thiele, 2014).

Figure 1 illustrates our application of critical tests to a hierarchy of properties,

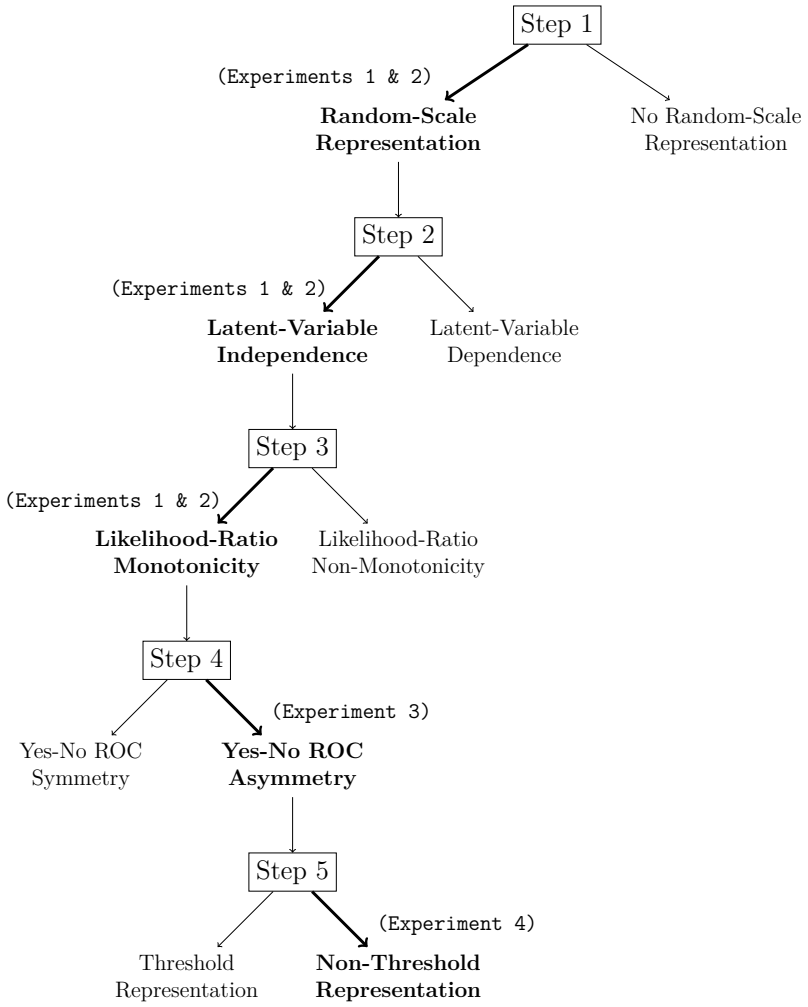

Figure 1. Overview of the critical testing conducted in the present manuscript. Each step establishes a division in terms of whether or not a property is satisfied. The properties supported by the reported experiments are presented in bold. For details, see the main text.

which are organized in a way that reflects their increasing level of specificity (for very similar testing strategies, see Luce, 2010; Steingrimsson, 2016). Families and sub-families of SDT models can be defined by establishing which properties its members must satisfy. When testing for a specific property, we are evaluating the empirical adequacy of the families or sub-families of models that satisfy it. Allowing for details to be discussed later on, let us take brief overview of the properties being tested. At the top level, we identify a critical property of SDT models at large – the assumption of a *random-scale representation*. In the context of recognition memory, this representation holds that choices between studied items and new items is characterized by a joint distribution of latent-strength values and the probability distribution over rank orders that it induces. Experiments 1 and 2 implement a direct test on the family of SDT

models that assume a random-scale representation. A failure of the random-scale representation hypothesis would indicate a need to rethink the way in which latent-strength variables are assumed to underlie recognition-memory judgments.

At the second level, we identify a property that is almost universally assumed in existing SDT models: that the latent-strength variables associated with the different items in a given test trial (e.g., in an m-alternative forced-choice task) are *independent*. At the third level, we consider a property that is expected to hold for the latent-strength variables – *likelihood-ratio monotonicity*. Experiments 1 and 2 will be used to test the empirical adequacy of these two properties. At the fourth level, we focus on the *symmetry of ROC functions*, a property that has been at the center of many theoretical discussions (for a review, see Yonelinas & Parks, 2007). Experiment 3 implements a critical test for ROC symmetry that does not rely on parametric assumptions. Finally, at the fifth level, we distinguish between *threshold and non-threshold representations*. Experiment 4 implements a critical test that distinguishes between these two representations. To foreshadow our results, at each node in the family tree structure shown in Figure 1, we find evidence supporting the branch labelled in **bold** text. Altogether, the results obtained across these studies provide an empirical foundation for SDT that has previously been unavailable, opening new avenues for future research.

Signal Detection Theory and ROC Functions

Let us consider a general scenario in which a decision maker is tasked with classifying incoming stimuli as belonging to one of a number of pre-specified stimulus classes. For example, such a scenario is encountered in a visual *yes-no* task in which the decision maker classifies stimuli as *signal* or *noise* by responding "`yes`" or "`no`" respectively. A second example would be a recognition-memory yes-no task in which the decision maker attempts to identify which stimuli she previously encountered in a study phase, and which ones she did not, once again by answering "`yes`" or "`no`". The challenge faced by the decision maker is that these classifications always involve some degree of uncertainty or confusability (for discussions, see Lu & Dosher, 2008; Wixted,

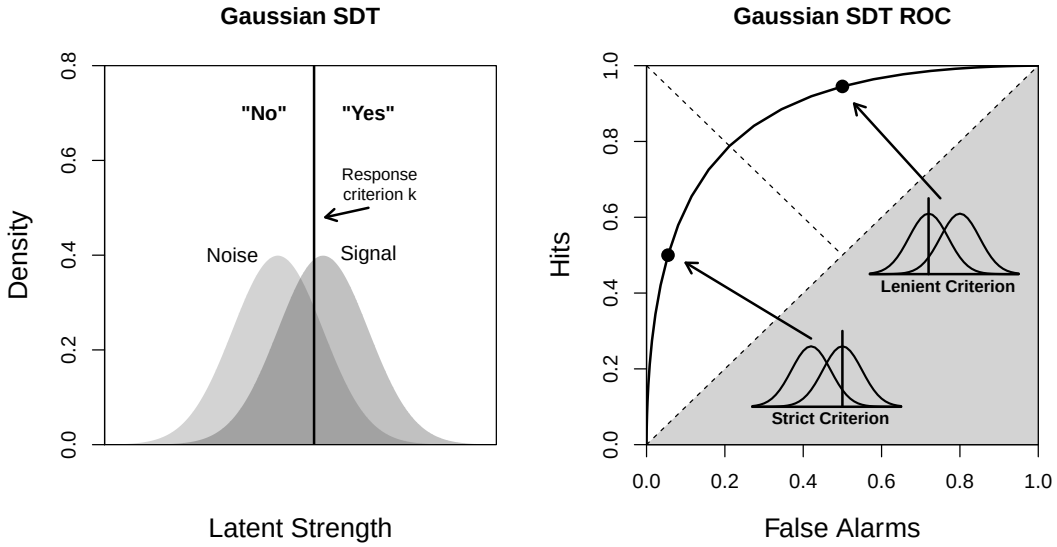

Figure 2. *Left Panel*: Illustration of the Gaussian SDT model Latent-strength values above the response criterion result in a "yes" response, values below a "no" response. *Right Panel*: The models' respective yes-no ROC, along with two ROC points associated with different response criteria (one strict, the other lenient). Dashed lines delimiting the gray areas indicate chance-level performance.

2020). According to SDT, the decision maker classifies stimuli based on the latent-strength values λ associated with each of them. These values are assumed to be realizations of random variables established on a latent metric space. In a simple scenario in which the decision maker attempts to discriminate between two stimulus classes – *signal* (S) and *noise* (N) – the distributions of these random variables are established on a latent unidimensional scale which we refer to as *latent strength*. In the context of recognition memory, signal and noise items correspond to studied and new items, respectively (these terms will be used interchangeably throughout this paper).

In a yes-no task, the latent-strength λ value associated with a given stimulus is evaluated by comparing it with a pre-established *response criterion* κ.[2] A "`signal`" or "`yes`" response is produced when $\lambda \geq \kappa$, otherwise a "`noise`" or "`no`" response is given.

[2] Characterizations of SDT typically assume that the response criterion κ is fixed across test trials. Although we can relax this restriction and allow κ to be a random variable (e.g., Kellen, Klauer, & Singmann, 2012; Rosner & Kochanski, 2009), we refrain from doing so here for two reasons: First, the present discussion of κ takes place in a context in which SDT with fixed vs. random κ are not empirically distinguishable (see Rosner & Kochanski, 2009). Second, our discussions will almost invariably focus on forced-choice judgments, which are assumed to be based on comparisons between latent-strength values (i.e., κ plays no role). The only exception is when we establish a critical test comparing 'threshold' versus 'non-threshold' representations (see Step 5 in Figure 1). In this specific case, it turns out that the question of whether or not κ is fixed across trials is completely inconsequential.

The degree of overlap between latent distributions reflects the degree of confusability between the different stimulus classes. The value of criterion κ relative to these latent distributions denotes the degree of response bias, with larger values indicating a more strict response criterion. The left panel of Figure 2 illustrates this description of SDT, under the assumption that the latent signal and noise distributions are Gaussian. The right panel of Figure 2 illustrates the predicted relationship between the probabilities of "`yes`" responses for signal and noise stimuli – referred to as *hit* (H) and *false-alarm* probabilities (FA) – as the response criterion κ varies from strict to lenient. The functional relationship between hit and false-alarm probabilities, $\mathrm{H} = \rho(\mathrm{FA})$, which can be used to characterize the decision-maker's performance across different response biases, is commonly referred to as the *[yes-no] ROC function*.

When discussing some of the more formal aspects of SDT, such as establishing the kind of ROC functions that it can predict, or the relationships between different types of judgments, it is convenient to adopt a *universal representation* in which latent variables are defined on a [0,1] interval (for previous applications, see Iverson & Bamber, 1997; Rouder, Pratte, & Morey, 2010; Rouder et al. 2014). Let f_N and f_S be the noise and signal density functions, respectively, with corresponding cumulative distribution functions F_N and F_S. Let F^{-1} denote the inverse of the cumulative distribution function. We will also assume that the noise distribution is *uniform* on the $[0, 1]$ interval, such that $f_N(\kappa) = 1$ and $F_N(\kappa) = \kappa$ for all $\kappa \in [0, 1]$. No constraints are being imposed on f_S, with the exception that $\int_0^1 f_S(t)\ \mathrm{d}t = 1$. Unless noted, this universal representation will be assumed throughout the manuscript. We acknowledge that this representation is not 'standard' in the sense that latent variables are typically represented on the entire real line (e.g., see the left panel of Figure 2) rather than on the unit interval. However, its adoption does not limit the scope of any the theoretical results discussed (as shown below, we can describe any yes-no ROC function with it).

According to SDT, the hit and false-alarm probabilities correspond to:

$$\mathrm{FA} \quad = \quad P(\lambda_N \geq \kappa) = 1 - F_N(\kappa), \tag{1}$$

$$\text{H} \quad = \quad P(\lambda_S \geq \kappa) = 1 - F_S(\kappa). \tag{2}$$

As previously stated, the ROC function ρ expresses the hit probability H as a function of the false-alarm probability FA. Rearranging (1), we see that $\kappa = F_N^{-1}(1 - \text{FA})$, which leads us to a more explicit formulation of the yes-no ROC function:

$$\text{H} = \rho(\text{FA}) = 1 - F_S(F_N^{-1}(1 - \text{FA})). \tag{3}$$

But what kind of ROC functions are permissible under SDT? The shape of an ROC function is given by its slope, $\rho'(\text{FA})$, at each point. Taking the derivative of (3) with respect to FA:

$$\rho'(\text{FA}) = \frac{f_S(F_N^{-1}(1 - \text{FA}))}{f_N(F_N^{-1}(1 - \text{FA}))} = \frac{f_S(\kappa)}{f_N(\kappa)}. \tag{4}$$

Given that $f_N(\kappa) = 1$ under a universal representation, it follows that:

$$\rho'(\text{FA}) = f_S(\kappa). \tag{5}$$

It turns out that the slope of the ROC function at any given point corresponds to the signal density f_S. Note that because $f_S \geq 0$, the ROC function is necessarily monotonically increasing. It also follows that *any* ROC shape (e.g., linear, concave) can be produced by establishing a density f_S that perfectly matches the slope of the function. In other words, there is no (monotonically-increasing) yes-no ROC function that is at odds with the tenets of SDT.

Additional constraints can be introduced via the assumption that the likelihood ratio $\frac{f_S(\kappa)}{f_N(\kappa)}$ is monotonically increasing, reflecting the notion that signal items are more likely to take on larger latent-strength values than noise items (Criss & McClelland, 2006; Glanzer, Hilford, & Maloney, 2009; Osth & Dennis, 2015; Zhang & Mueller, 2005). This property, *likelihood-ratio monotonicity*, holds if and only if the ROC function is concave (i.e., it has a monotonically non-increasing slope).

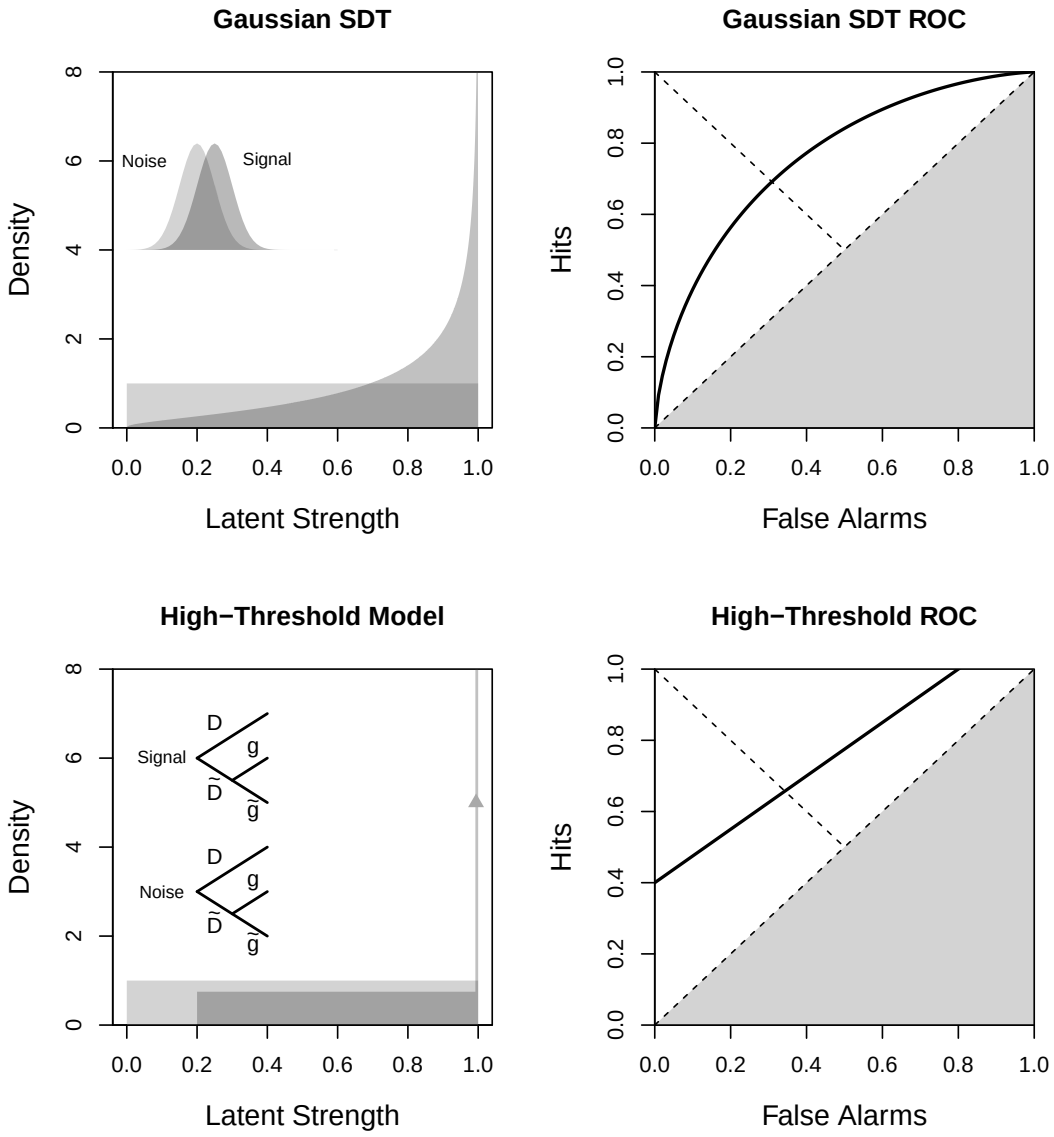

Figure 3. Examples of models that can be cast in a *universal SDT representation*. *Left Column*: Universal representation of models, along with their traditional representations (embedded in each panel). The darker density corresponds to the signal distribution, the lighter density the noise distribution. The parameters (e.g., D, g) and their complements ($\tilde{D}$, $\tilde{g}$) in each tree correspond to the probabilities associated with the different binary branches. For the high-threshold distribution, the signal density is a mixture between a rectangular distribution and a Dirac pulse (highlighted by a triangle). *Right Column*: The models' respective ROCs. Dashed lines delimiting the gray areas indicate chance-level performance.

ROC Shape Only Permits the Testing of SDT Family Members

The fact that SDT can accommodate any monotonically-increasing ROC function raises important issues in the way SDT modeling is traditionally approached in the domain of recognition memory. Researchers often rely on ROC shape to compare different models, such as the un/equal variance Gaussian SDT model (Egan, 1958), the

Finite-Mixture Model (DeCarlo, 2002), the Dual-Process Model (Yonelinas, 1997), or the High-Threshold Model (Bröder & Schütz, 2009). Rotello (2018) provides a thorough review. The problem with these comparisons is that they often confound the SDT assumption of 'latent-strength-based judgments' with the Gaussian assumption and the ROCs that it can yield. Manifestations of this confound include evaluations of the number and nature of different memory processes that are entirely based on *how curved* an ROC function is (e.g., Parks, Murray, Elfman, & Yonelinas, 2011), or attempts to dismiss a dual-process account through the investigation of the fit residuals associated with different candidate models (e.g., Dede, Squire, & Wixted, 2014; Kellen & Singmann, 2016). These comparisons overlook – or at the very least downplay – the fact that there is always some latent-strength distribution f_S that can perfectly capture the observed ROC shape (for a similar point, see Rouder et al., 2014).[3]

One prominent example is the long-documented comparison between the unequal-variance Gaussian SDT model and the *High-Threshold Model* (e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012; Macmillan & Creelman, 2005). This specific comparison has been framed as distinguishing between *continuous* and *discrete* accounts of recognition memory.[4] Figure 3 shows the universal SDT representation of both models and their corresponding ROCs. In the case of the High-Threshold Model, it is clear that it can be equivalently represented in terms of continuous distributions.[5] This representability indicates that the success of the High-Threshold Model, especially when relative to an alternative such as the Gaussian SDT model (e.g., Kellen et al., 2013),

---

[3] One reaction to this point is that the discussions mentioned above reflect researchers' interest in specific models, *auxiliary assumptions included*. Although we acknowledge the legitimacy of such modeling practices, we argue that the meta-theoretical principles that underlie it severely limit what can be meaningfully said and done. We will turn our attention to this issue in the General Discussion.

[4] We will discuss threshold models in detail later on, when testing a 'generalized' threshold model (see Step 5 in Figure 1).

[5] Macmillan and Creelman (2005, Chap. 4) and Swets (1986) showed that the High-Threshold Model can be represented as a SDT model with rectangular distributions. Our universal representation is slightly different because both distributions are bound to the [0,1] interval, which is entirely covered by a uniform noise distribution. Unlike Macmillan and Creelman or Swets, our formulation does not allow us to establish an upper region of strength values that *only* the signal distribution covers. We therefore have to represent the signal distribution as a mixture between a uniform distribution and Dirac pulse located at the upper boundary 1. Importantly, note that these differences do not indicate a limitation of either representation in their ability to characterize the high-threshold model.

cannot be used to argue against SDT. The only thing these model comparisons can achieve is determining which parametric assumptions perform best in a given situation (see also Malejka & Bröder, 2019).

Step 1: Random-Scale Representation

A fundamental feature shared by all SDT models considered in the recognition-memory literature at large is the existence of a *random-scale representation*. That is, the different classes of stimuli (e.g., studied and new items) can be characterized by a joint latent distribution, and that choice probabilities can be induced from the comparison of a joint realization of latent-strength values. In the domain of forced-choice judgments, it can be shown that this random-scale representation implies a set of order constraints originally formalized by Block and Marschak (1960) and Falmagne (1978). These results, known as the *Block-Marschak inequalities*, were originally discussed in the context of the class of *'random-utility'* or *'random-scale'* models, which includes SDT, Thurstonian models (Thurstone, 1927; Torgerson, 1958), Luce's Choice Theory (Luce, 1959), among others (for reviews, see Marley, 1990; Marley & Regenwetter, 2017).

Given its origins, it is useful to begin by discussing the Block-Marschak inequalities in the broader context of 'multi-alternative decision making', and only then discuss the specific case of forced-choice judgments in recognition memory. Let a, b, c, d, ..., denote alternatives or options in an option set T, and let $A \subseteq T$ be a non-empty option subset. Moreover, let $(\lambda_a, \lambda_b, \lambda_c, \lambda_d, ...)$, be a vector drawn from a latent joint distribution. Assuming that ties are impossible, it is easy to see that this joint distribution of latent strengths implies a probability distribution over rank orders (e.g., $P(\lambda_b > ... > \lambda_d > ... > \lambda_a, > \lambda_c > ...)$). Now, let us consider a decision maker that, when presented with the options in subset A, chooses the one with the *largest* latent-strength value λ. It follows that the probability of choosing $x \in A$, $P_x^{\langle A \rangle}$, corresponds to

$$P_x^{\langle A \rangle} = P(\lambda_x = \max_{z \in A}(\lambda_z)), \quad x \in A \subseteq T, \tag{6}$$

which can be induced from the probability distribution over rankings. For example, if $A = \{a, b, c, d\}$, then

$$\begin{aligned} P_a^{\langle A \rangle} &= P(\lambda_a > \lambda_b > \lambda_c > \lambda_d) + P(\lambda_a > \lambda_b > \lambda_d > \lambda_c) + \\ &\quad P(\lambda_a > \lambda_c > \lambda_b > \lambda_d) + P(\lambda_a > \lambda_c > \lambda_d > \lambda_b) + \\ &\quad P(\lambda_a > \lambda_d > \lambda_b > \lambda_c) + P(\lambda_a > \lambda_d > \lambda_c > \lambda_b) \end{aligned} \tag{7}$$

A random-scale representation is said to hold when choice probabilities over option subsets A are based on a joint distribution of latent variables and (6). This representation introduces a number of inequality constraints at the level of choice probabilities when comparing different option subsets. For example, consider the probability of choosing option a when presented with subset $A \setminus \{b\}$, in which option b is not included:

$$\begin{aligned} P_a^{\langle A \setminus \{b\} \rangle} &= P(\lambda_a > \lambda_c > \lambda_d) + P(\lambda_a > \lambda_d > \lambda_c) \\ &= P(\lambda_b > \lambda_a > \lambda_c > \lambda_d) + P(\lambda_a > \lambda_b > \lambda_c > \lambda_d) + \\ &\quad P(\lambda_a > \lambda_c > \lambda_b > \lambda_d) + P(\lambda_a > \lambda_c > \lambda_d > \lambda_b) + \\ &\quad P(\lambda_b > \lambda_a > \lambda_d > \lambda_c) + P(\lambda_a > \lambda_b > \lambda_d > \lambda_c) + \\ &\quad P(\lambda_a > \lambda_d > \lambda_b > \lambda_c) + P(\lambda_a > \lambda_d > \lambda_c > \lambda_b) \end{aligned} \tag{8}$$

The latter choice probability corresponds to the sum of probabilities of rank orders in which option a is ranked above c and d, while option b can be assigned any rank. Because (8) includes all the probability summands in (7) and then some, it follows that $P_a^{\langle A \setminus \{b\} \rangle} \geq P_a^{\langle A \rangle}$. Block and Marschak (1960) showed that the example above is a small part of a system of inequalities that is implied when choices can be induced from a probability distribution over rank orders. For example, if $A = \{a, b, c, d\}$:

$$\begin{aligned} P_a^{\langle A \rangle} &\geq 0, \\ P_a^{\langle A \setminus \{b\} \rangle} &\geq 0, \end{aligned}$$

$$\begin{aligned} P_a^{\langle A\setminus\{b\}\rangle} - P_a^{\langle A\rangle} &\geq 0, \\ P_a^{\langle A\rangle} + P_a^{\langle A\setminus\{b,c\}\rangle} - P_a^{\langle A\setminus\{b\}\rangle} - P_a^{\langle A\setminus\{c\}\rangle} &\geq 0, \\ P_a^{\langle A\setminus\{b,c,d\}\rangle} - P_a^{\langle A\setminus\{b,c\}\rangle} - P_a^{\langle A\setminus\{b,d\}\rangle} - P_a^{\langle A\setminus\{c,d\}\rangle} - P_a^{\langle A\rangle} & \\ P_a^{\langle A\setminus\{b\}\rangle} + P_a^{\langle A\setminus\{c\}\rangle} + P_a^{\langle A\setminus\{d\}\rangle} &\geq 0, \end{aligned} \tag{9}$$

with analogous inequalities being enforced over the probabilities of choosing options b, c, or d. This system of inequalities can be described in a compact form:

$$\sum_{B:A\subseteq B\subseteq T} (-1)^{|B\setminus A|} P_x^{\langle B\rangle} \geq 0, \tag{10}$$

for all non-empty $A \subseteq T$ and $x \in A$, and with $|B \setminus A|$ denoting the cardinality of option subset B *minus* subset A. As later proved by Falmagne (1978), choice probabilities are consistent with a random-scale representation *if and only if* they satisfy the system of *Block-Marschak inequalities* (Block & Marschak, 1960). In other words, the statements "*a decision-maker is consistent with some random-scale representation*" and "*the choice probabilities of a decision-maker satisfy the Block-Marschak inequalities*" are formally equivalent.

Two important aspects of Falmagne's (1978) result need to be highlighted: First, its generality – aside from the impossibility of ties, no parametric assumptions are being imposed over the latent distributions. Second, the system of inequalities defined by (10) provides a complete description of the facets of a convex polytope on a unit hypercube, with vertices representing the possible ranking orders on 0/1 coordinates (for a minimal description of this polytope, see Fiorini, 2004).[6] Choice probabilities conforming to this system of inequalities correspond to mixtures of these vertices, and therefore will be inside the convex polytope. The same holds for averages of said choice probabilities, which means that the inequalities cannot be spuriously violated because of aggregation (e.g., across participants; see Estes, 1956; Heathcote, Brown, & Mewhort, 2000; for a

[6] It turns out that many hypotheses in psychology can be appropriately represented by convex polytopes, which has important implications for their testing (for detailed tutorial discussions, see Heck & Davis-Stober, 2019; Marley & Regenwetter, 2017; Regenwetter & Cavagnaro, 2019).

recent overview, see Regenwetter & Robinson, 2017).[7]

From an empirical perspective, the Block-Marschak inequalities are far from trivial: As shown by McCausland, Davis-Stober, Marley, Park, and Brown (2020), the space of choice probabilities that satisfy them is small under realistic choice experiments, allowing researchers to apply strict tests to the hypothesis that a random-scale representation holds (see also McCausland & Marley, 2014). Nor are they theoretically trivial: For instance, one of the properties that emerges from the random-scale representation is *regularity*, according to which the probability of choosing a given option (e.g., option a) cannot be increased by introducing additional options into the subset under consideration. The inequality $P_a^{\langle A\setminus\{b\}\rangle} \geq P_a^{\langle A\rangle}$ discussed earlier exemplifies this property. Violations of regularity have been observed in a variety of domains, as reported in the rapidly-expanding literature on 'context effects' (e.g., Spektor, Kellen, & Hotaling, 2018; Trueblood, Brown, Heathcote, & Busemeyer, 2013). For example, according to the *attraction effect*, the probability of choosing one option (e.g., apples) over another (e.g., oranges) *increases* when introducing an inferior option (e.g., apples of worse quality but not cheaper).

Block-Marschak Inequalities in The Context of m-Alternative Recognition Judgments

In the context of recognition memory, we are dealing with scenarios that are much simpler than the ones found in typical multi-alternative decision making scenarios. The option sets used in m-alternative forced-choice (m-AFC) memory tasks are comprised of alternatives coming from only two stimulus classes – *signal* and *noise*. More specifically, the option set presented at a given test trial is always comprised of *one* signal alternative and $m-1$ noise alternatives. Alternatives coming from the same stimulus class are all associated with the same (marginal) latent-strength distribution.

---

[7] Although the Block-Marschak inequalities cannot be spuriously violated, they could be spuriously satisfied. This possibility is reflected in the fact that we can always concoct scenarios in which individual choice probabilities outside of the polytope fall within it (or very close to it) when aggregated. However, the small volumes of the polytopes in question (we will compute these volumes later on) suggest that the natural occurrence of such scenarios is quite implausible.

The latent variables associated with the $m-1$ noise stimuli are not only assumed to be identically distributed – they are assumed to be *exchangeable* (for an introduction to exchangeability, see Bernardo, 1996):[8] Let λ_S be the latent-strength value of the signal option and let $\lambda_{N,i}$ be the latent-strength values of the i^{th} noise option. Under exchangeability, the probability distribution over rank orders is the same across permutations of the noise options:

$$P(\lambda_{N,1} > ... > \lambda_{N,m-1}) = P(\lambda_{N,\xi(1)} > ... > \lambda_{N,\xi(m-1)}), \tag{11}$$

where ξ denotes some permutation of the indices $i = 1, ..., m-1$. The rationale behind the assumption of exchangeability is straightforward: In an m-AFC recognition task, the indexing of noise options, which are all assumed to belong to the same stimulus class, is completely arbitrary. In fact, when removing a noise option from a option set; i.e., going from m to $m-1$ alternatives total, there is no substantive distinction between the removal of the ith noise option vis-à-vis the jth noise option.[9] Because noise options are exchangeable, the characterization of people's choices can be reduced to how often they choose the signal stimulus when the latter is presented alongside $m-1$ noise stimuli.

Another characteristic of m-AFC judgments in recognition memory is that the probability of choosing the signal option should not go below $\frac{1}{m}$; i.e., *chance performance*. This lower boundary corresponds to an individual that is completely unable to differentiate between studied and new items (i.e., we are assuming that individuals are not actively trying to make incorrect judgments).

Altogether, the use of only two stimulus classes (signal and noise) and the exchangeability of noise options results in a simplification of the Block-Marschak inequalities. To better observe this simplification, let $A = \{a, b, c, d\}$, and let a denote

[8] Note that *independent and identically distributed* (i.i.d.) variables are exchangeable, but the converse is not true. For example, consider a bivariate Gaussian distribution with mean vector $\boldsymbol{\mu} = [0, 0]$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \varrho \\ \varrho & 1 \end{bmatrix}$. The two variables established by this distribution are not independent, but they are exchangeable.

[9] To be clear, we are not denying the possibility of setting up richer experimental designs in which we can distinguish between *different classes* of noise stimuli (e.g., high- and low-frequency words). We are only assuming exchangeability among options belonging to the same stimulus class.

the signal option whereas b, c, and d denote the exchangeable noise options. Now, let $P_S^{\langle m\rangle}$ denote the probability of selecting the signal option (a) in an option subset of size m. This probability, which has a lower bound of $\frac{1}{m}$, is the same across all possible option subsets of size m:

$$
\begin{aligned}
P_S^{\langle 4\rangle} &= P_a^{\langle A\rangle} \\
P_S^{\langle 3\rangle} &= P_a^{\langle A\setminus\{b\}\rangle} = P_a^{\langle A\setminus\{c\}\rangle} = P_a^{\langle A\setminus\{d\}\rangle} \\
P_S^{\langle 3\rangle} &= P_a^{\langle A\setminus\{b,c\}\rangle} = P_a^{\langle A\setminus\{b,d\}\rangle} = P_a^{\langle A\setminus\{c,d\}\rangle} \\
P_S^{\langle 1\rangle} &= P_a^{\langle A\setminus\{b,c,d\}\rangle} = 1
\end{aligned}
$$

Going back to the inequalities given in (9):

$$
\begin{array}{rcl}
\textbf{Multi-Alternative Decision Making} & \Longrightarrow & \textbf{m-AFC Recognition} \\
P_a^{\langle A\rangle} \geq 0 & \Longrightarrow & P_S^{\langle 4\rangle} \geq \frac{1}{4} \\
P_a^{\langle A\setminus\{b\}\rangle} \geq 0 & \Longrightarrow & P_S^{\langle 3\rangle} \geq \frac{1}{3} \\
P_a^{\langle A\setminus\{b\}\rangle} - P_a^{\langle A\rangle} \geq 0 & \Longrightarrow & P_S^{\langle 3\rangle} - P_S^{\langle 4\rangle} \geq 0 \\
P_a^{\langle A\rangle} + P_a^{\langle A\setminus\{b,c\}\rangle} - P_a^{\langle A\setminus\{b\}\rangle} - P_a^{\langle A\setminus\{c\}\rangle} \geq 0 & \Longrightarrow & P_S^{\langle 4\rangle} + P_S^{\langle 2\rangle} - 2P_S^{\langle 3\rangle} \geq 0 \\
P_a^{\langle A\setminus\{b,c,d\}\rangle} - P_a^{\langle A\setminus\{b,c\}\rangle} - P_a^{\langle A\setminus\{b,d\}\rangle} - P_a^{\langle A\setminus\{c,d\}\rangle} & & \\
+P_a^{\langle A\setminus\{b\}\rangle} + P_a^{\langle A\setminus\{c\}\rangle} + P_a^{\langle A\setminus\{d\}\rangle} - P_a^{\langle A\rangle} \geq 0 & \Longrightarrow & P_S^{\langle 1\rangle} - 3P_S^{\langle 2\rangle} + 3P_S^{\langle 3\rangle} - P_S^{\langle 4\rangle} \geq 0
\end{array}
$$

Now that we clarified the transition to the specific context of m-AFC recognition, we can provide a more general statement about the system of Block-Marschak inequalities that applies to it. In a sequence of m-AFC trials, with $m \in \{1, 2, ..., M\}$:

$$
\begin{aligned}
P_S^{\langle m\rangle} &\geq \tfrac{1}{m}, \quad \text{for } 1 \leq m < M, \\
P_S^{\langle m\rangle} - P_S^{\langle m+1\rangle} &\geq 0, \quad \text{for } 1 \leq m < M, \\
P_S^{\langle m-1\rangle} - 2P_S^{\langle m\rangle} + P_S^{\langle m+1\rangle} &\geq 0, \quad \text{for } 2 \leq m < M, \\
P_S^{\langle m-2\rangle} - 3P_S^{\langle m-1\rangle} + 3P_S^{\langle m\rangle} - P_S^{\langle m+1\rangle} &\geq 0, \quad \text{for } 3 \leq m < M, \\
P_S^{\langle m-3\rangle} - 4P_S^{\langle m-2\rangle} + 6P_S^{\langle m-1\rangle} - 4P_S^{\langle m\rangle} + P_S^{\langle m+1\rangle} &\geq 0, \quad \text{for } 4 \leq m < M, \\
P_S^{\langle m-4\rangle} - 5P_S^{\langle m-3\rangle} + 10P_S^{\langle m-2\rangle} - 10P_S^{\langle m-1\rangle} + 5P_S^{\langle m\rangle} - P_S^{\langle m+1\rangle} &\geq 0, \quad \text{for } 5 \leq m < M, \\
&\textit{etc.}
\end{aligned}
\tag{12}
$$

The system of linear inequalities described in (12) also constitutes a convex polytope (for a proof, see Grünbaum, 2003, Chap. 3), established on a unit hypercube with $M-1$ non-redundant dimensions, each referring to a given $P_S^{\langle m \rangle}$. The polytope's vertex representation is included in the online Supplemental Materials hosted at the Open Science Framework (see Author Note).

The transition to the case of recognition memory also has a positive effect on the plausibility of the statistical assumptions that are going to be made. Both choice-probability estimation and hypothesis testing (which will be detailed later on) will be conducted under the premise that each discrete choice is an *independent and identically distributed* (*i.i.d.*) sample. For example, the choice probability $P_a^{\langle A \rangle}$ would be estimated from the data by considering the number of times option a is chosen out of *multiple presentations* of option set A (e.g., Regenwetter et al., 2011). One concern is that choices might not be i.i.d. when each individual participant is presented with the exact same option sets multiple times (for a discussion, see Regenwetter, Dana, Davis-Stober, & Guo, 2011; Birnbaum, 2011b). This concern is considerably less serious when estimating $P_S^{\langle m \rangle}$ from recognition judgments, given that the multiple m-AFC trials encountered by any given participant are each comprised of a unique set of items. In other words, no m-AFC trial is *exactly* replicated. The only thing being replicated across m-AFC trials is the fact that there is one studied item and $m-1$ non-studied items.

The Testability of the Block-Marschak Inequalities in Recognition Memory

Given our previous discussion on the flexibility of SDT family, and its ability to capture any yes-no ROC, it is reasonable to ask how much do the Block-Marschak inequalities constrain choice probabilities in m-AFC recognition trials. This is an important question, given that the support for any given model or family of models is more compelling the more strict or limited is the range of permissible predictions (Roberts & Pashler, 2000). As shown below, the inequalities turn out be very strict under reasonable experimental designs.

First, let us consider some concrete cases that the Block-Marschak inequalities *cannot* accommodate, beyond context effects (e.g., Trueblood et al., 2013). At the core of the Block-Marschak inequalities is the assumption that choice probabilities across different choice subsets can be induced from a probability distribution of rank orders, which in turn is determined by a joint distribution of latent-strength values. If the way in which individuals evaluate options differs across option subsets, then this assumption is violated, leading to choice probabilities that are at odds with the Block-Marschak inequalities. As a first example, consider the $P_S^{\langle m \rangle}$ vector $[.84, .75, .69, .55, .46, .39, .34]$, from $m = 2$ to $m = 8$. At first glance, this vector appears to be perfectly plausible, as performance is always above chance and regularity is satisfied. However, it violates several Block-Marschak inequalities. For instance, $P_S^{\langle 3 \rangle} - 2P_S^{\langle 4 \rangle} + P_S^{\langle 5 \rangle} = 0.74 - 1.32 + 0.55 = -0.08$. These choice probabilities are associated with a decision maker who perfectly follows an unequal-variance Gaussian SDT model with $\mu_S = 1.5$ and $\sigma_S^2 = 1.3$, but cannot rank more than four alternatives, perhaps due to some limitation in their working memory capacity (e.g., Cowan, 2001).[10] A similar example involves participants shifting from a relative comparison of latent strengths to a serial, self-terminating search, in which a sequence of options is compared with a response criterion κ until one of them surpasses it (see Starns, Chen, & Staub, 2017).[11]

For a more general evaluation of the testability of the Block-Marschak inequalities, we considered vectors of $P_S^{\langle m \rangle}$, going from $m = 2$ to $m = 8$. These vectors can be represented as points on a seven-dimensional unit hypercube. Because this space includes many choice-probability vectors deemed unreasonable in the context of a

---

[10] Forced-choice accuracy here is given by $P_S^{\langle m \rangle} = \int_{-\infty}^{\infty} F_N(x)^{m-1} f_S(x)\, \mathrm{d}x$, for $m \leq 4$, and $P_S^{\langle m \rangle} = \frac{4}{m} \int_{-\infty}^{\infty} F_N(x)^3 f_S(x)\, \mathrm{d}x$, for $m > 4$.

[11] As in the previous example, we assumed a Gaussian SDT model with parameters $\mu_S = 1.5$ and $\sigma_S^2 = 1.3$, as well as $\kappa = 0.75$. We assumed that if no option is recognized in the serial search (i.e., all their latent strengths are below κ), then a recognition judgment is produced by randomly selecting one of the options. The probability of a correct response under this serial-search process is given by $P_S^{\langle m \rangle} = \frac{1}{m} F_N(\kappa)^{m-1} F_S(\kappa) + \frac{1}{m} \sum_{i=0}^{m-1} F_N(\kappa)^i (1 - F_S(\kappa))$. In the present example, we assumed a mixture of comparative judgments and serial search, with the probability of the former taking place being $\omega = (.95, .90, .80, .40, .20, .10, .05)$. The predictions of this model also violate several inequalities. For instance, it predicts that $P_S^{\langle 3 \rangle} - 3P_S^{\langle 4 \rangle} + 3P_S^{\langle 4 \rangle} - P_S^{\langle 6 \rangle} = -0.06$.

forced-choice recognition task, it is important to narrow our focus: First, we will only consider choice probabilities at or above chance (i.e., $P_S^{\langle m \rangle} \geq \frac{1}{m}$). Second, we will only consider choice probabilities that satisfy regularity, such that accuracy decreases as the number of alternatives increases (i.e., $P_S^{\langle 2 \rangle} \geq P_S^{\langle 3 \rangle} \geq ... \geq P_S^{\langle 8 \rangle}$). The satisfaction these two properties, which correspond to the first two inequalities in (12), can be seen as the minimum requirement for any set of choice probabilities to be deemed 'reasonable'. With the support of `R` packages `geometry` (Habel et al., 2015) and `vertexenum` (Robere, 2018), we found that the satisfaction of these two properties restricts predictions to approximately $\frac{1}{13{,}021}$ of the probability space. When applying the same computations to the *entire system* of Block-Marschak inequalities described in (12), we found that only $\frac{1}{37{,}405{,}425}$ of this already restricted space are permitted. What this means is that the constraints imposed by Block-Marschak inequalities (beyond above-chance performance and regularity) are very strict.

Goodness of Fit and Statistical Power

The statistical testing of inequality constraints such as (12) is going to be conducted as follows: The likelihood of the (presumed i.i.d.) discrete choices observed in each m-AFC condition will be described by a joint binomial model, with each probability parameter corresponding to our estimate of a given $P_S^{\langle m \rangle}$. More specifically, we will estimate the probability parameters under which the choice data are most likely, while enforcing the inequality constraints established by (12), for example (for an overview of this approach, see Riefer & Batchelder, 1988). The misfit of the data under these parameter estimates will be quantified by the G^2 statistic. Unfortunately, the evaluation of this statistic is complicated by the fact that its sampling distribution under the null hypothesis that the inequalities hold does not follow a χ^2 distribution (for a review, see Davis-Stober, 2009). In the present work, we addressed this challenge by computing p-values using a double semi-parametric bootstrap procedure (see Kalish, Dunn, Burdakov, and Sysoev, 2016; for further details, see also van de Schoot, Hoijtink,

& Dekovíc, 2010; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004).[12]

We assessed statistical power by generating discrete choice data from choice probabilities $P_S^{\langle m \rangle}$, for $m \in \{2, ..., 8\}$. These probabilities were all above chance and satisfied regularity.[13] Based on the previous analysis on the volume of predictions consistent with the Block-Marschak inequalities, we know that these data-generating probabilities will in all likelihood violate them. The simulated choice frequencies were then fit by a joint binomial model whose probabilities were constrained to conform to the Block-Marschak inequalities. When simulating 500/1000 choice trials per m-AFC condition (corresponding approximately to the samples sizes of Experiments 2 and 1, respectively), we observed statistically-significant misfits ($p < .05$) 86%/95% of the times. These simulation results complement the previous volume analysis of the Block-Marschak inequalities: Not only are they highly restrictive, their violation can be reliably detected in experiments with reasonable sample sizes.

Reanalysis of Swets (1959)

As a first empirical test of the Block-Marschak inequalities as expressed in (12), we reanalyzed data from an auditory detection task originally reported by Swets (1959). Three participants with "considerable practice" in psychophysical studies engaged in 2-, 3-, 4-, 6-, and 8-AFC trials. The signal was a tone of 1000 Hertz. Each alternative had a duration of 100 milliseconds (ms), and were separated by a 600 ms interval. Participants received feedback after each response. Swets (1959) used an incomplete sequence of m alternatives, with $m = 5$ and $m = 7$ missing. This omission does not allow us to test the Block-Marschak inequalities exactly as described in (12) given that they require complete sequences of m. We overcame this problem by canceling the $P_S^{\langle 5 \rangle}$ and $P_S^{\langle 7 \rangle}$ terms in the set of inequalities through the weighted sum of inequalities in

[12] Note that this double-bootstrap procedure overcomes the biases observed in the application of typical bootstrap procedures to cases of order-constrained inference (van de Schoot et al., 2010).

[13] For each m, response probabilities were obtained by generating independent random values from a uniform distribution ranging between $\frac{1}{m}$ and .85. Cases in which regularity was not satisfied were discarded. These probabilities were then used to generate response frequencies from a joint binomial distribution.

which the to-be-cancelled terms have opposite signs.

As shown in Figure 4, only one of the three individual datasets deviates noticeably from the best-fitting expectations that respect the Block-Marschak inequalities (Subject 3; $G^2 = 4.17$, $p = .10$). The cause of misfit here is a slight violation of regularity between $P_S^{\langle 3 \rangle}$ and $P_S^{\langle 4 \rangle}$. Visual inspections were consistent with the bootstrap-based p-values obtained for each individual participant. Overall, the results show that at least two of the individual datasets reported by Swets (1959) were consistent with the null-hypothesis that the Block-Marschak inequalities are satisfied.

Experiment 1: Testing the Random-Scale Representation in Recognition Memory

In this experiment, we collected m-AFC recognition judgments, with m ranging between 2 and 8. In each test trial, participants were presented with m alternatives, one of them being a studied item, and requested to choose the item they believed to have been previously studied. One key aspect of this experiment is that we collected a small number of trials per m-AFC condition, and placed our focus on the aggregated data. This move was motivated by the fact that the inequalities cannot be spuriously rejected due data aggregation.

Participants, Materials, and Procedure

One-hundred and ten participants took part in this study online. The participants were recruited through `Figure Eight` (www.figure-eight.com), and received a fixed \$2.50 reward in exchange for their participation. The experiment took roughly 10 minutes to complete. The experiment began with a study phase in which participants were presented with a list of 70 common nouns, each presented for 2000 ms, with a 400 ms interval between each word. The study list was presented twice in random order without a break between the two presentations. An additional primacy/recency buffer of five words was presented at the beginning and end of the study phase. These buffer words were not tested. After the study phase, participants initiated the test phase, which was comprised of 70 test trials. The test trials were comprised of 10 trials per

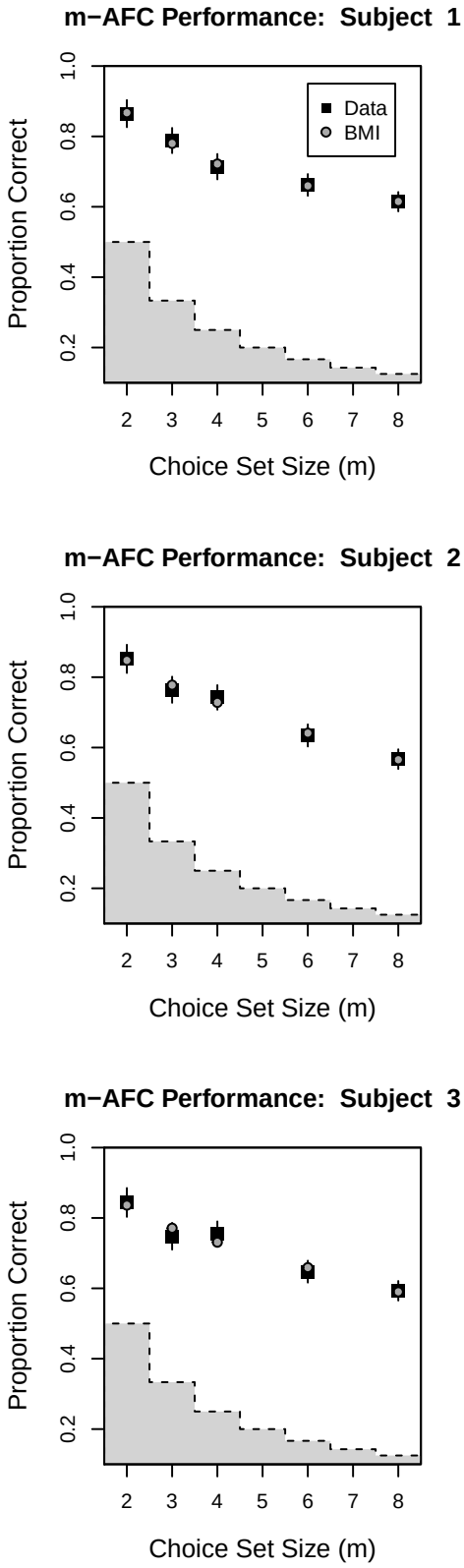

Figure 4. Analysis of individual data from Swets (1959). The bars represent the marginal 95% confidence intervals. BMI = Best-fitting predictions that respect the Block-Marschak inequalities.The dashed lines delimiting the gray areas indicate chance-level performance.

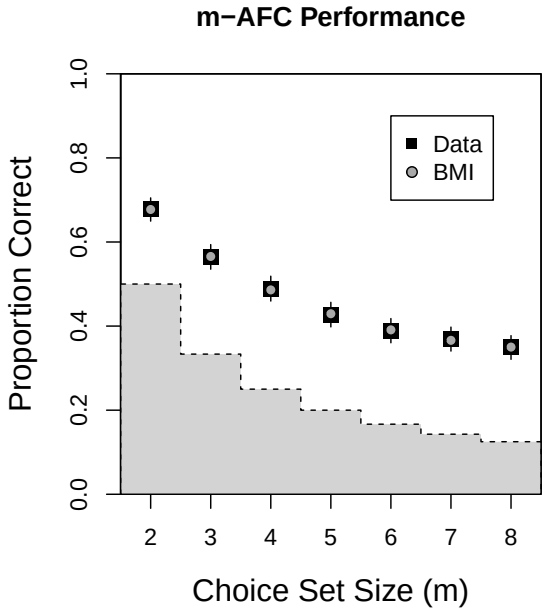

Figure 5. Experiment 1 Results. Observed and predicted performance in the m-AFC trials. Bars represent the marginal 95% confidence intervals. BMI = Best-fitting estimates that respect the Block-Marschak inequalities. The dashed lines delimiting the gray areas indicate chance-level performance.

choice set-size condition (randomly intermixed), with set sizes ranging from $m = 2$ to $m = 8$ in steps of 1 (i.e., roughly 1000 trials per m-AFC condition in total). Words were presented in the center of the screen. In trials involving smaller set sizes and/or shorter words, all test items were presented next to each other. In trials involving larger set sizes and/or longer words, test-item presentation was split into more than one row, with several words in each row presented next to each other. Participants were informed that their task was to select the one studied word from the m presented words. They selected the word of their choice by simply clicking on it. They moved on to the next test trial after confirming their choice. After completing all of the test trials, participants filled in a short demographic survey, were thanked, and received their monetary reward.

Results and Discussion

As shown in Figure 5, forced-choice accuracy was clearly above chance for all choice set sizes m. Regularity was also satisfied, with performance decreasing as m increased. But do the data respect the complete system of Block-Marschak inequalities? The model fits shown in Figure 5 indicate a near-perfect fit, with $G^2 = 0.13$, $p = .98$.

Note that our previous simulations indicate that this experiment is suitably powered to detect violations. These fit results show that the data are in very close agreement with a quite strict set of inequality constraints, providing empirical support for a very large family of SDT models in the literature.

Step 2: Latent-Variable Independence

As previously discussed, the Block-Marschak inequalities make no assumptions regarding the presence or absence of dependencies among the latent variables associated with each option. This is also true in the case of recognition memory discussed here, in which noise variables are assumed to be exchangeable. However, many SDT models go one step further and assume that the latent variables are *independent* (Green & Swets, 1966; Kellen & Klauer, 2018; Macmillan & Creelman, 2005; Wickens, 2002). For reference, latent-variable independence means that the joint cumulative distribution function of latent-strength values corresponds to the product of their marginal cumulative distribution functions (e.g., $F_{S,N}(x, y) = F_S(x) \times F_N(y)$ for all permissible x and y values). In this section, we discuss a direct test of this assumption and establish an additional predictive test based on the relationships between forced choice, ranking, and yes-no judgments that follow from latent-variable independence.[14]

Sattath and Tversky (1976) formally showed that the assumption of latent-variable independence implies a number of multiplicative inequalities at the level of choice probabilities (see also Shaw, 1980; Suck, 2002; Suppes et al., 1989). Let $A, B \subseteq T$ be option subsets including alternative a (i.e., $a \in A \cap B$). Then:

$$P_a^{\langle A \cup B \rangle} \geq P_a^{\langle A \rangle} \times P_a^{\langle B \rangle}. \tag{13}$$

In the context of an m-AFC task, we obtain the following set of multiplicative

[14] Please note that the issue of latent-variable independence is distinct from the question of whether or not discrete choices are independent and identically distributed (i.i.d.).

inequalities:

$$
\begin{aligned}
&P_S^{\langle 3\rangle} \geq (P_S^{\langle 2\rangle})^2, \\
&P_S^{\langle 4\rangle} \geq (P_S^{\langle 2\rangle})^3, \;\; P_S^{\langle 2\rangle} \times P_S^{\langle 3\rangle}, \\
&P_S^{\langle 5\rangle} \geq (P_S^{\langle 2\rangle})^4, \;\; (P_S^{\langle 2\rangle})^2 \times P_S^{\langle 3\rangle}, \;\; (P_S^{\langle 3\rangle})^2, \;\; P_S^{\langle 2\rangle} \times P_S^{\langle 4\rangle}, \\
&\textit{etc.}
\end{aligned}
\tag{14}
$$

These multiplicative inequalities are *necessary* under latent-variable independence, although *not sufficient* (see Sattath & Tversky, 1976). What this means is that these inequalities cannot be used to dismiss models assuming dependent latent variables. However, the satisfaction of the multiplicative inequalities under strict testing conditions can still be used to motivate the exploration and testing of models that assume independent latent variables (for a similar testing approach, see McCausland & Marley, 2014).

Some of the multiplicative inequalities implied by latent-variable independence might not be obvious at first glance: For example, consider the values $P_S^{\langle 2\rangle} = .90$, $P_S^{\langle 3\rangle} = .70$, and $P_S^{\langle 4\rangle} = .60$. These values might seem reasonable; after all they are all above chance, satisfy regularity, and satisfy one of the Block-Marschak inequalities, $P_S^{\langle 2\rangle} - 2P_S^{\langle 3\rangle} + P_S^{\langle 4\rangle} = 0.10 \geq 0$. However, they cannot be accommodated by the sub-family of SDT models that satisfy latent-variable independence. To see this, simply note that it violates the first inequality in (14), as $.70 < .90^2$.

Using the same simulation procedure described in Footnote 13, we generated vectors of correct-response probabilities $P_S^{\langle m\rangle}$, from $m = 2$ to $m = 8$. These probabilities were all above chance and satisfied regularity. We then checked how many these vectors also conformed to the multiplicative inequalities described above (i.e., were consistent with latent-variable independence). We found that only 7% of them did, which means that latent-variable independence is not easily satisfied, even when focusing on 'reasonable' choice probabilities.

In order to determine our ability to detect violations of multiplicative inequalities, we used the non-conforming probabilities obtained in the previous simulation to

generate artificial choice data, with 500/1000 trials per m-AFC condition. When fitting these artificial data with a joint binomial model that conforms to the multiplicative inequalities, we found it to yield statistically-significant misfits ($p < .05$) in 85%/91% of cases. These simulation results suggest that experiments with the sample sizes considered (which are comparable to Experiments 1 and 2) are well suited to detect violations of latent-variable independence.

When applying the same model-fitting procedure to the data from Experiment 1, we obtained a perfect fit ($G^2 = 0$, $p = 1$), which indicates that the observed choice proportions are entirely consistent with the hypothesis that latent-variable independence holds. This result is quite impressive given that most 'reasonable' choice probabilities do not satisfy the multiplicative inequalities, and that the sample size used in Experiment 1 is able to detect violations (Roberts & Pashler, 2000).

The Relationship Between Forced-Choice, Ranking, and Yes-No Judgments

Now that we have established some of the inequality constraints imposed by latent-variable independence, and found the data from Experiment 1 to be in perfect alignment with them, we turn to some of the implications that follow from this property. Under the assumption of latent-variable independence, it can be shown that each $P_S^{\langle m \rangle}$ corresponds to the $m-1$th moment of the yes-no ROC function ρ. Once again, relying on the universal SDT representation:

$$\begin{aligned} P_S^{\langle m \rangle} &= P(\lambda_S > \max(\lambda_{N,1}, \lambda_{N,2}, ..., \lambda_{N,m-1})) \\ &= \int_0^1 F_N(t)^{m-1} f_S(t)\, \mathrm{d}t \\ &= \int_0^1 t^{m-1}\, \mathrm{d}F_S(t) \\ &= \mathbb{E}(\lambda_S^{m-1}), \end{aligned} \tag{15}$$

with $\mathbb{E}(\cdot)$ being the expectation operator. Moments are quantities describing a function (the first moment is the mean, the second *central* moment is the variance, etc.). Because the ROC function is bounded between 0 and 1, it can be fully described by its moments

(see Feller, 1971, Chap. 7). Note that $P_S^{\langle 2\rangle} = \mathbb{E}(\lambda_S)$, which means that the probability of a correct response in 2-AFC trials corresponds to the area under the yes-no ROC function. This equality is the famous *Area Theorem* established by Green and Moses (1966). The formal result in (15), which includes Green and Moses' Area Theorem as a special case, was coined the *Generalized Area Theorem* by Iverson and Bamber (1997).

Based on (15), we can determine the probabilities associated with *ranking judgments*, which have also been modeled using SDT (e.g., Kellen & Klauer, 2014; Kellen et al., 2012; McAdoo & Gronlund, 2016). In a typical ranking task, the decision maker is requested to rank the options according to her belief that they are the signal (rank 1 being the highest, m the lowest). Let $R_i^{\langle m\rangle}$ denote the probability of the signal stimulus being assigned rank i among m alternatives. Under the assumption that the latent variables are independent, signal-ranking probabilities are given by:

$$\begin{aligned} R_i^{\langle m\rangle} &= \binom{m-1}{i-1} \int_0^1 (1-F_N(t))^{i-1} F_N(t)^{m-i} f_S(t)\, \mathrm{d}t \\ &= \binom{m-1}{i-1} \int_0^1 (1-t)^{i-1} t^{m-i}\, \mathrm{d}F_S(t), \end{aligned} \tag{16}$$

with the binomial coefficient $\binom{m-1}{i-1}$ counting the number of ways that the signal option can be outranked by $i-1$ out of $m-1$ noise options.

The expansion of the integrand $(1-t)^{i-1}t^{m-i}$ in Equation 16 provides us with a simple way to express signal-ranking probabilities $R_i^{\langle m\rangle}$ in terms of forced-choice probabilities $P_S^{\langle m\rangle}$. For example, consider the probability of the signal being assigned rank 3 for $m = 4$:

$$\begin{aligned} R_3^{\langle 4\rangle} &= 3\int_0^1 (1-t)^2 t\, \mathrm{d}F_S(t) \\ &= 3\int_0^1 t^3 - 2t^2 + t\, \mathrm{d}F_S(t) \\ &= 3 \times \left(P_S^{\langle 4\rangle} - 2P_S^{\langle 3\rangle} + P_S^{\langle 2\rangle}\right). \end{aligned} \tag{17}$$

Note that the $P_S^{\langle m\rangle}$ summands in (17) correspond to one of the linear functions in the Block-Marschak inequalities (see third line of (12)) scaled by $\binom{m-1}{i-1} = 3$. This

correspondence reasserts the intimate relationship between the Block-Marschak inequalities and ranking probabilities across subsets – a violation of some of the inequalities would imply *negative* signal-ranking probabilities, a nonsensical scenario. Based on these observations, it shouldn't come as a surprise that there is also a close relationship between ranking and yes-no judgments. The theoretical result below will help us understand this relationship precisely.

Analogous to $R_i^{\langle m \rangle}$, let $Q_i^{\langle m \rangle}$ be the probability of some noise option being assigned rank i among m options. Given the exchangeability of noise options, it follows that

$$Q_i^{\langle m \rangle} = \frac{1 - R_i^{\langle m \rangle}}{m - 1}. \tag{18}$$

Using formal results reported by Feller (1971, Chap. 7), Iverson and Bamber (1997) showed that as $m \to \infty$, the cumulative sums $\sum_{i=1}^{t} R_i^{\langle m \rangle}$ and $\sum_{i=1}^{t} Q_i^{\langle m \rangle}$, $t = 1, \ldots, m$, converge to the universal representations of the signal and noise's cumulative distribution functions, F_S and F_N. Figure 6 illustrates the accuracy of this approximation using the cumulative sums of $R_i^{\langle m \rangle}$ and $Q_i^{\langle m \rangle}$ across different choice set sizes m.

Based on the latter result, it is possible to reconstruct the underlying yes-no ROC function from recognition-memory studies that have collected ranking judgments. Such experiments were conducted by Kellen and Klauer (2014) and McAdoo and Gronlund (2016). In both sets of experiments, the strength of the studied items was manipulated via study repetition (e.g., weak items being studied once and strong items thrice). In Kellen and Klauer's experiments the items were common words, whereas in McAdoo and Gronlund's experiments they were human faces. As shown in Figure 7, the reconstructed yes-no ROCs appear to be slightly concave with ROCs corresponding to strong studied items dominating their weak counterparts. Also, these ROCs appear to be asymmetric relative to the negative diagonal. Note, however, that the small option set sizes in each of these studies ($m = 3$ and $m = 4$) only allows for rather crude reconstructions (see Figure 6). Figure 8 illustrates the reconstructed ROCs obtained with the four-alternative "answer-until-correct" forced-choice paradigm used by

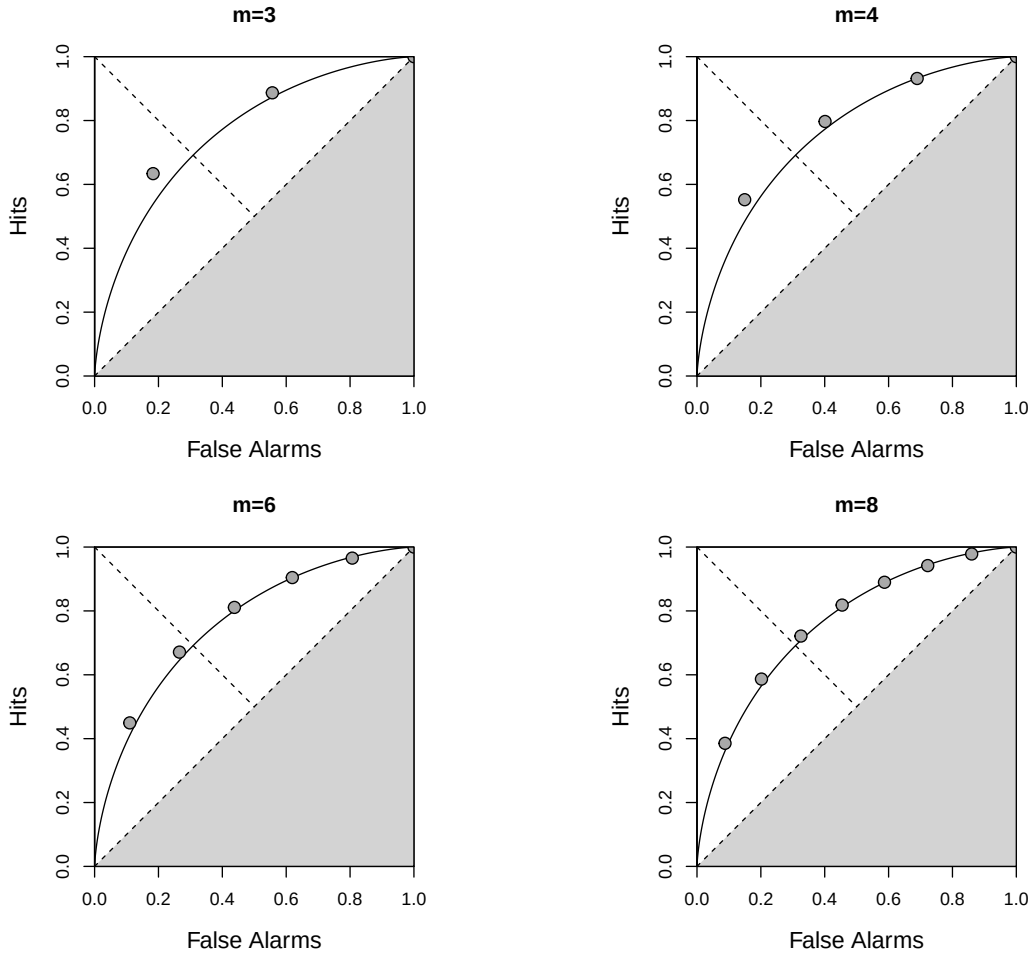

Figure 6. Theoretical and reconstructed yes-no ROC functions (lines and points, respectively) for different choice set sizes m.

Chechile, Sloboda, and Chamberland (2012). According to SDT, the choices made in such a paradigm should follow the ranking of the different alternatives. Again, we observe concave and asymmetric ROCs, with ROCs from "stronger" study conditions dominating their "weaker" counterparts.

Step 3: Likelihood-Ratio Monotonicity

The relationship between ranking judgments and the yes-no ROC function provides an important insight into the constraint that the slope of the latter should be monotonically decreasing (i.e., the ROC is concave). As shown by Equation 4, a decreasing ROC slope follows from the fact that the likelihood ratio $\frac{f_S(\kappa)}{f_N(\kappa)}$ is monotonically increasing, which is in line with the widely-held assumption that small/large latent-strength values are more likely under the noise/signal distribution

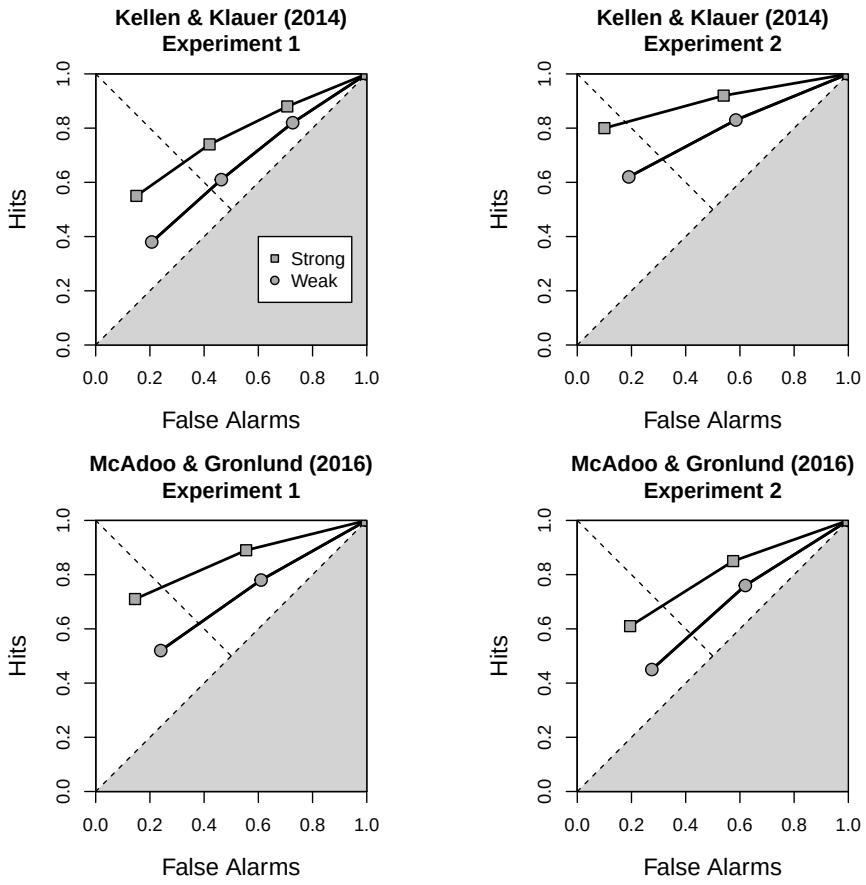

Figure 7. Reconstructed yes-no ROC functions based on ranking judgments.

(e.g., Criss & McClelland, 2006; Glanzer et al., 2009; Osth & Dennis, 2015).

Now, let us go back to the reconstruction of the ROC function based on rankings. In order for the reconstructed ROC function to be concave (i.e., for likelihood-ratio monotonicity to hold), the signal-rank probabilities $R_i^{\langle m \rangle}$ need to be monotonically decreasing with respect to rank position, such that:

$$R_i^{\langle m \rangle} - R_j^{\langle m \rangle} \geq 0, \tag{19}$$

for $1 \leq i < j < m$. To see how ROC concavity (i.e., likelihood-ratio monotonicity) imposes this order constraint, simply note that the reconstructed ROC function is piecewise linear, and that the slope of the linear segment connecting the ith and $(i+1)$th ROC point corresponds to the ratio $\frac{R_{i+1}^{\langle m \rangle}}{Q_{i+1}^{\langle m \rangle}}$. It should also be noted that a mixture of concave functions (e.g., ROCs) will always yield a concave function, which means that (19) cannot be spuriously violated by aggregation.

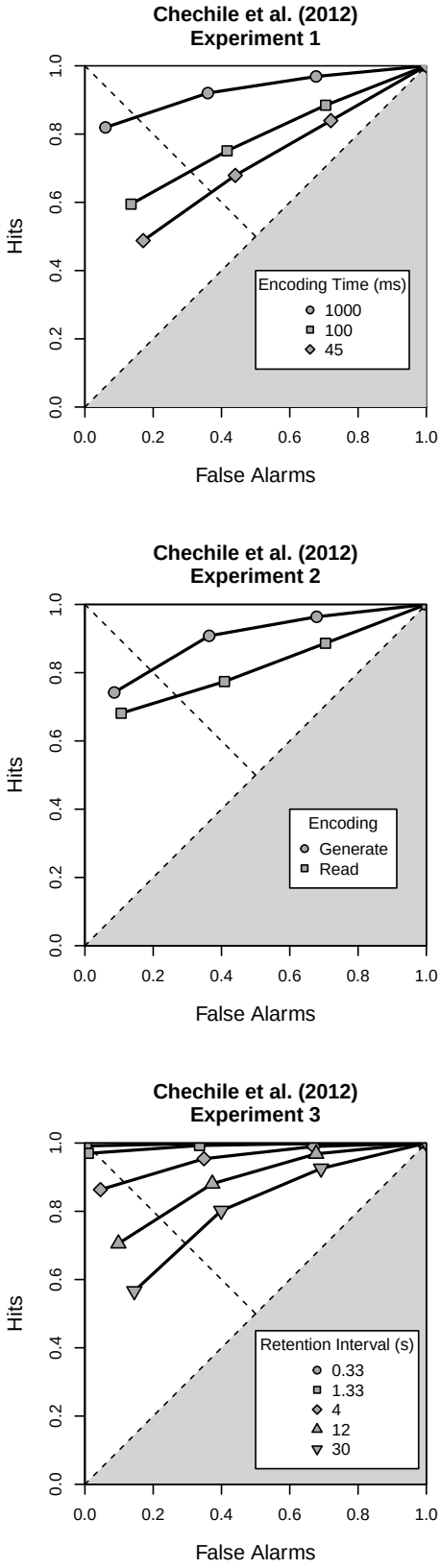

Figure 8. Reconstructed yes-no ROC functions based on "answer-until-correct" judgments reported by Chechile et al. (2012).

The order constraint in (19) leads us to the following insight: In terms of ranking judgments, likelihood-ratio monotonicity implies that *more egregious errors are less probable than more moderate errors*. For instance, assigning rank 5 to a signal option is less probable than assigning rank 4, which in turn is less probable than assigning rank 3, and so forth.

As shown in (17), rank probabilities $R_i^{\langle m\rangle}$ can be expressed in terms of forced-choice probabilities $P_S^{\langle m\rangle}$, which allows us to recast the inequality constraint described in (19). For example, in the case of forced-choice probabilities up to $m = 4$:

$$\begin{aligned}
\underbrace{P_S^{\langle 4\rangle}}_{R_1^{\langle 4\rangle}} - \underbrace{3\left(P_S^{\langle 3\rangle} - P_S^{\langle 4\rangle}\right)}_{R_2^{\langle 4\rangle}} &\geq 0, \\
\underbrace{3\left(P_S^{\langle 3\rangle} - P_S^{\langle 4\rangle}\right)}_{R_2^{\langle 4\rangle}} - \underbrace{3\left(P_S^{\langle 4\rangle} - 2P_S^{\langle 3\rangle} + P_S^{\langle 2\rangle}\right)}_{R_3^{\langle 4\rangle}} &\geq 0, \\
\underbrace{3\left(P_S^{\langle 4\rangle} - 2P_S^{\langle 3\rangle} + P_S^{\langle 2\rangle}\right)}_{R_3^{\langle 4\rangle}} - \underbrace{\left(-P_S^{\langle 4\rangle} + 3P_S^{\langle 3\rangle} - 3P_S^{\langle 2\rangle} + P_S^{\langle 1\rangle}\right)}_{R_4^{\langle 4\rangle}} &\geq 0.
\end{aligned} \tag{20}$$

This system of inequalities can be imposed alongside the Block-Marschak inequalities, allowing us to test for a random-scale representation comprised of latent variables that satisfy likelihood-ratio monotonicity. When imposing these additional constraints over a sequence of m-AFC judgments from $m = 2$ to $m = 8$, we observe a considerable reduction of the volume of permitted choice probabilities.[15] Specifically, this volume now corresponds to $\frac{1}{413{,}121{,}934{,}659}$ of the already restricted volume of $P_S^{\langle m\rangle}$ that are above chance and satisfy regularity. Simulations showed that with a sample size of 500/1000 trials per m-AFC condition, the percentage of cases in which this extended set of inequality constraints is rejected ($p < .05$), is around 95%/98% (for details on this simulation, see Footnote 13).

We tested the constraints imposed by likelihood-ratio monotonicity along with the Block-Marschak inequalities, using the data from Experiment 1. The resulting fit was

[15] The intersection between two convex polytopes – such as the ones defined by the random-scale representation and likelihood-ratio monotonicity – is itself a convex polytope (see Grünbaum, 2003, Chap. 3). The vertex representation of the latter is reported in the online Supplemental materials (see Author Note).

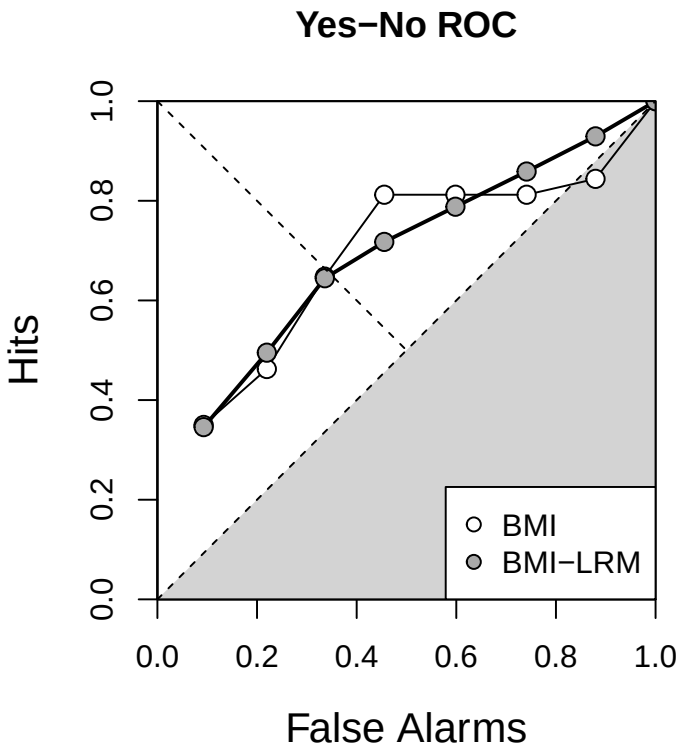

Figure 9. Reconstructed yes-no ROC of the data from Experiment 1 using the predicted forced-choice performance under the Block-Marschak inequalities without (BMI) and with likelihood-ratio monotonicity (BMI-LRM; white and gray points, respectively). The dashed line delimiting the gray area indicates chance-level performance.

still very good ($G^2 = 0.88, p = .93$). We also reconstructed the yes-no ROC function from Experiment 1 and assessed the impact of the constraints introduced by likelihood-ratio monotonicity. To do this, we converted the best-fitting $P_S^{\langle m \rangle}$ values obtained under each set of constraints (i.e., random-scale representation with and without likelihood-ratio monotonicity) into ranking probabilities $R_i^{\langle m \rangle}$. The reconstructed ROCs illustrated in Figure 9 show that the constraints imposed by likelihood-ratio monotonicity removed a couple of small violations of concavity. Note that the small difference in misfit found between the two tests suggests that these violations of concavity are well within what one would expect under the null hypothesis that all of the inequalities coming from random-scale representation *and* likelihood-ratio monotonicity are satisfied.

Experiment 2: Further Testing of Latent-Variable Independence and Likelihood-Ratio Monotonicity

When applied to the results of Experiment 1, the constraints imposed by a random-scale representation, as well as latent-variable independence and likelihood-ratio monotonicity, were all found to hold. The aim of Experiment 2 is to replicate and extend these findings, using a slightly modified design in which yes-no judgments for single items are also collected. These yes-no judgments will enable us to conduct a predictive test of independence and likelihood-ratio monotonicity: If the reconstructed concave ROC function accurately captures the relationship between hits and false alarms, then it should be able to capture the observed yes-no ROC point.

Participants, Materials, and Procedure

One-hundred and three new participants took part in this study online. As before, the participants were recruited through `Figure Eight` and received a fixed $2.50 reward in exchange for their participation. This experiment was identical to Experiment 1, with two changes. First, we introduced twenty single-item trials (10 studied and 10 non-studied) in which participants were requested to judge whether the item was previously studied, responding "`yes`" or "`no`". Second, we alleviated the task demands by reducing the number of m-AFC trials to five per m (i.e., roughly 500 trials per m-AFC condition in total).

Results and Discussion

The data are shown in Figure 10. As before, we found performance to be above chance and regularity satisfied. Note that performance was generally better than in Experiment 1, a result that can be attributed to the fewer number of m-AFC trials collected per participant and its relation with output interference (Criss, Malmberg, & Shiffrin, 2011; Murdock & Anderson, 1975): The more items one encounters throughout the test phase, the more performance should be impaired (see Murdock & Anderson, 1975, Table 7).

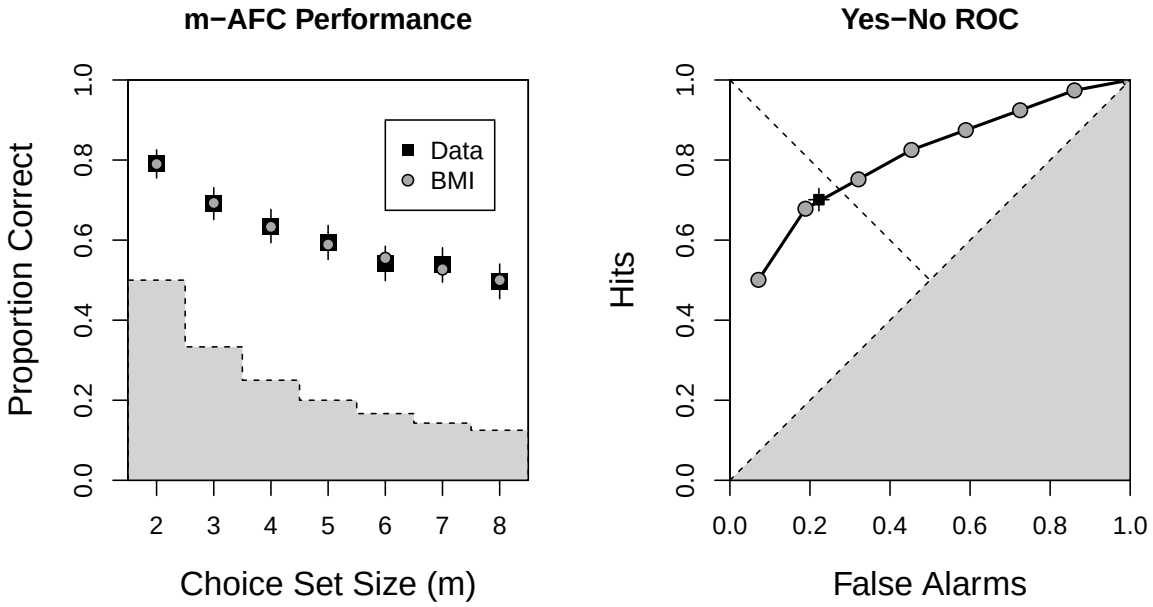

Figure 10. Experiment 2 Results. *Left Panel*: Observed and predicted performance in the m-AFC trials. Bars represent the marginal 95% confidence intervals. BMI = Best-fitting estimates that respect the Block-Marschak inequalities. *Right Panel*: Reconstructed yes-no ROC using the predicted forced-choice performance under the Block-Marschak inequalities and likelihood-ratio monotonicity. Black square represents the observed hit and false-alarm probability. In both panels, the dashed lines delimiting the gray areas indicate chance-level performance.

The data were once again consistent with the Block-Marschak inequalities ($G^2 = 0.70$, $p = .91$). As before, enforcing likelihood-ratio monotonicity leads to a virtually identical fit ($G^2 = 0.74$, $p = .95$), and the multiplicative inequalities were once again perfectly satisfied ($G^2 = 0$, $p = 1$). The right panel of Figure 10 shows the reconstructed yes-no ROC. Again, this function takes on a concave, asymmetric shape. Most importantly, the yes-no data point is entirely consistent with the ordered ROC points in the sense that they can all be described by a single monotonically-increasing function with monotonically decreasing slope. Altogether, these results provide support for the sub-family of SDT models that satisfy latent-variable independence and likelihood-ratio monotonicity.

Step 4: Yes-No ROC Symmetry

The reconstructed yes-no ROCs obtained so far are all *asymmetric* relative to the negative diagonal (see Figures 7-10). In the domain of recognition memory, ROC asymmetries of this kind have been a major motivation for the development of

extensions of the equal-variance Gaussian SDT model. These extensions involve a myriad of theoretical concepts such as variable encoding, attentional failure, or additional retrieval processes (for a review of different models, see Yonelinas & Parks, 2007). In more applied domains, the violation of ROC symmetry has important implications for our ability to assess and compare different decision makers (for a discussion, see Rotello et al., 2015).

Assessments of ROC symmetry almost invariably rely on confidence-rating judgments (e.g., Yonelinas & Parks, 2007). However, there are a number issues that speak against this approach: For instance, a recent critical test by Kellen and Klauer (2015) showed that confidence ratings do not behave as expected under a large sub-family of models that includes the Gaussian SDT model (for a replication of these results, see McAdoo et al., 2018; for other issues, see Benjamin, Tullis, & Lee, 2013; Van Zandt, 2000). Given these issues, some researchers have considered alternatives such as constructing ROCs using binary yes-no judgments obtained across different response-bias conditions (e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012). But unfortunately, this alternative approach often leads to extremely noisy data that can fail to meet some basic selective-influence assumptions (both response criterion and memory discriminability might be affected; see Kellen, Klauer, & Bröder, 2013; Van Zandt, 2000). As a solution to these challenges, we propose a direct test of ROC symmetry that hinges on a simple equality prediction across response probabilities, and does not depend upon any parametric assumptions.

Formally, an ROC function ρ is symmetric if and only if the inclusion of point $\{\text{FA}, \text{H}\}$ implies the inclusion of point $\{1 - \text{H}, 1 - \text{FA}\}$ (Iverson & Bamber, 1997; Killeen & Taylor, 2004). Figure 11 provides an illustration. This constraint can be expressed in terms of the following equality:

$$\rho(\text{FA}) + \rho^{-1}(1 - \text{FA}) = H + (1 - \text{H}) = 1, \tag{21}$$

where ρ^{-1} is the inverse of the ROC function. Under a universal representation, this

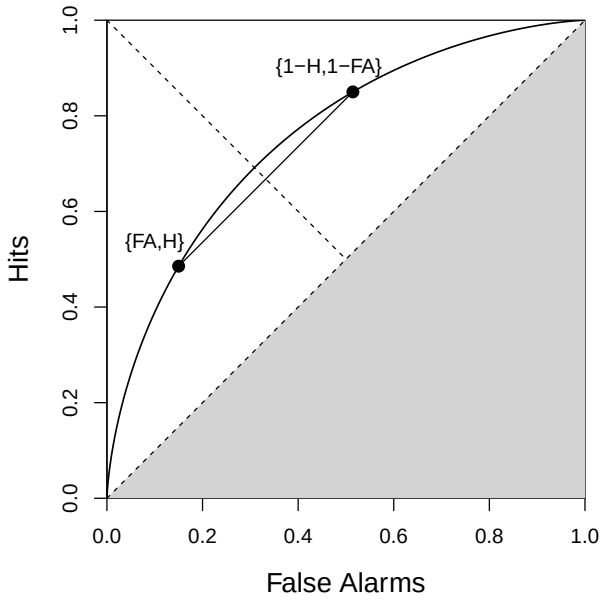

Figure 11. Illustration of a symmetrical yes-no ROC function and the constraint that it must include points $\{\text{FA}, \text{H}\}$ and $\{1 - \text{H}, 1 - \text{FA}\}$.

equality constraint can also be expressed in terms of F_S and F_S^{-1}:

$$F_S(t) + F_S^{-1}(1 - t) = 1. \tag{22}$$

Iverson and Bamber (1997) formally showed that ROC symmetry implies an equality between m-AFC judgments and judgments made in an *modified* m-alternative forced-choice task ($m^\star$-AFC) in which individuals are requested to choose the single noise option from among the $m - 1$ signal options.[16] To see this, let us denote the probability of a correct response (noise being chosen) by $P_N^{\langle m\star\rangle}$, which corresponds to the probability that the single λ_N value is smaller than all the $m - 1$ λ_S values. Again, using the universal SDT representation:

$$\begin{aligned} P_N^{\langle m\star\rangle} &= P(\lambda_N < \min(\lambda_{S,1}, \lambda_{S,2}, ..., \lambda_{S,m-1})), \\ &= \int_0^1 f_N(t)(1 - F_S(t))^{m-1} \, \mathrm{d}t. \end{aligned} \tag{23}$$

[16] Both m-AFC and $m^\star$-AFC judgments should not be confused with the judgments made in an *oddity task* (O'Connor, Guhl, Cox, & Dobbins, 2011). In the latter task participants are requested to choose the 'odd item' without knowing whether the m-alternative choice set includes $m - 1$ signal or $m - 1$ noise stimuli.

$$= \int_0^1 t^{m-1} \, \mathrm{d}(1 - F_S^{-1}(1-t)).$$

If symmetry holds, then $F_S(t) = 1 - F_S^{-1}(1-t)$ (see (22)) and therefore $P_N^{\langle m\star\rangle} = P_S^{\langle m\rangle}$ for all m (compare the penultimate and last lines of Equations 15 and 23, respectively). The predicted equality in choice accuracy between m-AFC and $m^\star$-AFC under ROC symmetry allows us to dismiss the latter if any $P_N^{\langle m\star\rangle}$ systematically differs from $P_S^{\langle m\rangle}$ for some m. Importantly, testing these predicted equalities does not require the elicitation of confidence judgments nor the use of response-bias manipulations.

With the type of asymmetry typically observed in recognition-memory ROCs, including the one reconstructed in Experiment 1, we expect $P_S^{\langle m\rangle} > P_N^{\langle m\star\rangle}$. For example, an unequal-variance Gaussian SDT model with parameters $\mu_S = 1$ and $\sigma_S^2 = 1.3$, which yields an asymmetric ROC, expects the following forced-choice probabilities: $P_S^{\langle 4\rangle} = .55$, $P_S^{\langle 5\rangle} = .49$, and $P_S^{\langle 6\rangle} = .45.$, whereas $P_N^{\langle 4\star\rangle} = .51$, $P_N^{\langle 5\star\rangle} = .45$, and $P_N^{\langle 6\star\rangle} = .40$.

Interestingly, the *opposite prediction* is made by some well-known parametric distributions. For instance, take the double-exponential distribution discussed by Yellott (1977). Yellott showed that under rather benevolent conditions (including latent-variable independence), the latent variables underlying preferences must be double-exponentially distributed, which in turn implies that choice probabilities must conform to *Luce's Choice Theorem* (Luce, 1959; see also Luce, 1977). For example, if the signal and noise distribution have location parameters 0 and 1, respectively, then $P_S^{\langle 4\rangle} = .48$, $P_S^{\langle 5\rangle} = .40$, and $P_S^{\langle 6\rangle} = .35.$, which are all respectively smaller than $P_N^{\langle 4\star\rangle} = .55$, $P_N^{\langle 5\star\rangle} = .50$, and $P_N^{\langle 6\star\rangle} = .47$.

Experiment 3: Testing ROC Symmetry

In this experiment, forced choice recognition judgments were obtained under signal-recognition instructions (m-AFC trials) and under noise-recognition instructions ($m^\star$-AFC trials).

Participants, Materials, and Procedure

Three-hundred and fifty-nine new participants were recruited online, again through `Figure Eight`. As before, a $2.50 reward was given in exchange for participation. The task participants engaged in (m-AFC vs. $m^\star$-AFC) was manipulated between subjects. We recruited 180 participants in the m-AFC condition and 179 participants in the $m^\star$-AFC condition. Of those, we had to exclude 10 participants in the $m^\star$-AFC condition who did not follow the instructions and indicated incorrectly in a post-experiment survey that their task was to select the *studied items*. The study and test phases were similar to the previous experiment, with the exception that we only considered forced-choice trials, and only for choice-set sizes $m = 4$, 5, and 6 (six test trials per m per participant, which resulted in roughly 1000 trials per $m/m^\star$-AFC condition in total). The focus on a limited number of choice set sizes was motivated by previous simulations showing that differences would be more easily detected with these set sizes. The reduction in the total amount of test trials was necessary when trying to have the same number of trials per condition. After all, each m-AFC trial involves only one studied item, whereas each $m^\star$-AFC trial involves $m - 1$ studied items.

Results and Discussion

The proportion of correct responses are illustrated in Figure 12. These proportions are all above chance and satisfy regularity, and although the $m^\star$-AFC judgments violate the Block-Marschak inequalities, as $P_N^{\langle 4\star\rangle} - 2P_N^{\langle 5\star\rangle} + P_N^{\langle 6\star\rangle} = -0.03$, this was not statistically significant ($G^2 = 0.74$, $p = .18$). We tested whether the equality hypothesis following from the assumption of ROC symmetry holds by comparing the goodness of fit of two joint binomial models. First, we fitted a model that imposed the inequality constraints $P_S^{\langle 4\rangle} \geq P_N^{\langle 4\star\rangle}$, $P_S^{\langle 5\rangle} \geq P_N^{\langle 5\star\rangle}$, and $P_S^{\langle 6\rangle} \geq P_N^{\langle 6\star\rangle}$. We then compared this fit with the one from another model imposing the equality constraints $P_S^{\langle 4\rangle} = P_N^{\langle 4\star\rangle}$, $P_S^{\langle 5\rangle} = P_N^{\langle 5\star\rangle}$, and $P_S^{\langle 6\rangle} = P_N^{\langle 6\star\rangle}$. The difference in fit between the two models was statistically significant, with $\Delta G^2 = 16.62$, $p < .001$, indicating that the equality constraint imposed by the hypothesis of ROC symmetry does not provide a

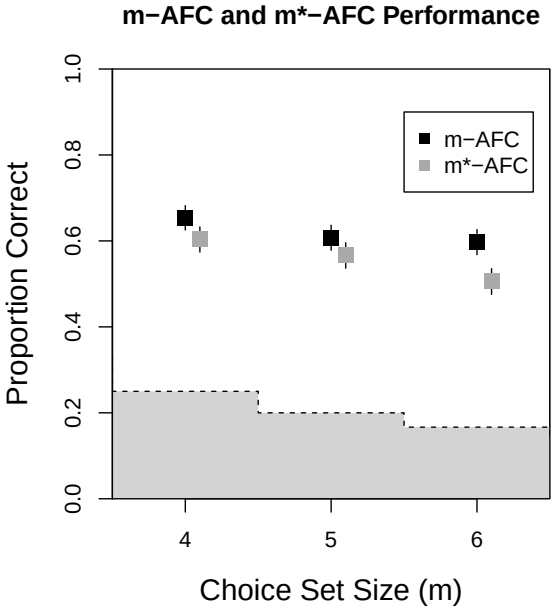

Figure 12. Experiment 3 results. The bars represent the marginal 95% confidence intervals.

reasonable characterization of the data. Furthermore, note that these results are also at odds with the latent-strength distributions assumed by Luce's Choice Theory (Luce, 1959, 1977; Yellott, 1977) as they predict the opposite pattern, namely $P_S^{\langle m \rangle} \leq P_N^{\langle m\star \rangle}$. The present results suggest that the yes-no ROC is asymmetric, corroborating what has been so far observed in ROCs constructed with response-bias manipulatons (e.g., Dube & Rotello, 2012; Kellen et al., 2013) or confidence ratings (e.g., Klauer & Kellen, 2015; Pratte, Rouder, & Morey 2010), or with ROCs reconstructed from forced-choice/ranking judgments (see Figures 7-10).

Finally, it is worth noting that the present approach for testing ROC symmetry has been applied in other domains: Trippas et al. (2018) conducted a large-scale meta-analysis on ROCs obtained with syllogistic-reasoning judgments, and found them to be *symmetric*. This meta-analytic result was corroborated by a subsequent comparison between m-AFC and $m^\star$-AFC judgments, which were not reliably different from each other. At least in the domains of syllogistic reasoning and recognition memory, the a/symmetry found in ROCs is in line with outcomes of this critical test.

Step 5: Threshold and Non-Threshold Representations

The results from the critical tests conducted up to this point are consistent with a random-scale representation that is comprised of independent latent-strength distributions that satisfy likelihood-ratio monotonicity, and that yield asymmetric yes-no ROCs. In this section, we will further divide this sub-family of models by introducing the distinction between '*threshold*' and '*non-threshold*' representations. This distinction hinges on *conditional independence*, a property that we will elaborate upon below (for earlier discussions on this property, see Kellen & Klauer, 2014, 2015; Province & Rouder, 2012; Rouder & Morey, 2009; Rouder et al., 2014).

According to threshold models, people's judgments are governed by a small number of mutually-exclusive discrete mental states that are entered with some probability. In turn, each of these mental states is associated with a probability distribution over the response alternatives – a *state-response mapping* (see Rouder & Morey, 2009; Klauer & Kellen, 2010). As an example, let us consider one of the best-studied threshold models in the literature, the *High-Threshold Model* (e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012). This model, which is illustrated in Figure 13 (see also the bottom row of Figure 3), postulates that signal and noise items can fall into one of three mutually-exclusive discrete mental states $\mathfrak{M}$. The probability with which each state can be reached differs between stimulus classes. In fact, some of the states can only be reached by certain stimulus classes:

- $\mathfrak{M}_1$: **Noise-Detection State**

  $\mathrm{P}(\textit{Signal Stimuli}) = 0$

  $\mathrm{P}(\textit{Noise Stimuli}) = q_N$

- $\mathfrak{M}_2$: **Uncertainty State**

  $\mathrm{P}(\textit{Signal Stimuli}) = 1 - p_S$

  $\mathrm{P}(\textit{Noise Stimuli}) = 1 - q_N$

- $\mathfrak{M}_3$: **Signal-Detection State**

  $\mathrm{P}(\textit{Signal Stimuli}) = p_S$

$$\mathrm{P}(\textit{Noise Stimuli}) = 0$$

Conditional independence enforces the same state-response mapping to all stimuli that happen to be in the same mental state $\mathfrak{M}$, irrespective of the stimulus class they belong to. For instance, the same guessing probabilities g and $1 - g$ apply to signal and noise stimuli that reach the uncertainty state $\mathfrak{M}_2$ (see Figure 13).

Conditional independence also establishes that a state's response mapping is unaffected by experimental manipulations that vary the probability of that state being reached by stimuli of a given class (for discussions, see Kellen & Klauer, 2018; Rouder & Morey, 2009; Rouder et al., 2014). For example, consider an experimental design that includes a *study-strength manipulation* that is applied to the class of studied items (e.g., some items are studied once, others thrice), such that we can refer to the subclasses of *weak* and *strong* studied items.[17] The response mapping associated with each mental state will be the same for both subclasses of studied items, a situation that implies a number of *equality constraints* at the level of response probabilities. For instance, incorrect "`new`" responses can only result from the mapping of the uncertainty state $\mathfrak{M}_2$ (see the top tree in Figure 13). It follows that the probability distribution of such responses is *the same* for weak and strong studied items (for tests, see Chen, Starns, & Rotello, 2015; Kellen & Klauer, 2015; McAdoo, Key, & Gronlund, 2018; Province & Rouder, 2012).

In recent years, many studies have pitted the High-Threshold model against non-threshold SDT models (e.g., Bröder & Schütz, 2009; Chen et al., 2015; Dube & Rotello, 2012; Kellen & Klauer, 2014, 2015; Kellen et al., 2013; Province & Rouder, 2012; Starns, Dubé, & Frelinger, 2018). One limitation found in these model comparisons is that their implications do not generalize beyond this specific threshold model. For instance, Kellen and Klauer (2014) reported a critical test that rejected the High-Threshold Model (for a replication, see McAdoo & Gronlund, 2016). However, the

[17] Importantly, the different subclasses introduced by the study-strength manipulation (*weak* vs. *strong*) are not to be correlated with external (e.g., word color, font), internal (e.g., word frequency, category membership), or contextual features (e.g., experimental block). Otherwise, conditional independence cannot be assumed (e.g., one can guess differently for green/weak and red/strong words). For a similar point, see Kellen and Klauer (2015).

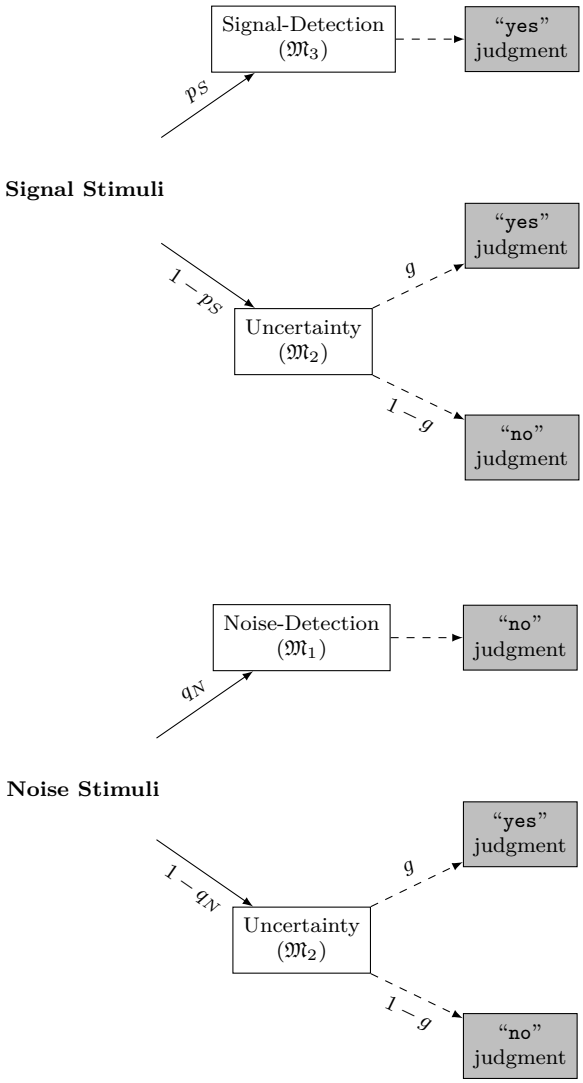

Figure 13. Processing tree representation of the High-Threshold Model for the yes-no task. White squares correspond to the latent mental states, whereas gray squares correspond to the observed responses. State-response mappings are represented by the dashed branches. The parameters associated with each branch correspond to its respective probability (when omitted, the probability of a branch is 1). Parameters: p_S = Probability that a signal stimulus is signal-detected. q_N = Probability that a noise stimulus is noise-detected. g = Probability of guessing "`yes`", conditional on the stimulus being in the uncertainty state.

results of their critical tests can be successfully accommodated by the *Low-Threshold Model* originally proposed by Luce (1963; see also Kellen, Erdfelder, Malmberg, Dubé, & Criss, 2016; McAdoo & Gronlund, 2020; Starns et al., 2018). One distinctive property of this Low-Threshold model is that assumes that noise stimuli can be incorrectly signal-detected. It should be noted that an extension of the Low-Threshold model has been proposed by Krantz (1969). This extended model includes a *super-signal-detection* state that can only be reached by signal stimuli.

The limited scope of the existing critical tests for threshold representations is

unsatisfactory. Ideally, one would be able to directly test properties associated with a large family of threshold models, some of them more complex than the High- and Low-Threshold models discussed so far. In the section below, we achieve this desideratum by specifying a novel recognition-memory paradigm that allows for the direct testing of a *Generalized Threshold Model* (GTM), a model that includes all of its predecessors as special cases.

A Generalized Threshold Model and a Critical Test

The GTM assumes that a test item can be in one of five mutually-exclusive mental states $\mathfrak{M}$. These states can be reached by signal and noise stimuli with some probability, which is determined by the occurrence or non-occurrence of a number of detection processes (see Figure 14). These processes can either lead to an item being detected as studied ('*signal-detection*' or '*super-signal-detection*' states) or non-studied ('*noise-detection*' or '*super-noise-detection*' states). Some states can be reached by both signal and noise stimuli, whereas others are exclusive to one of the stimulus classes. In the following, we list the mental states included in the GTM and their respective probabilities (see also Figure 14). We also reference any previous threshold model that has postulated such state [in brackets]:

- $\mathfrak{M}_1$: **Super-Noise-Detection State**

  $\mathrm{P}(\textit{Signal Stimuli}) = 0$

  $\mathrm{P}(\textit{Noise Stimuli}) = (1 - q_S) \times q_N \times q_N^{\star}$

- $\mathfrak{M}_2$: **Noise-Detection State**

  $\mathrm{P}(\textit{Signal Stimuli}) = (1 - p_S) \times p_N$

  $\mathrm{P}(\textit{Noise Stimuli}) = (1 - q_S) \times q_N \times (1 - q_N^{\star})$ [High-Threshold Model]

- $\mathfrak{M}_3$: **Uncertainty State**

  $\mathrm{P}(\textit{Signal Stimuli}) = (1 - p_S) \times (1 - p_N)$ [all threshold models]

  $\mathrm{P}(\textit{Noise Stimuli}) = (1 - q_S) \times (1 - q_N)$ [all threshold models]

- $\mathfrak{M}_4$ **Signal-Detection State**

$\mathrm{P}(\textit{Signal Stimuli}) = p_S \times (1 - p_S^*)$ [all threshold models]

$\mathrm{P}(\textit{Noise Stimuli}) = q_S$ [Low-Threshold Model]

- $\mathfrak{M}_5'$: **Super-Signal-Detection State**

  $\mathrm{P}(\textit{Signal Stimuli}) = p_S \times p_S^*$ [Low-Threshold Model]

  $\mathrm{P}(\textit{Noise Stimuli}) = 0$

Note that in order to ensure the prediction of above-chance performance in recognition-memory tasks, one must assume that the probability of correctly detecting the class of a stimulus is always greater than the probability of an *incorrect* detection (i.e., $q_S \leq p_S$ and $p_N \leq q_N$).

The GTM includes all the threshold models previously discussed as special cases. For example, when $p_S^* = p_N = q_S = q_N^* = 0$, the GTM reduces to the High-Threshold Model illustrated in Figure 13, a model that has been at the center of many recent discussions (e.g., Bröder & Schutz, 2009; Dube & Rotello, 2012; Kellen & Klauer, 2014, 2015; Province & Rouder, 2012). Alternatively, the constraint $p_N = q_N = q_N^* = 0$ reduces the GTM to Krantz' extended Low-Threshold Model (Krantz, 1969). However, note that the GTM goes beyond any of these models. For instance, it allows signal stimuli to be incorrectly noise-detected (for a theoretical motivation, see Moran, 2016).

As in all preceding threshold models, conditional independence is assumed to hold. In the specific case of the GTM, conditional independence also imposes a constraint on p_N, the conditional probability of a studied item reaching the noise-detection state. As argued by Moran (2016), the assumption in High- and Low-Threshold models that individuals cannot actively reject studied items is nothing more than a convenient idealization. For example, consider a scenario in which a participant was often distracted during the study phase. There is no reason to assume that the studied items that the participant missed cannot be subjected to the same kind of meta-cognitive inferences assumed to underlie the noise-detection of non-studied items (for discussions, see Bröder & Schütz, 2009: Klauer & Kellen, 2010).[18] Because such inferences are

[18] The GTM nevertheless includes state $\mathfrak{M}_1'$, which is exclusive to noise stimuli. This inclusion is motivated by the existence of certain kinds of active rejection that could only apply to non-studied

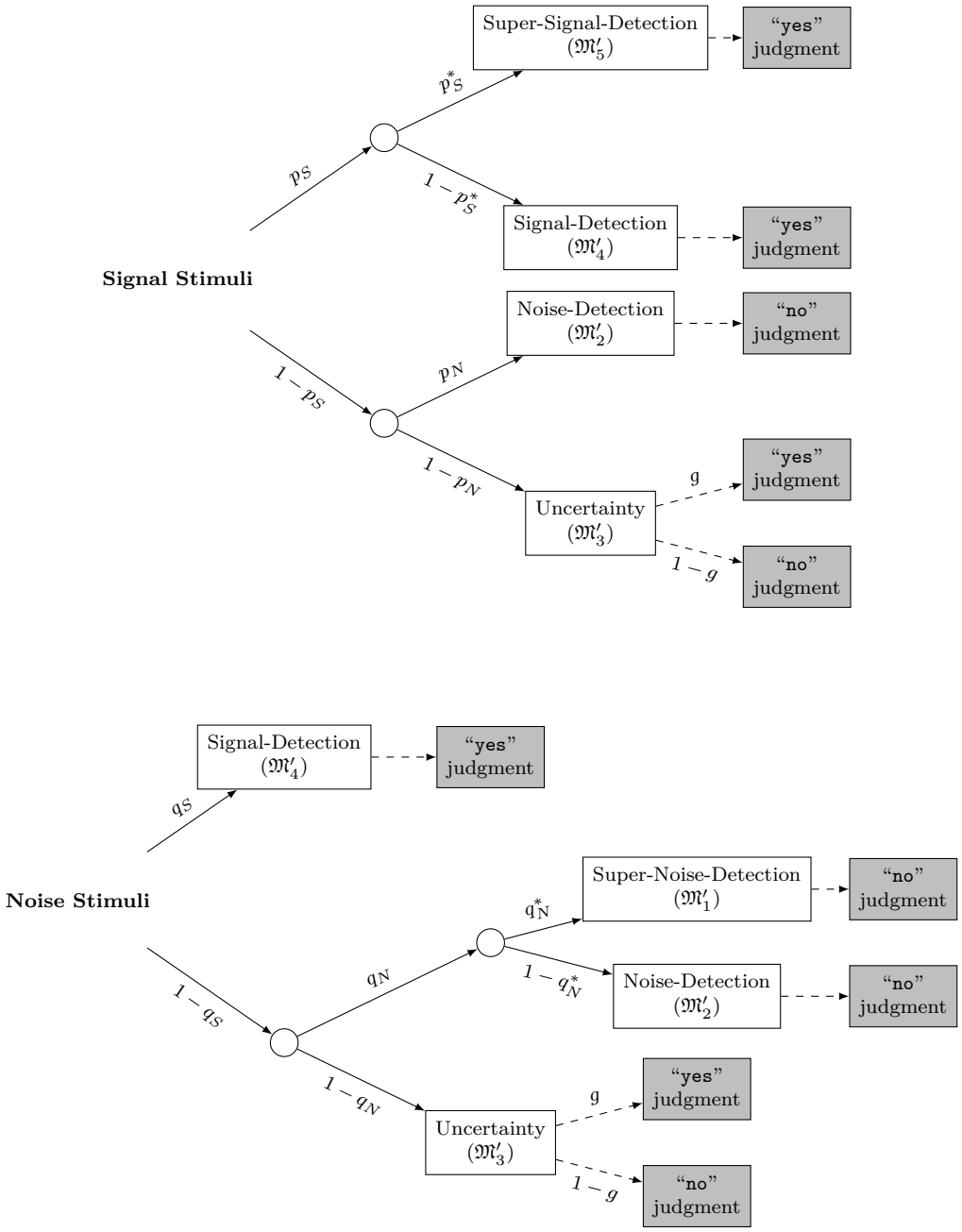

Figure 14. Processing tree representation of the Generalized Threshold Model (GTM) for the yes-no task. White squares correspond to the latent mental states, whereas gray squares correspond to the observed responses. State-response mappings are represented by the dashed branches. The parameters associated with each branch correspond to its respective probability (when omitted, the probability of a branch is 1). Parameters: p_S/q_S = Probability that a signal/noise stimulus is signal-detected. p_N/q_N = Probability that a signal/noise stimulus is noise-detected, conditional on not being signal-detected. p^*_S/q^*_N = Probability that a signal/noise stimulus is super-signal/noise-detected, given that it was signal/noise-detected. g = Probability of guessing "`yes`", conditional on the item being in the uncertainty state.

predicated on the absence of any kind of remembering (e.g., "*I would have remembered it, if I had seen it*"; see Strack & Bless, 1994), it follows that the probability of a

---

items (see Gallo, 2006, Chap. 5).

studied item reaching the noise-detection state $\mathfrak{M}'_2$ is conditionally independent from the probability of successfully remembering studied items; i.e., p_N is unaffected by the experimental manipulation of p_S and $p^{\star}_S$ (but see Footnote 17 for boundary conditions).

When dealing with test trials involving multiple alternatives (as done in the task detailed below), the mental-state probabilities associated with each of the alternatives are assumed to be independent (i.e., latent-variable independence holds). For example, if a signal stimulus and a noise stimulus are presented together in a 2-AFC trial, the probability of the former being in state $\mathfrak{M}'_4$ *and* the latter being in state $\mathfrak{M}'_2$ corresponds to $\overbrace{p_S \times p^*_S}^{\text{signal stimulus}} \times \overbrace{q_N \times (1 - q^*_N)}^{\text{noise stimulus}}$. The assumption of latent-variable independence has a precedent in previous work involving High- and Low-Threshold models (e.g., Luce, 1963; Province & Rouder, 2012).

Multiple-alternative subsetting task. In order to pit the GTM against non-threshold SDT models of recognition memory, we developed a novel multiple-alternative subsetting task (for a similar task in preferential choice, see Regenwetter, Marley, & Joe, 1998). At each *subsetting trial*, participants are shown five items. They are informed that the total number of studied items can range between one and four, and that this number varies randomly across subsetting trials. Participants are requested to select which items they believe to have been previously studied (i.e., select a subset). The total number of items selected per trial can range between one and four. The left panel of Figure 15 provides an illustration.

Some of the subsetting trials were immediately succeeded by a *followup 2-AFC trial*. Participants were informed that these followup 2-AFC trials would occur randomly, and that their occurrence was unrelated to the accuracy of their responses in the immediately-preceding subsetting trial. The two alternatives presented in these followup 2-AFC trials were always two items that received the *same* judgment (i.e., they were both judged to be studied or non-studied) in the immediately preceding subsetting trial. The participant's task here is to choose the item that they are more willing to reverse their previous judgment (e.g., which one would they rather judge to be studied; see the right panel of Figure 15).

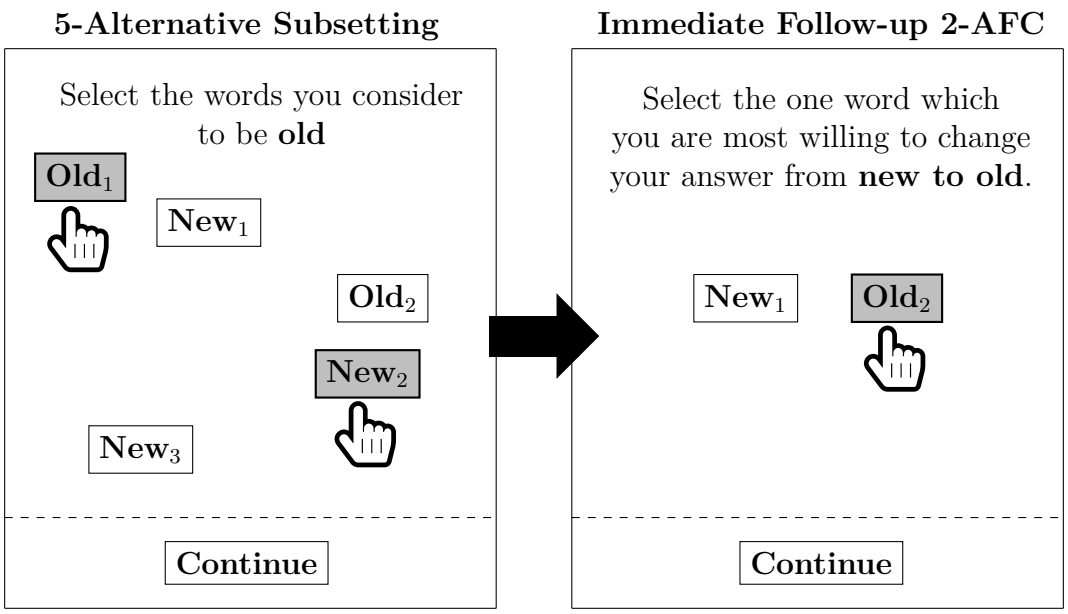

Figure 15. The two types of trials used in the experiment. The grey shades symbolize the selected words (i.e., the words judged as 'old'). In the subsetting task, the horizontal displacement was randomly determined for each word on each trial anew. Note that the follow-up trial on the right is an example of a new-to-old trial, which occurred probabilistically.

GTM Predictions. According to the GTM, the unrecognized test items presented in the followup 2-AFC trials are either in the uncertainty state $\mathfrak{M}'_3$ or being actively rejected (i.e., states $\mathfrak{M}'_1$ and $\mathfrak{M}'_2$). When confronted with two unrecognized items, a decision is assumed to be made based on their respective states: If the two items are in different mental states, then the one with the largest index will be chosen (e.g., Luce, 1963; Province & Rouder, 2012). For example, if one item was detected as new ($\mathfrak{M}'_1$ or $\mathfrak{M}'_2$) whereas the other one was in uncertainty $\mathfrak{M}'_3$, then the latter item would be chosen. Alternatively, if two items are in the same mental state, then one of them will be chosen with probability $\frac{1}{2}$ (for a discussion on the comparison of items based on their mental states, see Erdfelder, Küpper-Tetzel, & Mattern, 2011).

When the pair presented in the followup 2-AFC trial consists of an unrecognized studied item and an unrecognized new item – a $\langle S, N \rangle$ pair – the probability of a

correct response (i.e., the studied item being selected) is given by:[19]

$$P_S^{\langle S,N\rangle} = \frac{(1-p_S)(1-q_S)[q_N q_N^* + p_N q_N(1-q_N^*)\frac{1}{2} + (1-p_N)q_N(1-q_N^*) + (1-p_N)(1-q_N)\frac{1}{2}]}{(1-p_S)(1-q_S)}. \tag{24}$$

The products $(1-p_S)(1-q_S)$ in the numerator and denominator cancel out. Also, conditional independence establishes that the probabilities of noise-detection processes (given by parameters p_N, q_N, and $q_N^\star$) are unaffected by study-strength manipulations. Together, these constraints enforce the prediction that the probability of a correct response in a followup $\langle S, N\rangle$ pair *does not depend* on the signal-detection probabilities. In context of the study-strength manipulation referred to earlier, this means that the probability of an unrecognized studied item being chosen in a followup 2-AFC trial is predicted to be the same for ⟨weak-studied, new⟩ and ⟨strong-studied, new⟩ pairs. Henceforth, we will denote these pairs of unrecognized items by $\langle WS, N\rangle$ and $\langle SS, N\rangle$ pairs, respectively. A similar argument can be used to show that the probability of choosing an unrecognized weak item over an unrecognized strong item in a $\langle WS, SS\rangle$ pair; i.e., $P_{WS}^{\langle WS,SS\rangle}$, is predicted to be $\frac{1}{2}$. These predictions concerning unrecognized pairs are summarized by the following hypothesis, which we denote by $\mathcal{H}_{\text{GTM}}$:

$$\mathcal{H}_{\text{GTM}} : P_{WS}^{\langle WS,N\rangle} = P_{SS}^{\langle SS,N\rangle}, \;\; P_{WS}^{\langle WS,SS\rangle} = \frac{1}{2}.$$

This hypothesis is in line with other recent tests conducted on threshold models, which establish equality and/or chance-level constraints at the level of choice probabilities (see Malejka & Bröder, 2016; McAdoo, 2019; Starns et al., 2018; Voormann, Rothe-Wulff, Starns, & Klauer, 2020).

'Non-Threshold' SDT Predictions. According to SDT, the inclusion of items into a subset will depend on their latent-strength values. We will assume that these judgments are based on a comparison of the test items' latent-strength values with a response criterion κ, such that any test item whose latent-strength value surpasses it is

[19] So far, we used the superscript $\langle m\rangle$ to denote the number of alternatives in a given trial. In the present context, the actual stimulus-class composition of alternatives matters, which led us to include this information in the superscript.

judged to be studied (i.e., included in the subset). Latent-variable independence is assumed to hold.[20] Now, let us consider the case in which the decision maker is presented with an unrecognized $\langle S, N\rangle$ pair in a followup 2-AFC trial. Given the way items that are assumed to be judged in the subsetting task, it follows that the latent-strength values of both unrecognized items are below the response criterion κ. Conditional on this inequality, the probability of selecting the studied item over the non-studied item in this pair is given by the probability that the latent-strength value of the former is greater than the latter's. Formally:

$$P_S^{\langle S,N\rangle} = \frac{\int_l^\kappa f_S(t)F_N(t)\ dt}{F_S(\kappa)F_N(\kappa)}. \tag{25}$$

for $l \leq \kappa \leq u$, with l and u denoting the lower and upper bounds of the latent-strength scale. The reason why we are not fixing these bounds to 0 and 1 respectively is that the formal result discussed below will *not* rely on the universal representation used so far. This change is motivated by the specific approach used here to contrast models assuming threshold and non-threshold representations: We will formally establish a property at the level of the latent distributions that – whenever present – implies a violation of conditional independence (i.e., implies a non-threshold representation) at the level of the followup 2-AFC choice probabilities. We will then show that this property holds across a wide range of parametric distributions.

Theorem. *Let $F_N(t)$ and $F_\mu(t)$, for $t \in (l, u)$, be the CDFs of the noise and signal distributions, the latter being parametrized by parameter μ. Moreover, assume that F_μ is differentiable in μ for every admissible t. Finally, let $H_\mu(t) = \frac{\frac{\partial}{\partial\mu}F_\mu(t)}{F_\mu(t)}$. If $H_\mu(t)$ is monotonically increasing in t for all μ, then $P_S^{\langle S,N\rangle}$ is monotonically increasing in μ.*

Proof. First, let us express $P_S^{\langle S,N\rangle}$ as function of μ, using a more convenient formulation:

$$P_S^{\langle S,N\rangle}(\mu) = 1 - \frac{\int_l^\kappa F_\mu(t)f_N(t)\ \mathrm{d}t}{F_\mu(\kappa)F_N(\kappa)}.$$

[20] We note that a formally equivalent model has been previously proposed in the domain of social-choice modeling (see Marley, 1993; Regenwetter et al., 1998).

If $P_S^{\langle S,N\rangle}(\mu)$ is monotonically increasing, then the fraction on the second term, which we will denote here as $g(\mu)$, is monotonically decreasing (i.e., its derivative with respect to μ is negative). Rearranging terms and differentiating under the integral sign:

$$\frac{\partial}{\partial\mu}g(\mu) = \int_l^\kappa \frac{f_N(t)}{F_N(\kappa)F_\mu(\kappa)^2} \cdot \left(\frac{\partial}{\partial\mu}F_\mu(t)F_\mu(\kappa) - \frac{\partial}{\partial\mu}F_\mu(\kappa)F_\mu(t)\right)\, \mathrm{d}t.$$

The first multiplicative term can be ignored, as it is always positive. In order for the second term to be negative, the following inequality needs to hold

$$\frac{\frac{\partial}{\partial\mu}F_\mu(t)}{F_\mu(t)} \leq \frac{\frac{\partial}{\partial\mu}F_\mu(\kappa)}{F_\mu(\kappa)},$$

which corresponds to

$$H_\mu(t) \leq H_\mu(\kappa)$$

Given that $t \leq \kappa$, this inequality can only hold if H_μ is monotonically increasing. □

What the theorem above shows is that, if H_μ is monotonically increasing, then $P_S^{\langle S,N\rangle}$ should increase as a function of parameter μ. With this formal result as a backdrop, let us consider the predicted choice probabilities for unrecognized $\langle SS, N\rangle$ and $\langle WS, N\rangle$ pairs: The study-strength manipulation is expected to make strong studied items more memorable than than their weak counterparts. In terms of the latent-strength distributions of studied items and their respective parameters μ, this manipulation means that $\mu_{WS} \leq \mu_{SS}$, which leads us the following hypothesis, which we denote by $\mathcal{H}_{\text{SDT}}$:[21]

$$\mathcal{H}_{\text{SDT}} : P_{WS}^{\langle WS,N\rangle} \leq P_{SS}^{\langle SS,N\rangle},\;\; P_{WS}^{\langle WS,SS\rangle} \leq \frac{1}{2}$$

But how general is this hypothesis? Quite so, actually. It turns out that function H_μ is monotonically increasing for a large family of parametric distributions. For

[21] By replacing the noise distribution with the weak-signal distribution, the theorem above can be used to show that $P_{WS}^{\langle WS,SS\rangle}$ decreases along with the memory-strength μ of the strong-signal distribution. The upper boundary of this probability is $\frac{1}{2}$, which corresponds to the boundary case in which there is a null study-strength effect and the latent variables associated with weak and strong studied words are identical (i.e., $\mu_{WS} = \mu_{SS}$).

instance, take the case of the family of shift distributions $f_\mu(t) = f(t-\mu)$, which includes many well-known distributions such as the Gaussian, logistic, exponential, ex-Gaussian, Weibull, Gumbel, etc. As shown by Chechile (2011), the members of this family have monotonically decreasing *reverse-hazard* functions $r(t) = \frac{f_\mu(t)}{F_\mu(t)}$. It is easy to see that $H_\mu(t) = -r(t)$ for members of this family of distributions. However, note that the monotonicity of H_μ is not limited to shift-type distributions, as it can be shown to hold in cases such as the Gamma distribution (see Kellen & Klauer, 2014, 2015).

There is one notable exception though: $\mathcal{H}_{\text{SDT}}$ can be violated by a Gaussian SDT model that allows σ_S^2 to differ between weak and strong studied items. However, we do not see this situation as problematic. A closer inspection of this specific SDT model shows that such violations *have nothing to do* with this model's merits, such as its ability to describe asymmetric, concave ROCs, or the relationship between study-strength manipulations and ROC asymmetry (Heathcote, 2003; Lockhart & Murdock, 1970; Ratcliff, McKoon, & Tindall, 1994). Instead, violations of $\mathcal{H}_{\text{SDT}}$ are made possible by the model's unfortunate ability to make pathological predictions, such as below-chance accuracy or partially convex ROCs (all of which follow from violations of likelihood-ratio monotonicity; see DeCarlo, 2002; Kellen & Klauer, 2011). In any case, this exception does not put into question the test results reported below.

Hypothesis $\mathcal{H}_{\text{GTM}}$ is included within $\mathcal{H}_{\text{SDT}}$ as a boundary case. This relationship reflects Rouder et al.'s (2014) argument that threshold models are nothing more than a sub-family of SDT models for which conditional independence holds. We also see the same kind of theoretical characterizations found in previous critical tests (Kellen & Klauer, 2014, 2015): Conditional independence imposes the prediction that the 'magnitude' of errors is invariant to the probability of said errors being committed. Even though recognition-failure is greater for weak studied items than for strong studied items, the probability of correcting such failures in a followup 2-AFC trial is the same for both stimulus subclasses. In contrast, non-threshold representations that violate conditional independence predict that the probability and magnitude of errors go hand in hand, such that the probability of correcting a failed recognition in a followup 2-AFC

trial is expected to be greater for strong studied items than weak studied items.

Under any realistic experimental design, the number of followup 2-AFC trials that one could ever hope to collect per participant is going to be extremely small. This unfortunate reality severely compromises the ability to test $\mathcal{H}_{\text{GTM}}$ and $\mathcal{H}_{\text{SDT}}$ at the individual level. However, these hypotheses can nevertheless be tested using aggregate data, as neither of them can be spuriously rejected when aggregating individual data that are in conformity with them. One fortunate consequence of this robustness to aggregation is that it also applies to the aggregation of heterogeneous data coming from the same individual respondents. As discussed in Footnote 2, one could in principle relax the assumption in SDT that the response criterion κ is fixed across test trials (e.g., Kellen et al., 2012; Rosner & Kochanski, 2009). The aggregation of responses from followup 2-AFC trials, when the latter were obtained under different κ (i.e., in the presence of criterion noise), will not spuriously reject $\mathcal{H}_{\text{SDT}}$ the same way that aggregating responses from heterogeneous respondents would not.

Experiment 4: Testing the Generalized Threshold Model

Participants, Materials, and Procedure

Four-hundred and one participants took part in this study online. Participants were recruited through `Prolific` (www.prolific.co), and received a fixed £1.5 reward in exchange for their participation. The experiment took roughly 12 minutes to complete.

After instructions were given, the experiment began with a study phase in which participants were presented a list of 70 common words. Half of the words were presented once (weak words), whereas the other half was presented three times (strong words). Words were presented in random order with a minimal distance of five words between the repetitions of strong words. Each word was presented for 2000 ms, with a 400 ms interval between words. A primacy/recency buffer of five words was presented at the beginning and end of the study phase. These buffer words were not tested. After the study phase was completed, participants initiated the test phase, which was comprised of 28 subsetting trials and a variable amount of followup 2-AFC trials. The

total number of follow-up trials varied across participants as it depended on chance.

Subsetting trials consisted of five words, each presented on a separate row with a random horizontal displacement to it (see Figure 15, Left Panel). The reason behind the use of a jittered display was to help distinguishing the different words clearly and prevent them from being perceived as grouped units. Participants were asked to select between one and four words, which they judged to be old. This request was made along with the information that each subsetting trial included between one and four studied words, with the exact composition being randomly determined across trials.

After each subsetting trial, there was a chance that a followup 2-AFC trial would take place: If two or more words were not recognized, then a followup 2-AFC trial comprised of unrecognized words would take place with probability .40. If only one word was not recognized, then a followup 2-AFC trial comprised of recognized words would take place with probability .40. In either case, the words included in a followup trial were randomly selected among the ones available. In each of these followup 2-AFC trials, participants were requested to select the word for which they were more willing to reverse their judgment (for an illustration, see Figure 15, Right Panel). Participants had to select one of the words in order to continue with the experiment.

After completing the test phase, participants filled in a short demographic survey, were thanked, and received their monetary reward. The demographic survey included a question in which participants could state that they did not take the experiment seriously, and that we should better not analyze their data (without having negative consequences for their pay). Data from participants who responded affirmatively were discarded, which left us with a total of 395 participants. We also excluded data from thirty-nine individuals whose observed performance in the subsetting trials was not consistent with the experimental manipulation of study strength (i.e., failed to show a better performance for the strong words relative to the weak words in the subsetting trials; for a similar exclusion criterion, see Kellen & Klauer, 2015), leaving us with data from a total of 356 participants.

Results and Discussion

Among the retained participants, the average recognition rates in the subsetting trials were 19%, 52%, and 77% for new, weak, and strong words, respectively. These differences show a clear study-strength effect, which is necessary to ensure a diagnostic comparison of followup 2-AFC trials (if there is no study-strength effect, then $\mathcal{H}_{\text{SDT}}$ reduces to $\mathcal{H}_{\text{GTM}}$).

On average, the retained participants engaged in 8.20 followup 2-AFC trials, 5.50 of which involved unrecognized word pairs. Among the latter, an average of 1.10 trials were $\langle WS, SS\rangle$ pairs, 1.20 $\langle SS, N\rangle$ pairs, and 3.20 $\langle WS, N\rangle$ pairs. In these follow-up 2-AFC trials, unrecognized weak words were selected over unrecognized new words 60% of the times ($P_{WS}^{\langle WS,N\rangle}$; 95% CI = [.57,.63]), whereas unrecognized strong words were selected over unrecognized new words 68% of the times ($P_{SS}^{\langle SS,N\rangle}$; 95% CI = [.63,.72]). Moreover, unrecognized weak words were selected over unrecognized strong words only 44% of the times ($P_{WS}^{\langle WS,SS\rangle}$; 95% CI = [.39,.49]). In order to quantify the discrepancy between the observed 2-AFC choices and $\mathcal{H}_{\text{GTM}}$, we fitted them with a joint binomial model which assumed that $P_{WS}^{\langle WS,N\rangle} = P_{SS}^{\langle SS,N\rangle}$ and $P_{WS}^{\langle WS,SS\rangle} = \frac{1}{2}$. This model grossly misfitted the data ($G^2 = 13.74$, $p = .0004$). In contrast, a joint binomial model implementing $\mathcal{H}_{\text{SDT}}$ provided a perfect fit ($G^2 = 0$, $p = 1$).

Overall, the 2-AFC choices from Experiment 4 suggest that errors are less egregious in conditions where they are less frequent. These results violate of the conditional-independence assumption that unites the family of threshold models encompassed by the GTM, and are consistent with a non-threshold representation.

General Discussion

We began the present investigation by distinguishing theories from the families of models that instantiate them. This distinction highlighted the fact that models incorporate a number of choices with respect to different possible properties that can be assumed to hold. Testing any given model corresponds to an omnibus test of the set of assumptions defining it. This means that if that model fails to fit the data, it may not

be apparent which of the postulated properties are violated, and hence which alternative assumptions could have been made. With this distinction between theories and models as a backdrop, we pursued the strategy of identifying a set of properties that are relevant to a large family of SDT models of recognition memory. The choices with respect to each property partitions the set of possible models into different sub-families. By identifying and testing each relevant property, it is possible to find evidence either for or against each sub-family of models. The critical tests conducted here serve both to support the general framework of SDT as applied to recognition memory and to identify sub-families of models that are more strongly supported by the data. We considered five different properties which we discuss in turn.

As our first step, we focused on the existence of a random-scale representation which is fundamental to all SDT models considered in the literature. A random-scale representation implies that choice alternatives (i.e., a signal or noise item) are characterized by a joint distribution in latent variable space. One alternative to this picture is that the latent-strength distributions are context-dependent. In other words, the joint distribution changes as a function of the composition of the set of alternatives under consideration (e.g., Trueblood et al., 2013), a situation that would undermine the viability of the kind of SDT modeling that is typically found in the literature. We tested the hypothesis that a random-scale representation is possible by assessing whether the data from two m-alternative forced choice experiments satisfy the system of Block-Marschak inequalities (Block & Marschak, 1960; Falmagne, 1978). The forced-choice data from both Experiment 1 and Experiment 2 were found to be consistent with these inequalities, supporting (at least in this instance) the existence of a random-scale representation.

In our second step, we directed our attention to the hypothesis that the latent-strength distributions being postulated are independent. Latent-variable independence is a property that underwrites many extant models and enables a set of crisp and powerful relationships between m-AFC tasks, ranking tasks, and the yes-no ROC function. If this property does not hold, then more complex models would be

required to characterize the data. We applied a direct test of latent-variable independence to the forced-choice data from Experiments 1 and 2. In Experiment 2, we also invoked the aforementioned relationships between the different tasks and the yes-no ROC function, and established an additional predictive test. Overall, the results were consistent with the assumption that recognition judgments can be successfully described by a SDT model assuming independent latent-strength variables.

In our third step, we were concerned with the requirement that the likelihood ratio of latent signal and noise distributions is monotonically increasing. This property, likelihood-ratio monotonicity, which is often assumed but rarely tested, is formally equivalent to stating that the yes-no ROC function is concave. We showed that it is possible to derive a series of implications from likelihood-ratio monotonicity, which take the form of a set of inequality constraints at the level of signal-ranking or forced-choice probabilities. More specifically, ROCs are concave if and only if the signal-ranking probabilities are ordered, such that more egregious errors are less probable than more moderate errors. The data from both Experiments 1 and 2 were found to be in agreement with the constraints imposed by likelihood-ratio monotonicity.

Our fourth step concerned the symmetry of the yes-no ROC function. It has been generally agreed that ROCs in recognition memory are positively asymmetric. However, this finding might be due to the way latent-strength values are mapped onto confidence values (e.g., Kellen et al., 2012; Klauer & Kellen, 2010) and/or to a misinterpretation of the effect of response-bias manipulations (e.g., Van Zandt, 2000). We constructed a critical test of ROC symmetry that sidesteps all of these issues. The test is based on the comparison of a traditional m-AFC task with a $m^{\star}$-AFC task. Data from Experiment 3 implementing this critical test were found to be at odds with the hypothesis that the yes-no ROC function is symmetric. Instead, the data were consistent with the hypothesis that yes-no ROC functions are positively asymmetric.

Our fifth and final step addressed the way in which the latent-strength distributions are affected by study-strength manipulations. We contrasted a large sub-family of threshold models with an alternative sub-family of non-threshold (often

described as continuous) models. We showed that the conditional-independence property underlying threshold representations implies a number of equality constraints at the level of response probabilities. Using a subsetting task in Experiment 4, we were able to establish a critical test which yielded results that were found to be inconsistent with a threshold representation.

The SDT modeling approach taken here is so far not a part of most researchers' toolboxes. We will therefore dedicate the remainder of the general discussion to two things: First, we will address certain kinds of skepticism regarding the value of critical-test approaches. Second, we will delineate potential future directions and the different ways in which the present work may contribute to lines of research where SDT plays a major role.

Meta-Theoretical Clarifications

One possible reaction to the arguments motivating the critical-test approach taken here is that they overlook the fact that many researchers are interested in models *as a whole*. After all, models – not the theories from which they descend – are what ultimately comes into contact with data, characterizing them in terms of a number of well-defined conceptual components (e.g., parameters that modulate a number of postulated processes or representations). Therefore, when someone argues that a model $\mathcal{M}_A$ outperforms model $\mathcal{M}_B$, that is *all* that is being said, with little or no interest in the theoretical propositions from which these models stem. When taking this "view from the trenches", the concerns we raise may appear somewhat esoteric and out of touch with a perfectly consistent way of engaging with models in psychology. Moreover, one could also argue that critical tests often resort to experimental designs that differ markedly from the ones that are otherwise adopted. Therefore, one would answer negatively the question of whether or not the failure of model $\mathcal{M}_A$ under experimental design $\mathcal{E}_1$ should in any way detract from its successes under experimental design $\mathcal{E}_2$.

These criticisms are predicated on a number of misconceptions that are important to dispel. The best place to begin is the agreement among philosophers and scientists at

large that the term "modeling" encompasses many different '*systems of practice*' that cannot be reduced to a monolithic set of principles, goals, and criteria (see Bailer-Jones, 2009; Chang, 2012). Although this acknowledgement loudly echoes Feyerabend's (1975) famous maxim that *anything goes*, we are always required to make a number of meta-theoretical commitments. These commitments will allow us to adjudicate what can be meaningfully stated (see Maraun, 1998). The commitment to consider models "as a whole" works to prevent us from scrutinizing the specific elements of each model that drive their respective successes or failures. We can say that model $\mathcal{M}_A$ outperforms model $\mathcal{M}_B$ according to some penalized-fit statistic, but we are unable to directly inquire about which specific properties of each model are driving their performance.[22] Indeed, one can establish a system of practice that only engages with models as a whole. But this would be an unnecessarily impoverished one, as it would force us to remain silent in cases where much could be said.

In reaction, one could argue that we are resorting to a straw-man argument, in the sense that nobody would really defend an embargo on dissecting models. But if that is the case, then the justification for the above-described reluctance towards the critical-test approach is undermined. For it is unclear what kind of intellectual gerrymandering could sustain a system of practices that would permit a more careful scrutiny of models while at the same time dismissing or downplaying the present focus on specific properties that demarcate sub-families of models. The importance of engaging in critical testing becomes obvious as soon as we begin disentangling the different properties of a model and trying to figure out which ones are doing the leg work.

Lastly, let us turn to the use of different experimental designs and their inferential value. Under specific systems of practice, such as the ones found in *cognitive psychometrics* (see Batchelder & Alexander, 2013), it is deemed perfectly reasonable to continuously apply a model $\mathcal{M}_A$ to data coming out of experimental design $\mathcal{E}_2$ even

[22] More formally: Let $\mathcal{M}$ be a model that is defined by a conjunction of postulates $\mathcal{P}_i, \ldots, \mathcal{P}_N$, and let $\mathcal{C}$ be an observational consequence of $\mathcal{M}$. Empirical verification of $\mathcal{C}$ can affect our credence on $\mathcal{M}$ as a whole, however these effects do not necessarily extend to each of the postulates that constitute it, only to those that contribute to the derivation of $\mathcal{C}$ (for discussions, see Rozeboom, 1970, 2008).

though its failures with data coming from experimental design $\mathcal{E}_1$ are well documented. The reason being that the cognitive-psychometric goal is *not* to test theories but to establish *measurement apparatuses* with desirable properties that trump their known shortcomings. For instance, consider the recent applications of streamlined evidence-accumulation models (e.g., van Ravenzwaaij, Donkin, & Vandekerckhove, 2017): Even though these models make predictions that have long been falsified, they nevertheless provide a convenient way to characterize differences across groups and experimental conditions. Outside of cognitive-psychometric system of practice, it is unclear to us how one could ever justify bestowing a specific experimental design with some kind of privileged status (beyond experimental knowledge; see Mayo, 1996), especially if one's interests lie in the theoretical propositions that underlie the set of candidate models: On one hand, such a privileged status would effectively discourage the derivation of novel observable consequences from theories, a fundamental step in every scientific domain that we can think of. On the other, it would assist in the development of self-reinforcing relations between models and experiments that ultimately inhibit the possibility of 'paradigm-shifting' developments (see Hacking, 1992).

Boundary Conditions

The present work focused on recognition memory, an area where SDT plays a major role. An obvious question is whether the same type of critical tests could be applied in other domains, such as visual working memory (Cowan, 2001; Donkin, Tran, & Nosofsky, 2014; van den Berg, Shin, Chou, George, & Ma, 2012). The answer is: *it depends*. In the case of visual working memory, it is well established that the fidelity of stimulus representations is negatively affected by increases in stimulus set-size (e.g., van den Berg et al., 2012). This result indicates that one cannot describe people's judgments across set sizes using the same joint latent distribution, which means that the Block-Marschak inequalities are not expected to hold. However, one could still test the Block-Marschak inequalities (and perhaps reconstruct yes-no ROCs) in designs

where the set size is fixed (e.g., Donkin et al., 2014).

Testing SDT in More Complex Designs

The critical tests discussed here only considered two classes of stimuli – signal and noise. Future work should also consider more complex designs in which multiple classes of stimuli, such as different types of non-studied items, are considered. Note that these 'enriched scenarios' are already covered in the original formulation of the Block-Marschak inequalities (see (9)). For example, a more complex design would allow us to test Malmberg's (2008) conjecture that retrieval processes are adjusted to the composition of test items, with more recollection-type processes being involved when most of the non-studied items are extremely familiar and/or similar to the studied items (see also Heathcote, Raymond, & Dunn, 2006). Such designs could also be used in more applied domains such as eyewitness identification, where it has been shown that the introduction of certain types of alternatives (e.g., decoys similar to a suspect) can affect performance in rather nuanced ways (see Wixted, Vul, Mickes, & Wilson, 2018).

Further critical testing of SDT can also be achieved by revisiting previously-published studies. For example, Wixted (1992) tested whether 2-AFC choice probabilities for a number of different stimulus-class pairings satisfy *strong stochastic transitivity* (Luce & Suppes, 1965): Let X, Y, and Z denote three distinct stimulus classes. Choice probabilities satisfy strong-stochastic transitivity if and only if

$$P_X^{\langle X,Y\rangle}, P_Y^{\langle Y,Z\rangle} \geq \frac{1}{2} \text{ implies that}$$
$$P_X^{\langle X,Z\rangle} \geq \max\left(P_X^{\langle X,Y\rangle}, P_Y^{\langle Y,Z\rangle}\right).$$

Using six different stimulus classes, Wixted (1992) found the data to be consistent with strong-stochastic transitivity. Hintzmann, Curran, and Caulton (1995) reported similar results with an experimental design involving a total of eight stimulus classes.

A random-scale representation implies a property known as the *triangle inequality*

(Niederée & Heyer, 1997), which is satisfied if and only if

$$P_X^{\langle X,Y\rangle} + P_Y^{\langle Y,Z\rangle} - P_X^{\langle X,Z\rangle} \leq 1.$$

Strong-stochastic transitivity implies the triangle inequality. When dealing with five stimulus classes or less, the triangle inequality being satisfied is both a necessary and sufficient condition for existence of a random-scale representation. But this is no longer the case when dealing with more than five stimulus classes (for an overview, see Regenwetter, Dana, & Davis-Stober, 2010). This means that the tests conducted by both Wixted (1992) and Hintzmann et al. (1995) did not include an exhaustive evaluation of all of the constraints implied by a random-scale representation. They also did not directly test whether latent-variable independence is violated (Suck, 2002; see also McCausland & Marley, 2013, 2014). Until recently, a reanalysis of these previously-published data would have been unfeasible – but recent algorithmic developments have made the challenge much more tractable (see Smeulders, Davis-Stober, Regenwetter, & Spieksma, 2019). Future research efforts should be placed on revisiting these studies.

The Usefulness of Ranking Judgments

Although we did not collect ranking judgments in the present work, researchers should not overlook the possibility of reconstructing ROC functions from them. Especially in areas of research where the predominance of confidence-rating ROCs might raise some concerns. For instance, Rotello et al. (2015) used confidence-rating ROC data to assess the performance of maltreatment referrals for black and white children. Instead of confidence ratings, one could reconstruct ROCs based on ranking judgments ('*please order these cases according to their likelihood of being cases of maltreatment*'). Among other things, one could use such an approach to evaluate performance in the absence of racial information and/or whether performance is affected by the separate/joint ranking of black and white children.

Ranking judgments can also play an important role in the study of eyewitness

identification (e.g., Wixted et al., 2018). Typical paradigms have focused on single choices that participants may or may not make. Requesting participants to rank alternatives, even when they do not believe that a suspect is among them, can provide additional information that is valuable for the theoretical characterization of eyewitness judgments (see Brewer, Weber, & Guerin, 2020; Carlson et al., 2019).

Confidence-Rating Judgments

Most of the ROCs reported in the literature are based on confidence-rating judgments (for reviews, see Wixted, 2007; Yonelinas & Parks, 2007). The fact that our results yield ROCs that are concave and asymmetric is consistent with the characteristics of ROCs obtained with confidence ratings. However, it would be unwise to interpret our results as legitimizing the general use of confidence ratings to estimate yes-no ROCs. Our reluctance comes from the fact that confidence judgments often do not behave as expected and/or can change the phenomena being studied (see Benjamin et al., 2013; Brainerd, Nakamura, Reyna, & Holliday, 2017; Kellen & Klauer, 2015; Miyoshi, Kuwahara, & Kawaguchi, 2018). Also relevant is the way in which confidence judgments are requested (e.g., one-step versus two-step procedure; see Moran, Teodorescu, & Usher, 2015; Stephens, Dunn & Hayes, 2019). On a more technical side, the modeling of confidence-rating judgments often requires auxiliary assumptions that can affect results (e.g., how response criteria can segment regions of a latent-strength continuum; see Moran & Goshen-Gottstein, 2015). Given these issues, we think that further work is necessary to better understand the agreement between performance as described by SDT and confidence-rating ROCs across a wide range of conditions.

Final Remarks

The present work demonstrated how we can isolate some of the properties underlying SDT models and directly evaluate them using critical tests. But beyond providing an empirical foundation for SDT modeling in recognition memory, the present work demonstrates the value of critical-test approaches. One of the merits of critical testing is its ability to establish a clear relationship between theoretical statements and

data. This contrasts with the model-fit comparisons that are typically found in the literature at large, in which the failure of a given model (rather than a family of models) is often discussed without an explicit reference to the theoretical properties found to be at odds with the data. Our hope is that the present work encourages researchers to expand their toolboxes and consider going beyond traditional model-comparison practices.

References

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 157–188). Elsevier.

Bailer-Jones, D. M. (2009). *Scientific models in philosophy of science*. University of Pittsburgh Press.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181.

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). *Psychological Bulletin*, *139*, 1204–1212. https://doi.org/10.1037/a0033894

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1601–1608.

Bernardo, J. M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, *4*, 111–122.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*, 463–501. https://doi.org/10.1037/0033-295x.115.2.463

Birnbaum, M. H. (2011a). Testing theories of risky decision making via critical tests. *Frontiers in Psychology*, *2*, 315.

Birnbaum, M. H. (2011b). Testing. mixture models of transitive preference: Comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*, 675–683.

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of response. In I. Olkin, S. Ghurye, W. Hoeffding, M. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 97–132). Stanford University Press.

Brainerd, C., Nakamura, K., Reyna, V., & Holliday, R. (2017). Overdistribution illusions: Categorical judgments produce them, confidence ratings reduce them. *Journal of Experimental Psychology: General*, *146*, 20–40.

Brewer, N., Weber, N., & Guerin, N. (2020). Police lineups of the future? *American Psychologist*, *75*, 76–91.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear - or are they? on premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*, 587–606.

Carlson, C. A., Jones, A. R., Whittington, J. E., Lockamyeir, R. F., Carlson, M. A., & Wooten, A. R. (2019). Lineup fairness: Propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive research: principles and implications*, *4*(1), 2.

Chechile, R. A. (2011). Properties of reverse hazard functions. *Journal of Mathematical Psychology*, *55*, 203–222. https://doi.org/10.1016/j.jmp.2011.03.001

Chechile, R. A., Sloboda, L. N., & Chamberland, J. R. (2012). Obtaining separate measures for implicit and explicit memory. *Journal of Mathematical Psychology*, *56*, 35–53. https://doi.org/10.1016/j.jmp.2012.01.002

Chen, T., Starns, J. J., & Rotello, C. M. (2015). A violation of the conditional independence assumption in the two-high-threshold model of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1215–1222.

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

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*, 316–326.

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the

subjective likelihood model (SLiM). *Journal of Memory and Language*, *55*, 447–460.

Davis-Stober, C. P. (2009). Multinomial models under linear inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710–721.

Dede, A. J. O., Squire, L. R., & Wixted, J. T. (2014). A novel approach to an old problem: Analysis of systematic errors in two models of recognition memory. *Neuropsychologia*, *54*, 51–56.

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

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

Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton University Press.

Dunn, J. C., & Anderson, L. (2018). Signed difference analysis: Testing for structure under monotonicity. *Journal of Mathematical Psychology*, *85*, 36–54.

Dunn, J. C., & Kalish, M. L. (2018). *State-trace analysis*. Springer.

Dunn, J. C., & Rao, L.-L. (2019). Models of risky choice: A state-trace and signed difference analysis. *Journal of Mathematical Psychology*, *90*, 61–75.

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note*.

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, *6*, 7–22.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

Falmagne, J. C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, *18*, 52–72.

Falmagne, J.-C. (1985). *Elements of psychophysical theory*. Oxford University Press.

Feller, W. (1971). *An introduction to Probability Theory and its applications (Vol. II).* John Wiley & Sons.

Feyerabend, P. K. (1975). *Against method*. Verso.

Fiorini, S. (2004). A short proof of a theorem of Falmagne. *Journal of Mathematical Psychology*, *48*, 80–82.

Frigg, R., & Hartmann, S. (2018). Models in science. *Stanford Encyclopedia of Philosophy*.

Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. Psychology Press.

Garcia-Marques, L., Garcia-Marques, T., & Brauer, M. (2014). Buy three but get only two: The smallest effect in a 2× 2 anova is always uninterpretable. *Psychonomic bulletin & review*, *21*, 1415–1430.

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*, 431–455.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, *66*, 228–234.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Grünbaum, B. (2003). *Convex polytopes*. Springer.

Habel, K., Grasman, R., Gramacy, R. B., Stahel, A., & Sterratt, D. C. (2015). *Geometry: Mesh generation and surface tesselation* [R package version 0.3-6]. https://CRAN.R-project.org/package=geometry

Hacking, I. (1992). The self-vindication of the laboratory sciences. In A. Pickering (Ed.), *Science as practice and culture* (pp. 29–64). University of Chicago Press.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1210–1230. https://doi.org/10.1037/0278-7393.29.6.1210

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory & Language*, *55*, 495–514.

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for bayesian inference. *Journal of mathematical psychology*, *91*, 70–87.

Hintzman, D. L., Curran, T., & Caulton, D. A. (1995). Scaling the episodic familiarities of pictures and words. *Psychological Science*, *6*(5), 308–313.

Iverson, G. J., & Bamber, D. (1997). The generalized area theorem in signal detection theory. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 301–318). Lawrence Erlbaum Associates.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*, 1–32.

Kalish, M. L., Dunn, J. C., Burdakov, O. P., & Sysoev, O. (2016). A statistical test of the equality of latent orders. *Journal of Mathematical Psychology*, *70*, 1–11.

Karabatsos, G. (2005). The exchangeable multinomial model as an approach to testing deterministic axioms of choice and measurement. *Journal of Mathematical Psychology*, *49*, 51–69.

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, *2*, 160–165.

Kellen, D., Erdfelder, E., Malmberg, K. J., Dubé, C., & Criss, A. H. (2016). The ignored alternative: An application of Luce's low-threshold model to recognition memory. *Journal of Mathematical Psychology*, *75*, 86–95.

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology*, *55*, 251–266.

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1795–1804.

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, *122*, 542–557.

Kellen, D., & Klauer, K. C. (2018). Elementary signal detection and threshold theory. In E. J. Wagenmakers (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive neuroscience (4th edition, vol. v)*. Wiley.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, *20*, 693–719.

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

Kellen, D., & Singmann, H. (2016). ROC residuals in signal-detection models of recognition memory. *Psychonomic Bulletin & Review*, *23*, 253–264.

Killeen, P. R., & Taylor, T. J. (2004). Symmetric receiver operating characteristics. *Journal of Mathematical Psychology*, *48*, 432–434.

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source memory: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*, 465–478. https://doi.org/10.3758/PBR.17.4.465

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review*, *76*, 308–324.

Krantz, D. H., Luce, R. D., Suppes, P., & Tverksy, A. (1971). *Foundations of measurement (Vol. I)*. Academic Press.

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

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319.

Lu, Z.-L., & Dosher, B. A. (2008). Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review*, *115*, 44–82.

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*, 61–79.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*, 215–233.

Luce, R. D. (2010). Behavioral assumptions for a class of utility models: A program of experiments. *Journal of Risk and Uncertainty*, *41*, 19–37.

Luce, R. D., & Suppes, P. (1965). Preference, utility and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology (vol. iii)* (pp. 249–410). Wiley.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)* Erlbaum.

Malejka, S., & Bröder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & Cognition*, *44*, 63–72.

Malejka, S., & Bröder, A. (2019). Exploring the shape of signal-detection distributions in individual recognition roc data. *Journal of Memory and Language*, *104*, 83–107.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384.

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology*, *8*, 435–461.

Marley, A. A. J., & Regenwetter, M. (2017). Choice, preference, and utility: Probabilistic and deterministic representations. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New Handbook of Mathematical Psychology (Vol. I)* (pp. 374–453). Cambridge University Press.

Marley, A. (1990). A historical and contemporary perspective on random scale representations of choice probabilities and reaction times in the context of Cohen and Falmagne's (1990, Journal of Mathematical Psychology, 34) results. *Journal of Mathematical Psychology*, *34*, 81–87.

Marley, A. (1993). Aggregation theorems and the combination of probabilistic rank orders. In M. S. Fligner & J. S. Verducci (Eds.), *Probability models and statistical analyses for ranking data* (pp. 216–240). Springer.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

McAdoo, R. M. (2019). Recognition memory is fundamentally continuous, and strategic discretization does not change this. *Unpublished Doctoral Dissertation*.

McAdoo, R. M., & Gronlund, S. D. (2016). Relative judgment theory and the mediation of facial recognition: Implications for theories of eyewitness identification. *Cognitive Research: Principles and Implications*, *1*, 11.

McAdoo, R. M., & Gronlund, S. D. (2020). Theoretical note: Exploring luce's (1963) low-threshold model applied to recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 247–256.

McAdoo, R. M., Key, K. N., & Gronlund, S. D. (2018). Stimulus effects and the mediation of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1814–1823.

McCausland, W. J., Davis-Stober, C., Marley, A. A. J., Park, S., & Brown, N. (2020). Testing the Random Utility Hypothesis Directly Testing Random Utility Directly. *The Economic Journal*, *130*(625), 183–207.

McCausland, W. J., & Marley, A. A. J. (2014). Bayesian inference and model comparison for random choice structures. *Journal of Mathematical Psychology*, *62*, 33–46.

McCausland, W. J., & Marley, A. (2013). Prior distributions for random choice structures. *Journal of Mathematical Psychology*, *57*, 78–93.

Miyoshi, K., Kuwahara, A., & Kawaguchi, J. (2018). Comparing the confidence calculation rules for forced-choice recognition memory: A winner-takes-all rule wins. *Journal of Memory and Language*, *102*, 142–154.

Moran, R. (2016). Thou shalt identify! the identifiability of two high-threshold models in confidence-rating recognition (and super-recognition) paradigms. *Journal of Mathematical Psychology*, *73*, 1–11.

Moran, R., & Goshen-Gottstein, Y. (2015). Old processes, new perspectives: Familiarity is correlated with (not independent of) recollection and is more (not equally) variable for targets than for lures. *Cognitive Psychology*, *79*, 40–67.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147.

Morgan, M. S., & Morrison, M. (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge University Press.

Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage, and retrieval of item information. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 145–194). Erlbaum.

Niederée, R., & Heyer, D. (1997). Generalized random utility models and the representational theory of measurement: A conceptual link. In A. A. J. Marley (Ed.), *Choice, Decision and Measurement: Essays in Honor of R. Duncan Luce* (pp. 155–189). Lawrence Erlbaum.

O'Connor, A. R., Guhl, E. N., Cox, J. C., & Dobbins, I. G. (2011). Some memories are odder than others: Judgments of episodic oddity violate known decision rules. *Journal of Memory and Language*, *64*, 299–315.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*, 260–311.

Parks, C. M., Murray, L. J., Elfman, K., & Yonelinas, A. P. (2011). Variations in recollection: The effects of complexity on source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 861–873.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347–353.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 224–232.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences USA*, *109*, 14357–14362.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 763–785.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42–56.

Regenwetter, M., Dana, J., Davis-Stober, C. P., & Guo, Y. (2011). Parsimonious testing of transitive or intransitive preferences: Reply to Birnbaum (2011). *Psychological Review*, *118*, 684–688.

Regenwetter, M., Marley, A. A. J., & Joe, H. (1998). Random utility threshold models of subset choice. *Australian Journal of Psychology*, *50*, 175–185.

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

Regenwetter, M., & Cavagnaro, D. R. (2019). Tutorial on removing the shackles of regression analysis: How to stay true to your theory of binary response probabilities. *Psychological methods*, *24*(2), 135–152.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Frontiers in psychology*, *1*, 148.

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

Robere, R. (2018). *Vertexenum: Vertex enumeration of polytopes* [R package version 1.0.2].

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Rosner, B. S., & Kochanski, G. (2009). The law of categorical judgment (corrected) and the interpretation of changes in psychophysical performance. *Psychological Review*, *116*, 116–128.

Rotello, C. M. (2018). Signal detection theories of recognition memory. In J. T. Wixted (Ed.), *Learning and Memory: A Comprehensive Reference, 2nd edition (Vol. 4: Cognitive Psychology of Memory)*. Elsevier.

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

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

Rouder, J. N., Province, J. M., Swagman, A. R., & Thiele, J. E. (2014). From ROC curves to psychological theory. *Manuscript submitted for publication*.

Rouder, J., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review*, *116*, 655–660.

Rozeboom, W. W. (1970). The Art of Metascience, or, What Should a Psychological Theory Be? In J. Royce (Ed.), *Toward Unification in Psychology* (pp. 53–164). University of Toronto Press.

Rozeboom, W. W. (2008). The problematic importance of hypotheses. *Journal of Clinical Psychology*, *64*, 1109–1127.

Sattath, S., & Tversky, A. (1976). Unite and conquer: A multiplicative inequality for choice probabilities. *Econometrica*, *44*, 79–89.

Shaw, M. L. (1980). Identifying attentional and decision-making components in information processing. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 277–296). Erlbaum.

Smeulders, B., Davis-Stober, C., Regenwetter, M., & Spieksma, F. C. (2018). Testing probabilistic models of choice using column generation. *Computers & operations research*, *95*, 32–43.

Spektor, M. S., Kellen, D., & Hotaling, J. M. (2018). When the good looks bad: An experimental exploration of the repulsion effect. *Psychological Science*, *29*, 1309–1320.

Starns, J. J., 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, J. J., Dubé, C., & Frelinger, M. E. (2018). The speed of memory errors shows the influence of misleading information: Testing the diffusion model and discrete-state models. *Cognitive Psychology*, *102*, 21–40. https://doi.org/https://doi.org/10.1016/j.cogpsych.2018.01.001

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Steingrimsson, R. (2016). Subjective intensity: Behavioral laws, numerical representations, and behavioral predictions in Luce's model of global psychophysics. *Journal of Mathematical Psychology*, *75*, 205–217.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2019). Belief bias is response bias: Evidence from a two-step signal detection model. *Journal of experimental psychology: learning, memory, and cognition*, *45*, 320–332.

Stephens, R. G., Matzke, D., & Hayes, B. K. (2019). Disappearing dissociations in experimental psychology: Using state-trace analysis to test for multiple processes. *Journal of Mathematical Psychology*, *90*, 3–22.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language*, *33*, 203–217. https://doi.org/10.1006/jmla.1994.1010

Suck, R. (2002). Independent random utility representations. *Mathematical Social Sciences*, *43*, 371–389.

Suppes, P., Krantz, D. H., Luce, R. D., & Tverksy, A. (1989). *Foundations of measurement (Vol. II)*. Academic Press.

Suppes, P. C. (2002). *Representation and invariance of scientific structures*. CSLI Publications.

Swets, J. A. (1959). Indices of signal detectability obtained with various psychophysical procedures. *The Journal of the Acoustical Society of America*, *31*, 511–513.

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

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Torgerson, W. S. (1958). *Theory and methods of scaling.* Wiley.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*(4), 321–359.

Trippas, D., Kellen, D., Singmann, H., Pennycook, G., Koehler, D. J., Fugelsang, J. A., & Dubé, C. (2018). Characterizing belief bias in syllogistic reasoning: A hierarchical-bayesian meta-analysis of roc data. *Psychonomic Bulletin & Review*, *25*, 2141–2174.

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision making. *Psychological Science*, *24*, 901–908.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. https://doi.org/10.1007/BF00122574

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological review*, *121*(1), 124–149.

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*, 8780–8785.

van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in sem models. *Structural Equation Modeling*, *17*(3), 443–463. https://doi.org/10.1080/10705511.2010.489010

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

van Fraassen, B. (1980). *The scientific image*. Oxford University Press.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The ez diffusion model provides a powerful test of simple empirical effects. *Psychonomic bulletin & review*, *24*(2), 547–556.

Voormann, A., Rothe-Wulf, A., Starns, J. J., & Klauer, K. C. (2020). Does speed of recognition predict two-alternative forced-choice performance? replicating and extending starns, dubé, and frelinger (2018). *Quarterly Journal of Experimental Psychology*.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after loftus. *Memory & cognition*, *40*(2), 145–160.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford University Press.

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

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

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive psychology*, *105*, 81–114.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 681–690.

Yellott, J. I. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, *15*, 109–144.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*, 747–763.

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

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*, 203–212.