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 of Zurich

Author Note

David Kellen, Samuel Winiger, and Henrik Singmann were supported by SNSF grant 100014_165591. John Dunn was supported by ARC grant DP130101535. The authors thank Rani Moran for valuable comments on an earlier draft. Data and R scripts can be found at the Open Science Framework: REDACTED

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 notions are often misunderstood, especially when its comes to its testability. The present work clarifies these misconceptions in the context of recognition memory, and shows that a general class of SDT models can be strictly tested through a set of behavioral constraints known as the *Block-Marschak inequalities*. We also discuss the connection between yes-no, forced-choice, and ranking judgments. This connection introduces additional behavioral constraints and provides a non-parametric way to reconstruct yes-no Receiver Operating Characteristic (ROC) functions. Two recognition-memory experiments tested different sets of constraints on forced-choice judgments, and used them to reconstruct yes-no ROCs. All constraints were generally supported by the data. The reconstructed ROCs were found to be concave and asymmetric. The plausibility of these ROCs was supported by a direct comparison with actual yes-no judgments, and by an independent, non-parametric test of ROC asymmetry (Experiment 3). 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

Signal Detection Theory (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). The theory characterizes a detection or identification task in which the decision maker must classify stimuli as belonging to one or more classes (e.g., respond "yes" when a visual signal is presented, or say "old" when encountering a previously-studied word). The challenge is that detection of any stimulus class always involves some degree of uncertainty or confusability due to the presence of many different forms of noise (e.g., Lu & Dosher, 2008; Kellen, Klauer, & Singmann, 2012). According to SDT, the decision maker (human or non-human) decides on the basis of sampled values coming from distributions established on a latent metric space. In a simple scenario in which the decision maker encounters a single stimulus and attempts to discriminate between two stimulus classes – *signal* (S) and *noise* (N) – the distributions are on a unidimensional latent strength scale. In order to determine a response, SDT assumes that each sampled strength value ϵ is compared to a pre-established response criterion κ. A "signal" or "yes" response is produced when $\epsilon \geq \kappa$, otherwise a "noise" or "no" response is given. The degree of overlap between distributions reflects the uncertainty associated with the different stimulus classes. Through this characterization, SDT is able to determine the decision maker's ability to discriminate between stimulus classes as well her response biases. The left panel of Figure 1 illustrates this model under the assumption that the signal and noise distributions are Gaussian. The right panel illustrates the predicted relationship between the probabilities of "yes" responses for signal and noise stimuli (referred to as hit and false-alarm probabilities) as one varies the response criterion κ. This relationship is commonly referred to as the yes-no *receiver operating characteristic* (ROC) curve.

The popularity of SDT can be attributed to two factors, namely its *empirical success* and its *theoretical inclusiveness*: First, SDT has been able to successfully characterize judgments across a wide variety of psychological domains (e.g., perception, memory, reasoning; see Green & Swets, 1966; Macmillan & Creelman, 2005; Rotello,

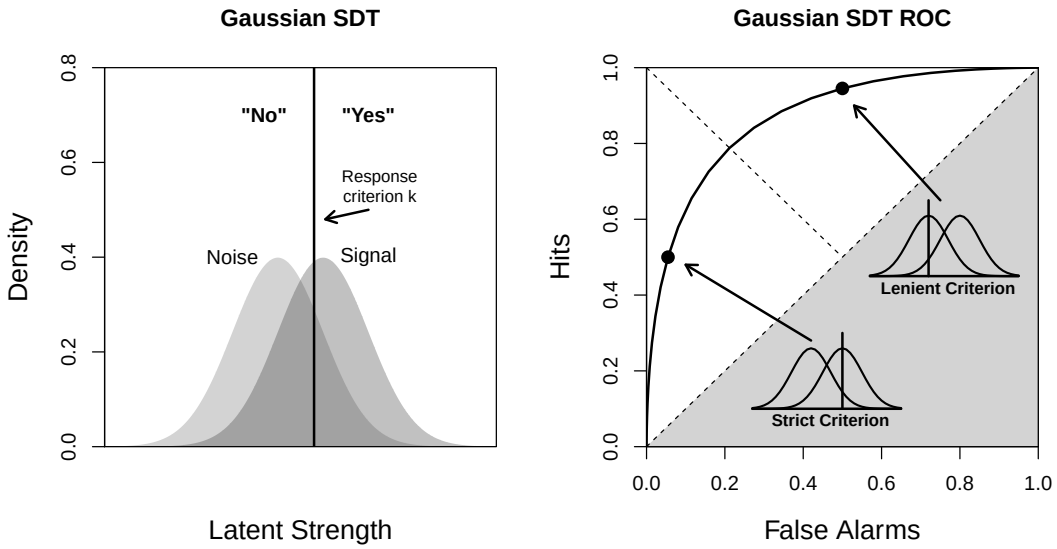

Figure 1. *Left Panel*: Illustration of the Gaussian SDT model Latent-strength values above the response criterion result in a "yes" response, values below a "no" response. *Right Panel*: The models' respective ROC, along with two ROC points associated with different response criteria (one strict, the other lenient). Dashed lines delimiting the gray areas indicate chance-level performance.

2018; Trippas et al., in press). Second, the SDT 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). Both factors have contributed to researchers taking the SDT assumptions for granted rather than attempting to test them directly. Also, some of the attempts to test SDT have hinged on the misunderstanding of the theoretical status of certain auxiliary assumptions, such as the commonly-held assumption that the latent distributions are Gaussian. Although the parametric representation provided in Figure 1 is by far the most popular one, there is no intrinsic link between SDT and the Gaussian assumption (e.g., see Green & Swets, 1966, p. 58). The goal of the present work is to clarify these misunderstandings and to provide a behavioral foundation for SDT modeling that is stripped from any strong parametric assumptions.

The approach taken here follows a long-standing tradition of directly testing the properties of latent representations through behavioral patterns that can be formally shown to be intrinsically connected to them (e.g., Bamber, 1979; Dunn & Kalish, 2018; Falmagne, 1985; Karabatsos, 2005; Krantz, Luce, Suppes, & Tversky, 1971; Luce, 2010;

Luce & Tukey, 1964; Steingrimsson, 2016; Suppes, Krantz, Luce, & Tversky, 1989). The most famous example of such an approach is Luce and Tukey's (1964) work on *additive conjoint measurement*, showing that the hypothesis that some observable attribute constitutes a measure (i.e., can be represented as an additive structure on the set of real numbers) can be evaluated through different cancellation tests in a factorial design. A more recent example would be Regenwetter, Dana, and Davis-Stober's (2011) testing of the axiom of transitivity, which is a core assumption in the vast majority of models of decision making. In the present case, we will test a set of behavioral constraints that SDT needs to comply with, irrespective of parametric assumptions.

The remainder of this paper is organized as follows: First, we will discuss how the shape of any given ROC cannot be used to test SDT or to motivate the development of more complex accounts of recognition memory that postulate *additional* processes. We will then show how SDT in its general form can be tested through the compliance of forced-choice judgments with the *Block-Marschak inequalities* (Block & Marschak, 1960; Falmagne, 1978). We will also show how forced-choice judgments can be used to reconstruct the yes-no ROC function without the need to collect yes-no judgments. We will then report three experiments testing SDT. The first experiment tests the compliance of forced-choice recognition-memory judgments with the aforementioned Block-Marschak inequalities and uses these judgments to reconstruct the associated yes-no ROC. The second experiment replicates the previous one and tests the empirical accuracy of the reconstructed yes-no ROC. A third experiment implements a direct test on the asymmetry of the yes-no ROC. Altogether, the results provide direct empirical support for the SDT assumption that choices are based on an evaluation of latent-strength values, and provide independent corroboration of the notion that yes-no ROCs in recognition memory are concave and asymmetric. Finally, we will discuss some of the theoretical implications that follow from our empirical results and how our approach provides new opportunities for SDT modeling.

ROC Shape Does Not Provide a Testbed for the General Class of SDT Models, Only Members of It

The interplay between a decision-maker's latent-strength distributions and her response biases can be captured with an ROC function, which we will denote here by ρ. This function describes how the hit probability (H) changes along with the false-alarm probability (FA). It is necessarily monotonically increasing, as any increase in FA cannot lead to a decrease in H and any increase in H cannot lead to a decrease in FA. As previously mentioned, according to SDT the ROC function describes exactly how both response probabilities are affected by the variation of the response criterion κ, from strict to lenient (e.g., right panel of Figure 1). In the domain of recognition memory, which will be the focus of the present work, ROCs are often used 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 rationale behind these comparisons can vary: In some cases, the motivation is not purely theoretical but rather a more practical or technical dissatisfaction with the Gaussian model. For instance, the mean and variance parameters of the signal distribution tend to be positively correlated, a situation that has led many to argue one would be better off assuming latent distributions in which means and variances are governed by the same parameter (e.g., DeCarlo, 1998; Lockhart & Murdock, 1970; Rouder, Province, Swagman, and Thiele, 2014). As discussed by Green and Swets (1966) in their seminal monograph, many alternative distributions could have been adopted, distributions that yield a wide variety of plausible ROC functions (e.g., Killeen & Taylor, 2004; Rouder, Pratte, & Morey, 2010).

More theoretical motivations can be found in the discussion of *dual-process accounts* that go *above and beyond* the latent-strength-based judgments postulated by SDT and introduce the possibility of recognition via *episodic recollection* (for reviews, see Wixted, 2007; Yonelinas & Parks, 2007). One problem with these discussions is that

they often confound the SDT assumption of latent-strength-based judgments with the Gaussian assumption and the ROCs that it can yield. This confound has led to discussions on the preponderance/nature of different processes that are entirely based on how curved an ROC is (e.g., Parks, Murray, Elfman, & Yonelinas, 2011), or attempts to dismiss a dual-process account through the investigation of model-fit residuals (e.g., Dede, Squire, & Wixted, 2014). None of these analyses are actually tapping into the core theoretical assumptions of the theories (e.g., one versus two processes). Instead, their results exclusively hinge on the auxiliary parametric assumptions that are required to implement them. The reality is that, when considering the SDT model in a more general form, stripped from any parametric assumptions, the model *cannot* be tested on the basis of ROC shape alone (e.g., Green & Swets, 1966; Iverson & Bamber, 1997; Killeen & Taylor, 2004; Rouder et al., 2014).

The inability to test SDT based on ROC shape can be demonstrated as follows: First, we define both noise and signal distributions with respective densities f_N and f_S on the [0,1] unit interval, and corresponding cumulative distribution functions F_N and F_S. Also, let F^{-1} denote the inverse of the cumulative distribution function. Without loss of generality, we assume that the noise distribution is *uniform* on the [0,1] interval. Thus, $f_N(\kappa) = 1$ and $F_N(\kappa) = \kappa$ for all $\kappa \in [0, 1]$. This specification of the latent distributions on the unit interval, referred to by Rouder et al. (2014) as a *universal representation*, is not 'standard' in the sense that latent variables tend to be represented on the entire real line rather than on the unit interval. However, it will prove useful when demonstrating some of the theoretical results that we discuss. On this representation, the hit and false-alarm probabilities are:

$$\text{FA} = P(\epsilon_N \geq \kappa) = 1 - F_N(\kappa) = 1 - \kappa, \tag{1}$$

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

Note that we can express the hit probability as a function of the false-alarm probability, $H = 1 - F_S(F_N^{-1}(1 - \text{FA}))$. The ROC function is thus:

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

Because $F_N^{-1}(1 - \text{FA}) = 1 - \text{FA} = \kappa$ under the universal representation, the ROC function takes on a simpler form:

$$\rho(\text{FA}) = 1 - F_S(1 - \text{FA}). \tag{4}$$

The only constraint being imposed here is that F_S has to be a cumulative distribution function defined on [0,1]. This means that there is always some cumulative distribution that can perfectly accommodate any ROC function.

The shape of the ROC is given by its slope. Taking derivatives of this function, we have:

$$\rho'(\text{FA}) = \frac{f_S(F_N^{-1}(1 - \text{FA}))}{f_N(F_N^{-1}(1 - \text{FA}))} = \frac{f_S(\kappa)}{f_N(\kappa)} = f_S(\kappa). \tag{5}$$

This shows that the slope of the ROC at any point is given by the density f_S. It follows that the distribution that perfectly describes the ROC function is the one with a density f_S that perfectly matches the ROC slope. It is reasonable to expect that the likelihood ratio $\frac{f_S(\kappa)}{f_N(\kappa)} = f_S(\kappa)$ monotonically increases with κ, reflecting the notion that signal items are more likely to take on larger strength values than noise items (Criss & McClelland, 2006; Glanzer, Hilford, & Maloney, 2008; Osth & Dennis, 2015). This *monotonic likelihood* assumption (e.g., Zhang & Mueller, 2005) implies that the ROC takes on a concave shape (i.e., it has monotonically decreasing slope; note that this includes linear ROCs).[1]

The existence of a distribution for any ROC data completely undermines the attempt to use them to test SDT in its general form. ROC data is therefore only

[1] Because f_S is always non-negative, the ROC has to be monotonically increasing, an assumption expressed at the beginning of this section. At this point we do not know of any circumstances where non-monotonic or monotonically decreasing ROCs would be plausible (e.g., DeCarlo, 2013).

relevant when evaluating specific parametric forms of the signal and noise distributions (e.g., setting them to be Gaussian). Figure 2 illustrates the traditional and universal SDT representations of two models that have often been pitted against each other, the Gaussian SDT model and the high-threshold model (Bröder & Schütz, 2009; Dube & Rotello, 2012; Province & Rouder, 2012).[2] Similar illustrations could be made for other popular models (DeCarlo, 2002; Yonelinas, 1997). The recasting of the high-threshold model as an SDT model is far from novel (see Swets, 1986), but researchers often ignore its implications. Rather than testing a fundamentally distinct model, one is testing a special type of SDT model that includes properties such as *conditional independence* (Kellen & Klauer, 2014, 2015; Province & Rouder, 2012; Rouder & Morey, 2009; Rouder et al., 2014). This notion is reflected in Kellen and Klauer's critical tests of conditional independence, in which the predictions of the high-threshold model corresponded to the null hypothesis, nested within the predictions of the class of SDT models.

A General Test for SDT Based on the Block-Marschak Inequalities

The inability to test SDT via the shape of ROCs has led some to argue that SDT should be seen as a general framework in which specific models and hypotheses can be tested, as in the case of conditional independence (Chen, Starns, & Rotello, 2015; Kellen & Klauer, 2014, 2015; Kellen, Singmann, Vogt, & Klauer, 2015; Province & Rouder, 2012; McAdoo & Gronlund, 2016; McAdoo, Key, & Gronlund, in press; Swagman, Province, & Rouder, 2015; for a theoretical overview, see Rouder et al., 2014). One advantage of this stance towards SDT is that it places the focus on the testing of behavioral constraints and on the construction of a behavioral corpus that any candidate theory needs to accommodate. But on the other hand, this stance overstates the inability to test SDT, overlooking other types of data such as judgments

[2] 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.

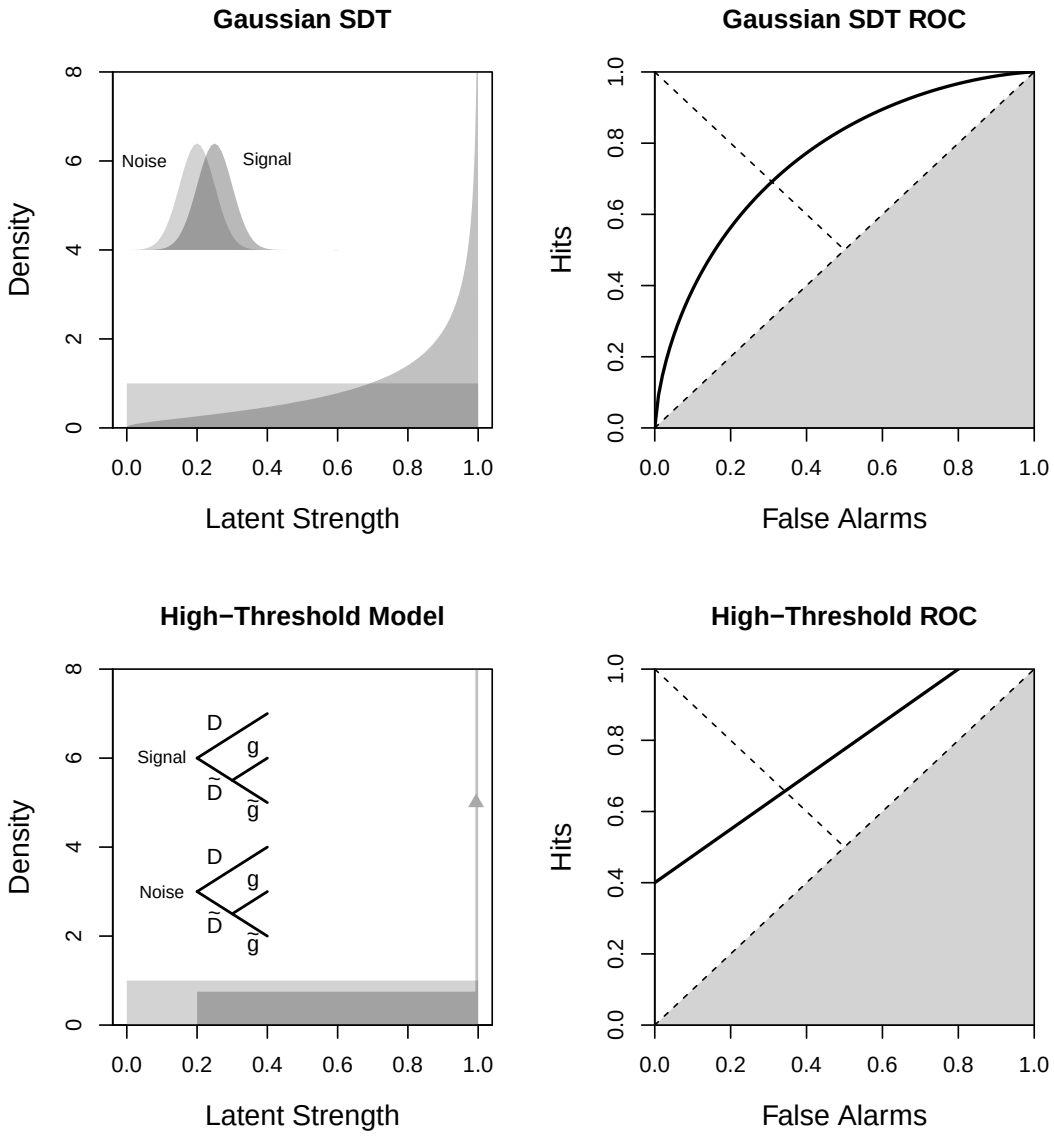

Figure 2. Examples of models that can be cast in a *universal SDT representation*. *Left Column*: Universal representation of models, along with their traditional representations (embedded in each panel). The darker density corresponds to the signal distribution, the lighter density the noise distribution. The parameters (e.g., D, g) 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. *Right Column*: The models' respective ROCs. Dashed lines delimiting the gray areas indicate chance-level performance.

from *multiple-alternative forced-choice* (AFC) tasks. The simplest case of such a task is two-alternative forced choice (2-AFC), in which individuals are requested to choose one stimulus out of two available stimuli. Although some degree of attention has been given to 2-AFC judgments (e.g., Jang, Wixted, & Huber, 2009; Province & Rouder, 2012), hardly any has been given to the relationships between forced-choice judgments across multiple choice sets (e.g., 2-AFC and 3-AFC).

In the domain of forced-choice judgments, it can be shown that the SDT model needs to satisfy a set of constraints originally formalized by Block and Marschak (1960) and Falmagne (1978). These results were originally discussed in the context of the class of *'random-utility'* or *'random-scale'* models, which include SDT, along with *the family of Thurstone models* (Thurstone, 1927; Torgerson, 1958), and *Luce's Choice Theory* (Luce, 1959), among others (for reviews, see Marley, 1990; Marley & Regenwetter, 2017). Let a, b, c, d, ..., denote alternatives in a choice set T, and let $B \subseteq T$ be a non-empty choice subset. Now, let $P_a^{\langle B \rangle}$ denote the probability that a decision maker chooses alternative a among the alternatives in subset B. Also, we use the operator '$\setminus\{\ \}$' to indicate choice alternatives removed from choice subset B. For example, $P_a^{\langle B \setminus \{b\} \rangle}$ denotes the probability of choosing a among subset B *minus* alternative b. Moreover, let us assume that each alternative is associated with a random variable ϵ, which are used as the basis for the comparisons being made. In other words, when presented with choice subset $B = \{a, b, c, d\}$, the decision maker's judgments are based on ϵ_a, ϵ_b, ϵ_c, and ϵ_d. The values taken by these random values correspond to samples from latent-strength distributions, each associated to a given option. The decision maker is assumed to choose the option associated with with the *largest latent-strength value*, with no possibility of ties.

Importantly, note that this set of assumptions implies that *the different alternatives can be ranked* according to their latent values. For instance, consider a situation in which the decision maker is requested to choose an option from a choice set, and subsequently requested to choose another option from the subset that excludes her first choice. A repetition of such requests until no options are left yields a sequence of choices that reflect the ranking of the options.

According to the *random-scale representation* following from these assumptions, the probability of choosing a among the alternatives in choice subset B corresponds to

$$P_a^{\langle B \rangle} = P(\epsilon_a = \max_{z \in B}(z)), \quad a \in B \subseteq T, \tag{6}$$

without the possibility of ties. Block and Marschak (1960) showed that the choice

probabilities coming from a random-scale representation need to comply with a system of linear inequalities:[3]

$$\begin{aligned} P_a^{\langle B\rangle} &\geq 0, \\ P_a^{\langle B\setminus\{b\}\rangle} &\geq 0, \\ P_a^{\langle B\setminus\{b\}\rangle} - P_a^{\langle B\rangle} &\geq 0, \\ P_a^{\langle B\rangle} + P_a^{\langle B\setminus\{b,c\}\rangle} - P_a^{\langle B\setminus\{b\}\rangle} - P_a^{\langle B\setminus\{c\}\rangle} &\geq 0, \\ P_a^{\langle B\setminus\{b,c,d\}\rangle} + P_a^{\langle B\setminus\{b\}\rangle} + P_a^{\langle B\setminus\{c\}\rangle} + P_a^{\langle B\setminus\{d\}\rangle} & \\ -P_a^{\langle B\rangle} - P_a^{\langle B\setminus\{b,c\}\rangle} - P_a^{\langle B\setminus\{b,d\}\rangle} - P_a^{\langle B\setminus\{c,d\}\rangle} &\geq 0, \\ etc. & \end{aligned} \tag{7}$$

As shown in the following, the system of inequalities in Equation 7 follows from the key notion that the rank-order probabilities for any subset of options is independent from the larger subset in which they are embedded. This notion is far from uncontroversial, as it has been rejected in a variety of domains, as reported in the rapidly-expanding literature on 'context effects' (for recent examples involving perceptual judgments, see Spektor, Kellen, & Hotaling, in press; Trueblood, Brown, Heathcote, & Busemeyer, 2013).

As an example, let us consider the third inequality, $P_a^{\langle B\setminus\{b\}\rangle} - P_a^{\langle B\rangle} \geq 0$, in a case where choice subset B is comprised of four alternatives a, b, c, and d. First, note that $P_a^{\langle B\rangle}$ corresponds to a sum of all rank-order probabilities for which ϵ_a has rank 1 (largest value) and the remaining alternatives can be ranked in any way:

$$\begin{aligned} P_a^{\langle B\rangle} \;=\; & P(\epsilon_a > \epsilon_b > \epsilon_c > \epsilon_d) + P(\epsilon_a > \epsilon_b > \epsilon_d > \epsilon_c) + \\ & P(\epsilon_a > \epsilon_c > \epsilon_b > \epsilon_d) + P(\epsilon_a > \epsilon_c > \epsilon_d > \epsilon_b) + \\ & P(\epsilon_a > \epsilon_d > \epsilon_b > \epsilon_c) + P(\epsilon_a > \epsilon_d > \epsilon_c > \epsilon_b) \end{aligned} \tag{8}$$

[3] Note that if subset B is solely comprised of alternative a, then $P_a^{\langle B\rangle} = 1$. This uncontroversial constraint is necessary for ensuring that all choice probabilities are between 0 and 1. Also, note that we are omitting here the constraints applied to other choice probabilities, such as $P_b^{\langle B\rangle}$, $P_b^{\langle B\setminus\{c\}\rangle}$, and so forth.

Analogously, we can express $P_a^{\langle B\setminus\{b\}\rangle}$ as the sum of rank-order probabilities for which a is ranked above c and d, and option b can take on any rank:

$$\begin{aligned} P_a^{\langle B\setminus\{b\}\rangle} &= P(\epsilon_a > \epsilon_c > \epsilon_d) + P(\epsilon_a > \epsilon_d > \epsilon_c) \\ &= P(\epsilon_b > \epsilon_a > \epsilon_c > \epsilon_d) + P(\epsilon_a > \epsilon_b > \epsilon_c > \epsilon_d) + \\ &\quad P(\epsilon_a > \epsilon_c > \epsilon_b > \epsilon_d) + P(\epsilon_a > \epsilon_c > \epsilon_d > \epsilon_b) + \\ &\quad P(\epsilon_b > \epsilon_a > \epsilon_d > \epsilon_c) + P(\epsilon_a > \epsilon_b > \epsilon_d > \epsilon_c) + \\ &\quad P(\epsilon_a > \epsilon_d > \epsilon_b > \epsilon_c) + P(\epsilon_a > \epsilon_d > \epsilon_c > \epsilon_b) \end{aligned} \tag{9}$$

The third inequality in Equation 7 follows because $P_a^{\langle B\setminus\{b\}\rangle}$ includes all of the rank-order probabilities in $P_a^{\langle B\rangle}$, implying that the former has to be larger or equal to the latter. Again, the key idea here is that the relative ranking of any subset of options (a, c, and d) is independent of other options present (e.g., b). The remaining Block-Marschak inequalities can be unpacked along the same lines: For instance, $P_a^{\langle B\setminus\{b,c\}\rangle}$ corresponds to the sum of rank-order probabilities for which $\epsilon_a > \epsilon_d$, irrespective of the ranks of ϵ_b and ϵ_c.

Falmagne (1978) proved that the system of Block-Marschak inequalities is both sufficient and necessary for the existence of a random-scale representation underlying choice probabilities (see also Barberá & Pattanaik, 1986; Fiorini, 2004). What this means is that the statements "*a decision-maker follows some random-scale representation*" and "*the choice probabilities of a decision-maker conform to the Block-Marschak inequalities*" are formally equivalent. This equivalence provides us with the opportunity to directly test the assumption of a random-scale representation, which underlies SDT, by testing whether the Block-Marschak inequalities hold empirically. Importantly, no auxiliary parametric assumptions are necessary to make any of this happen – we are testing SDT in its general form. Iverson (2006) argued for the use of the Block-Marschak inequalities for testing SDT, but the challenge has not been taken up so far.

In the forced-choice designs typically considered in the context of SDT modeling,

we are dealing with scenarios that are somewhat simpler than the ones found in multi-alternative decision making. Specifically, the choice sets are usually exclusively comprised of options coming from only two classes of items – signal and noise. In a typical m-alternative forced-choice (m-AFC) trial, the decision-maker tries to discriminate a single signal option from the other $m-1$ noise options, the latter being instances of the same random variable (i.e., they are identically distributed). Let $P_C^{\langle m \rangle}$ denote the probability of a correct response in an m-alternative forced-choice trial, which corresponds to a case in which the signal option had the highest latent value:

$$P_C^{\langle m \rangle} = P(\epsilon_S > \max(\epsilon_{N,1}, \epsilon_{N,2}, ..., \epsilon_{N,m-1})) \tag{10}$$

In this context, the Block-Marschak inequalities can be cast as follows for a sequence of m-AFC trials, with $m = 1, 2, ..., M$:

$$\begin{aligned}
P_C^{\langle m \rangle} - P_C^{\langle m+1 \rangle} &\geq 0, \quad \text{for } 1 \leq m < M, \\
P_C^{\langle m-1 \rangle} - 2P_C^{\langle m \rangle} + P_C^{\langle m+1 \rangle} &\geq 0, \quad \text{for } 2 \leq m < M, \\
P_C^{\langle m-2 \rangle} - 3P_C^{\langle m-1 \rangle} + 3P_C^{\langle m \rangle} - P_C^{\langle m+1 \rangle} &\geq 0, \quad \text{for } 3 \leq m < M, \\
P_C^{\langle m-3 \rangle} - 4P_C^{\langle m-2 \rangle} + 6P_C^{\langle m-1 \rangle} - 4P_C^{\langle m \rangle} + P_C^{\langle m+1 \rangle} &\geq 0, \quad \text{for } 4 \leq m < M, \\
P_C^{\langle m-4 \rangle} - 5P_C^{\langle m-3 \rangle} + 10P_C^{\langle m-2 \rangle} - 10P_C^{\langle m-1 \rangle} + 5P_C^{\langle m \rangle} - P_C^{\langle m+1 \rangle} &\geq 0, \quad \text{for } 5 \leq m < M, \\
&\textit{etc.}
\end{aligned} \tag{11}$$

By definition, $P_C^{\langle 1 \rangle} = 1$. Each inequality is comprised of a sum of I terms with increasing set size, where $I-1$ is the minimum set size m permitted in a given inequality. The ith P_C term, with $i = 1, ..., I$, is multiplied by coefficient $(-1)^{(i+1)}\binom{I-1}{i-1}$. To see the connection between the inequalities in Equations 9 and 11, let us associate the signal with option a compare the sums $P_a^{\langle B \rangle} + P_a^{\langle B \setminus \{b,c\} \rangle} - P_a^{\langle B \setminus \{b\} \rangle} - P_a^{\langle B \setminus \{c\} \rangle}$, and $P_C^{\langle m-1 \rangle} - 2P_C^{\langle m \rangle} + P_C^{\langle m+1 \rangle}$. A simple rearrangement shows that $P_a^{\langle B \rangle} = P_C^{\langle m+1 \rangle}$, $2P_C^{\langle m \rangle} = P_a^{\langle B \setminus \{b\} \rangle} + P_a^{\langle B \setminus \{c\} \rangle}$ (remember that noise values come from the same noise distribution), and $P_C^{\langle m-1 \rangle} = P_a^{\langle B \setminus \{b,c\} \rangle}$.

Because the Block-Marschak inequalities are defined at the level of choice probabilities without any reference to accuracy, they impose no constraints the minimum of $P_C^{\langle m \rangle}$. Because of this situation, we will always consider the Block-Marschak inequalities along with the constraint that performance is at least at chance level; i.e., $P_C^{\langle m \rangle} \geq \frac{1}{m}$ for all m.

The Testability of the Block-Marschak Inequalities

Given our previous discussion on the flexibility of SDT in its general form, it is not unreasonable to expect the Block-Marschak inequalities to effectively impose little to no constraint. This expectation turns out to be incorrect, as these inequalities can be quite restrictive. To see this, let us consider the sequence of $P_C^{\langle m \rangle}$ going from $m = 2$ to $m = 8$. The joint outcome space for this experimental design corresponds to a seven-dimensional hypercube with unit length. Now, this space includes many values deemed unreasonable in the context of a forced-choice task, such as values corresponding to below-chance performance. Moreover, one can also assume that the assumption of '*regularity*' holds: Performance should never increase when introducing an additional option to the choice set. In other words, $P_C^{\langle 2 \rangle} \geq P_C^{\langle 3 \rangle} \geq P_C^{\langle 4 \rangle} \geq ... \geq P_C^{\langle 8 \rangle}$. With the support of a reverse-search algorithm (Avis & Fukuda, 1992), we found that these constraints limit predictions to approximately $\frac{1}{13{,}021}$ of the outcome space. Note that the Block-Marschak inequalities imply regularity (it is established by the first inequality in Equation 11) but the converse is not true (e.g., $P_C^{\langle 2 \rangle} \geq P_C^{\langle 3 \rangle} \geq P_C^{\langle 4 \rangle}$ does not imply that $P_C^{\langle 2 \rangle} + P_C^{\langle 4 \rangle} \geq 2P_C^{\langle 3 \rangle}$). Once again, using the reverse-search algorithm we found that the volume defined by the Block-Marschak inequalities is approximately $\frac{1}{37{,}405{,}425}$ of the (already restricted) volume satisfying above-chance performance and regularity. This result demonstrates the potential of the Block-Marschak inequalities for testing SDT, as any reasonable attempt to a test a theory requires that it predict only a small fraction of the space of possible outcomes (Roberts & Pashler, 2000; see also Regenwetter et al., 2011). As an example, consider the $P_C^{\langle m \rangle}$ sequence $[.84, .75, .69, .55, .46, .39, .34]$, from $m = 2$ to $m = 8$. At first glance, this sequence

appears to be completely unproblematic, as performance is always above chance and regularity holds. However, the Block-Marschak inequalities are violated. For instance, $P_C^{\langle 3 \rangle} - 2P_C^{\langle 4 \rangle} + P_C^{\langle 5 \rangle} = 0.73 - 1.34 + 0.50 = -0.11$. These data can be shown to emerge from 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).[4]

Another important aspect of the Block-Marschak inequalities is the fact that they are *closed under averaging or aggregation*. Specifically, these inequalities define a convex polytope, which means that, if individuals in a given population have heterogeneous performances respecting the Block-Marschak inequalities, then their aggregated or averaged performance will also respect them (for a recent overview, see Regenwetter & Robinson, 2017). It is straightforward to see why this is the case here: The sum on the left-hand side of each inequality is always non-negative, and therefore its average with any other non-negative sums will also be non-negative. This closure under aggregation is very fortunate because it means that we do not have to worry about spurious rejections due to aggregation, as typically observed in the literature (e.g., Estes, 1956; Heathcote, Brown, & Mewhort, 2000). On top of that, it also enables the testing of the Block-Marschak inequalities in domains such as eyewitness testimony where analyses are generally conducted on aggregate data due to the fact that each witness provides a single response to a given suspect lineup, and that samples from latent distributions are assumed to be correlated (e.g., Wixted, Vul, Mickes, & Wilson, in press).[5]

Having established the constraints associated with Block-Marschak inequalities, and the ability to test them with aggregate data, we now turn to the issues of hypothesis testing and statistical power. The statistical testing of hypotheses comprised

---

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

[5] Although the Block-Marschak inequalities cannot be spuriously rejected based on aggregate data, they could be spuriously accepted. Specifically, one can conceive scenarios in which data from individuals outside of the polytope can fall within it when aggregated. We find this possibility unlikely due to the small volume of the polytope, which means that data outside of the polytope would have to conspire so that their aggregation would end up respecting the inequalities (for a discussion, see Birnbaum, 2011; Regenwetter, Dana, Davis-Stober, & Guo, 2011).

of inequalities is an ongoing area of research (see Davis-Stober, 2009; Hoijtink, 2011; Myung, Karabatsos, & Iverson, 2008; Silvapulle & Sen, 2011). The key challenge is that classic testing solutions cannot be applied. For instance, the G^2 statistic often used to quantify model misfit under the null hypothesis no longer follows a χ^2 distribution with a given number of degrees of freedom, but a mixture of χ^2 distributions (for a review, see Davis-Stober, 2009). Here, we adopted a classical-frequentist solution along the lines of Kalish, Dunn, Burdakov, and Sysoev (2016), with model fitting being cast as solving a quadratic programming problem: Finding the expected choice probabilities that best approximate the data (in this case, minimize the sum of squared errors) while enforcing the constraint that they need to fulfill the Block-Marschak inequalities. Model fit was evaluated via a semi-parametric bootstrap procedure in which data are generated based on model fits to bootstrapped data (for details, see Wagenmakers, Ratcliff, Gomez, & Iverson, 2004).[6]

Finally, we assessed statistical power by generating above-chance data that conform to the regularity assumption for $m = 2, ..., 8$.[7] With 500 and 1000 trials per m-AFC condition and a critical p-value of .05, we correctly rejected the Block-Marschak inequalities 86% and 95% of the times, respectively. We selected these sample sizes because they roughly match the sample sizes of Experiments 1 (roughly 1000 trials) and 2 (roughly 500 trials) reported below. These simulation results complement the previous volume analysis of the inequalities: Not only are they highly restrictive, their violation can be reliably captured when tested statistically.

Reanalysis of Swets (1959)

As a first empirical test of the Block-Marschak inequalities, 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

---

[6] Although we rely on a frequentist approach, Bayesian solutions are also available (see McCausland & Marley, 2014).

[7] For each m, response proportions were obtained by generating random values from a uniform distribution ranging between $\frac{1}{m}$ and .85. Cases in which regularity did not hold were discarded.

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.[8]

As shown in Figure 3, 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 $P_C^{\langle 3 \rangle}$ and $P_C^{\langle 4 \rangle}$. Overall, the results show that at least two out of the three individual datasets reported by Swets (1959) are generally compliant with the Block-Marschak inequalities.

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

As previously stated, the Block-Marschak inequalities make no assumptions regarding the dependencies between the latent-strength values compared (e.g., ϵ_S might be correlated with the ϵ_N values). However, in typical SDT implementations, it is assumed by default that the sampled latent-strength values ϵ are *independently* distributed within and across choice sets. The *independence assumption* is associated with its own set of constraints, which can also be tested. But more importantly, this assumption plays a key role in SDT, establishing the close relationship between different types judgments, namely yes-no, forced-choice, and ranking judgments.

The Relationship Between Forced-Choice Judgments and the Yes-No ROC

Under the assumption of independence, it can be shown that each $P_C^{\langle m \rangle}$ corresponds to the $m-1$th moment of the yes-no ROC function ρ. Once again, relying on the universal SDT representation:

---

[8] 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 as described in Equation 11 given that they require complete sequences of m. We overcame this problem by canceling the $P_C^{\langle 5 \rangle}$ and $P_C^{\langle 7 \rangle}$ terms in the set of Block-Marschak inequalities through the weighted sum of inequalities in which the to-be-cancelled terms have opposite signs.

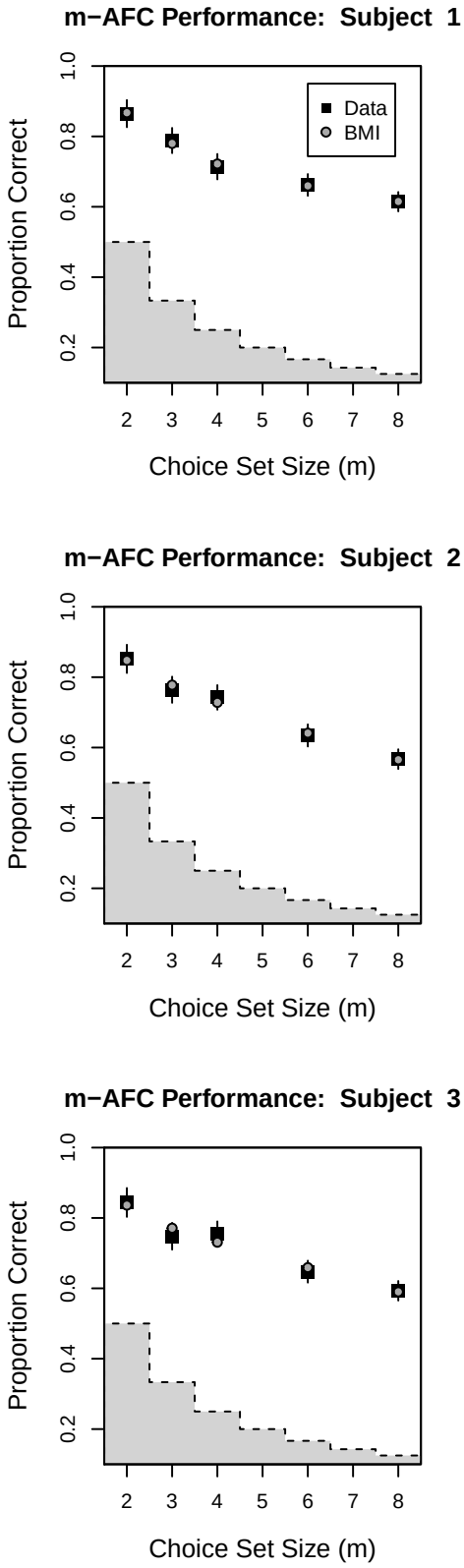

Figure 3. Analysis of individual data from Swets (1959). The bars represent 95% confidence intervals. BMI = Best-fitting predictions that respect the Block-Marschak inequalities.The dashed lines delimiting the gray areas indicate chance-level performance.

$$\begin{aligned} P_C^{\langle m \rangle} &= P(\epsilon_S > \max(\epsilon_{N,1}, \epsilon_{N,2}, ..., \epsilon_{N,m-1})) \\ &= \int_0^1 F_N(t)^{m-1} f_S(t) \\ &= \int_0^1 t^{m-1} \, \mathrm{d}F_S(t) \\ &= \mathbb{E}(\epsilon_S^{m-1}). \end{aligned} \tag{12}$$

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).[9] Note that $P_C^{\langle 2 \rangle} = \mathbb{E}(\epsilon_S)$, which corresponds to the area under the ROC function. This equality is the famous *Area Theorem* established by Green and Moses (1966) relating 2-AFC performance with the yes-no ROC (for a recent test in recognition memory, see Jang, Wixted, & Huber, 2009). The result in Equation 12, which includes Green and Moses' Area Theorem as a special case, was coined the *Generalized Area Theorem* by Iverson and Bamber (1997).

The introduction of an independence assumption introduces further constraints into SDT, some of which might not be obvious at first glance: For example, consider the values $P_C^{\langle 2 \rangle} = .90$ and $P_C^{\langle 3 \rangle} = .70$, $P_C^{\langle 4 \rangle} = .60$. These values seem reasonable, after all they are all above chance, regularity is satisfied, and $P_C^{\langle 2 \rangle} - 2P_C^{\langle 3 \rangle} + P_C^{\langle 4 \rangle} = 0.10$. However, they cannot be accommodated by any SDT model under the assumption of independence. To see this, simply note that the variance of ϵ_S is $\mathbb{E}(\epsilon_S^2) - \mathbb{E}(\epsilon_S)^2$, which in this specific case would be negative as $P_C^{\langle 3 \rangle} - (P_C^{\langle 2 \rangle})^2 = .70 - .81 = -0.11$. This example is part of a set of multiplicative inequalities introduced by the assumption of

---

[9] Iverson and Bamber (1997) proposed a ROC reconstruction method using the estimated moments (i.e., the observed $P_C^{\langle m \rangle}$) via Legendre polynomials. One problem with this approach is that it requires the $P_C^{\langle m \rangle}$ to be high precision values, which is unfeasible in any realistic paradigm where the granularity of probability estimates is determined by the number of trials. One remedy to this challenge is to rely on the $P_C^{\langle m \rangle}$ estimated from a parametric SDT model. Although we obtained satisfactory results with such an approach, it ultimately defeats the present purpose of testing SDT *without making parametric assumptions*. Later on, we will discuss a more robust approach for reconstructing ROCs that relies on ranking probabilities.

independence (Sattath & Tversky, 1976; see also Suppes et al., 1989). Let $B, C \subseteq T$ be subsets with some overlap (i.e., $B \cap C \neq \emptyset$) by at the very least both including alternative a. Then:

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

In the context of an m-AFC task, a series of multiplicative inequalities follows from independence:

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

These *independence inequalities* include the 'lower boundaries' in m-AFC discussed by Shaw (1980) as special cases (see also Macmillan & Creelman, 2005, p. 251). We checked how many correct-response probabilities from $m = 2$ to $m = 8$ that were above chance and respected the regularity assumption turned out to conform to the independence inequalities. We found that only 7% of them did. Although they are far less strict than the Block-Marschak inequalities, conforming to them by mere chance alone is nevertheless unlikely. We again estimated power using the previous simulation setup with above-chance performance and regularity. We found that with 500 and 1000 trials per m-AFC condition, we correctly rejected the independence inequalities 79% and 85% of the times, respectively.

The Relationship Between Forced-Choice and Ranking Judgments

Based on Equation 12, we can easily establish 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 independence assumption (and once again using the universal representation), ranking probabilities are given by:

$$\begin{aligned} R_i^{\langle m \rangle} &= \binom{m-1}{i-1} \int_0^1 (1 - F_N(t))^{i-1} F_N(t)^{m-i} f_S(t)\, \mathrm{d}t \\ &= \binom{m-1}{i-1} \int_0^1 (1-t)^{i-1} t^{m-i}\, \mathrm{d}F_S(t), \end{aligned} \tag{15}$$

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 15 provides us with a simple way to express ranking probabilities in terms of forced-choice probabilities. For example, consider the probability of the signal being assigned ranks 3 for $m = 4$:

$$\begin{aligned} R_3^{\langle 4 \rangle} &= 3 \int_0^1 (1-t)^2 t^2\, \mathrm{d}F_S(t) \\ &= 3 \int_0^1 t^3 - 2t^2 + t\, \mathrm{d}F_S(t) \\ &= 3 \times \left(P_C^{\langle 4 \rangle} - 2P_C^{\langle 3 \rangle} + P_C^{\langle 2 \rangle} \right). \end{aligned} \tag{16}$$

Note that the $P_C^{\langle m \rangle}$ terms included inside the parentheses of Equation 16 corresponds to one of the linear functions in the Block-Marschak inequalities (see second line of Equation 11) scaled by $\binom{m-1}{i-1} = 3$. This correspondence highlights the relationship between the Block-Marschak inequalities and ranking probabilities across set sizes – a violation of some of the inequalities would imply *negative* ranking probabilities, a nonsensical scenario.

The Relationship Between Ranking Judgments and the Yes-No ROC

At this point, it should not come as a surprise that there is a close relationship between ranking judgments and the yes-no ROC. In addition to $R_i^{\langle m \rangle}$, let us define the probability of a given noise option being assigned rank i among m options, $Q_i^{\langle m \rangle} = \frac{1 - R_i^{\langle m \rangle}}{m-1}$. Using the theoretical results developed by Feller (1971, Chap. 7),

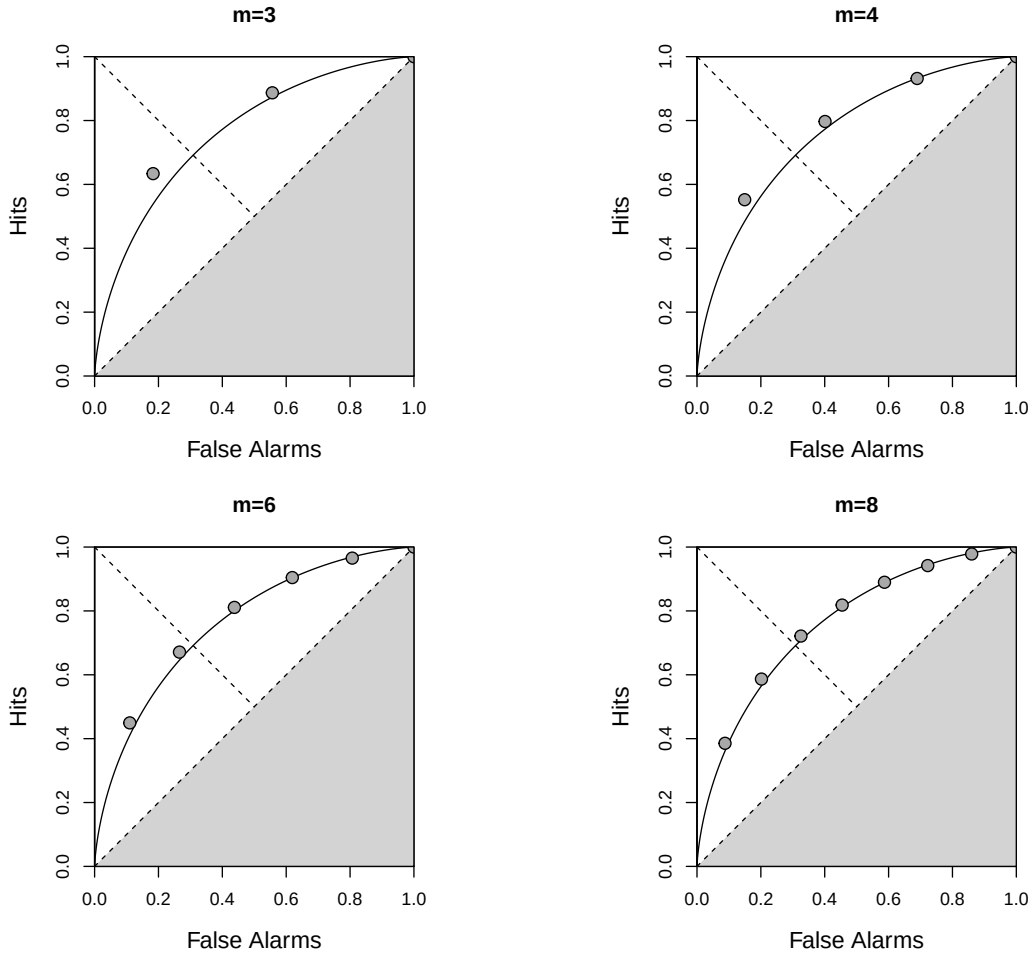

Figure 4. Theoretical and reconstructed yes-no ROC functions (lines and points, respectively) for different choice set sizes m.

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 4 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. Although the approximations are generally very good, one should keep in mind that the lower boundary of the approximation is given by $R_1^{\langle m \rangle}$ and $Q_1^{\langle m \rangle}$. In order to characterize the yes-no ROC for lower FA values, one is likely to require large m.

The relationship between ranking judgments and the yes-no ROC provides an important insight on the theoretical constraint that the slope of the yes-no ROC should be monotonically decreasing (e.g., Zhang & Mueller, 2005). Going back to Equation 5 we see that this constraint boils down to $\frac{f_S(\kappa)}{f_N(\kappa)}$ increasing along with κ. This strikes us as very reasonable given that one would expect small strength values to be mostly

associated with samples from the noise distribution, and higher values with samples from the signal distribution (Criss & McClelland, 2006; Glanzer et al., 2008; Osth & Dennis, 2015).[10] The approximation of the ROC via ranking judgments corroborates our impression: All $Q_i^{\langle m \rangle}$ quickly converge to $\frac{1}{m}$ as m increases, which means that each rank probability $R_i^{\langle m \rangle}$ corresponds to the ROC difference $\Delta\rho_i = \rho(\frac{i}{m}) - \rho(\frac{i-1}{m})$. If the ROC slope is monotonically decreasing, then the $\Delta\rho_i$ should be monotonically decreasing for increasing i. In terms of ranking judgments, this constraint implies that more egregious errors should be less likely 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. These constraints can be established just like the Block-Marschak inequalities: When expressing rank probabilities in terms of $P_C^{\langle m \rangle}$ as done in Equation 15, they merely correspond to $R_i^{\langle m \rangle} - R_{i+1}^{\langle m \rangle} \geq 0$, for $1 \leq i < m$.

When introduced along with the Block-Marschak inequalities for a sequence of m-AFC judgments from $m = 2$ to $m = 8$, these monotonic-likelihood constraints considerably reduce the volume of $P_C^{\langle m \rangle}$ values occupied by SDT. Specifically, it now corresponds to $\frac{1}{413{,}121{,}934{,}659}$ of the volume of possible $P_C^{\langle m \rangle}$ values that are above chance and respect regularity. With a sample size of 500/1000 trials per m, the probability of this extended set of restrictions being rejected by above-chance $P_C^{\langle m \rangle}$ values satisfying regularity was 95%/99% (for details, see Footnote 7). These simulations show that a study with these samples sizes is well powered to detect inequality violations.

Reconstructing ROCs from Previous Studies using Ranking Judgments

The existence of previous recognition-memory studies collecting ranking judgments allows us to reconstruct the underlying yes-no ROC. In the experiments reported by Klauer and Kellen (2014) and McAdoo and Gronlund (2016), the strength of the studied items was manipulated via study repetition. The studied/tested items in Klauer and Kellen's experiments were common words, whereas in the case of McAdoo

[10] This assumption is respected by threshold models like the one depicted in Figure 2. However, it is violated by the unequal-variance Gaussian SDT model, a situation that has long been understood as a problem that compromises the theoretical status of this model (see Green & Swets, 1966, Chap. 3).

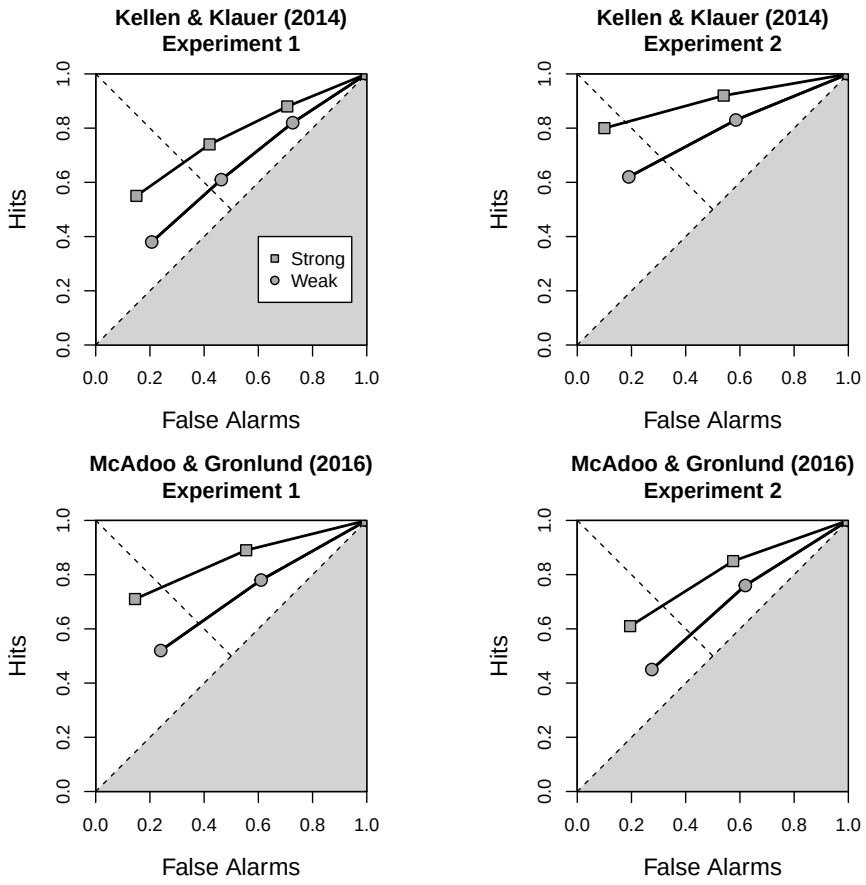

Figure 5. Reconstructed yes-no ROC functions based on ranking judgments.

and Gronlund's they were human faces. As shown in Figure 5, the reconstructed yes-no ROCs appear to have a slightly concave shape (but barely), with the ROCs concerning the strong studied items dominating their weak counterparts. Also, these ROCs appear to be asymmetric relative to the negative diagonal. Note, however, that the small choice set size in each of these studies ($m = 3$ and $m = 4$) only allows for rather crude reconstructions (see Figure 4).

Figure 6 illustrates the reconstructed ROCs obtained with the four-alternative "answer-until-correct" forced-choice paradigm used by 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" conditions dominating "weaker" ones. But again, many of these ROCs do not manifest the curvature that is often expected under certain parametrizations.

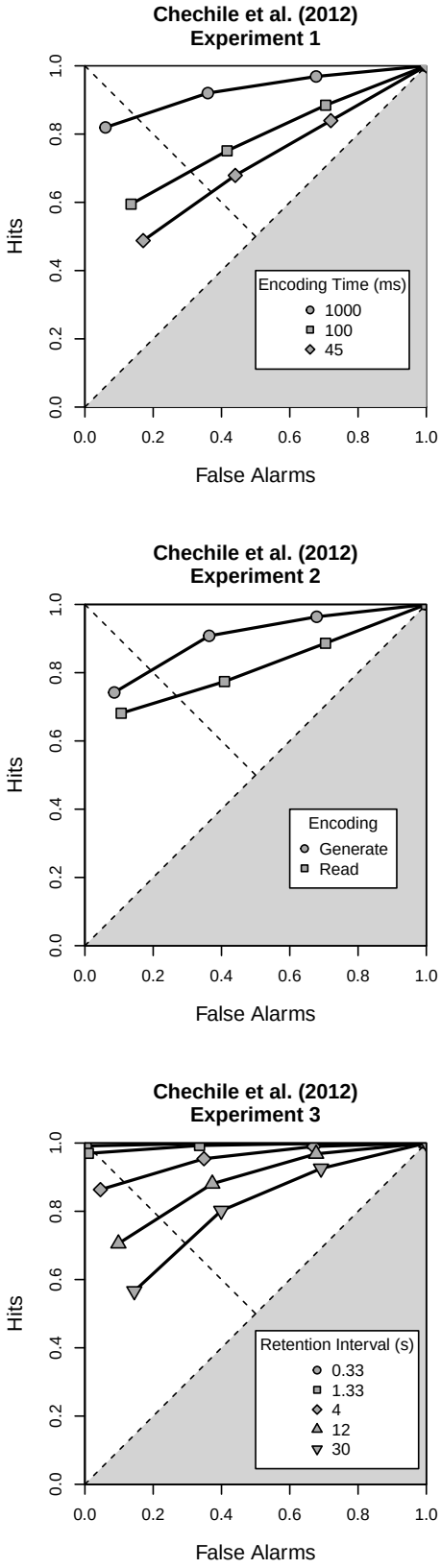

Figure 6. Reconstructed yes-no ROC functions based on "answer-until-correct" judgments.

Interim Discussion

At this point, it is useful to briefly review the main theoretical points made so far. First, we discussed how SDT as a general theoretical framework is often conflated with the auxiliary parametric assumptions used in its implementation. We also showed that, contrary to what is commonly assumed, SDT can in principle account for any ROC shape. This state of affairs suggests that in the absence of parametric assumptions, SDT is not really testable. However, using theoretical results by Block and Marschak (1960) and Falmagne (1978), we showed that even in such general form, SDT enforces a set of constraints on forced-choice judgments known as the Block-Marschak inequalities. If the Block-Marschak inequalities hold, then there exists a SDT (or random scale) representation of forced-choice judgments in which the alternatives are ranked according to their latent strength. If the Block-Marschak inequalities are rejected, then no SDT representation is possible. Because these inequalities are only satisfied by a small set of possible forced-choice accuracy values, they constitute a severe test on SDT that is extremely unlikely to succeed by mere chance alone, even when minimal conditions hold (e.g., above-chance performance and regularity).

We also discussed the assumption of independence in SDT (which is not part of the Block-Marschak inequalities), according to which the latent-strength values taken by the different options are mutually independent. This assumption, which is typically adopted in the context of SDT, yields its own set of multiplicative inequalities that also restrict to a considerable degree the range of possible forced-choice probabilities (Sattath & Tversky, 1976). Importantly, the independence assumption establishes a close connection between forced-choice judgments, ranking judgments, and the yes-no ROC function, which encompasses the famous area theorem (Feller, 1971; Iverson & Bamber, 1997). This connection enables us to reconstruct the yes-no ROC function with considerable precision as well as to motivate further constraints concerned with likelihood monotonicity.

Altogether, these theoretical results allow us to implement different tests on SDT that so far have been overlooked in the literature. In our first experiment, we will use a

large sample of participants to test the Block-Marschak inequalities (with and without the monotonic-likelihood constraints) and the multiplicative inequalities implied by independence. The estimated power of these tests given our sample size are 99%, 95%, and 85%, respectively. Based on the results from these tests, we will attempt to reconstruct the yes-no ROC. In our second experiment, we will test the accuracy of this reconstucted ROC by directly comparing it with actual yes-no judgment probabilities. A third experiment will focus on the symmetry of the yes-no ROC using a direct test that will be discussed in detail later on.

Experiment 1

Participants were presented with a randomized set of recognition-memory m-AFC trials, with m varying across the complete integer sequence between 2 and 8. In each trial, participants were 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 participant and AFC condition, and placed our focus on the aggregated data.

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.[11] The experiment began with a study phase in which participants were presented 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 choice set-size

[11] The reward provided for this period of time roughly corresponds to a $15.00 hourly wage, more than the US federal minimum wage.

condition (randomly intermixed), with set sizes ranging from $m = 2$ to $m = 8$ in steps of 1. Words were presented in the center of the screen. For smaller set sizes and/or trials with shorter words, all words were presented next to each other. For larger set sizes and/or trials with longer words, presentation was split into more than one row with several words in each row presented next to each other. Participants selected the word of their choice by clicking on it. After completing the test trials, participants filled in a short demographic survey, were thanked, and received their monetary reward.

Results and Discussion

As shown in the left panel of Figure 7, forced-choice accuracy was clearly above chance for all choice set sizes m. Moreover, the assumption of regularity also appeared to hold, with performance decreasing with increasing choice set size m. But do the data respect the Block-Marschak inequalities? Model fits shown in the left panel of Figure 7 indicate a near-perfect fit, with $G^2 = 0.13$, $p = .98$. Further introducing the monotonic likelihood constraint also led to a very good fit. ($G^2 = 0.88$, $p = .93$). Finally, we tested the independence inequalities and found all of them to hold perfectly ($G^2 = 0$, $p = 1$).

Given that none of the inequality constraints was rejected, we used the best-fitting $P_C^{\langle m \rangle}$ estimates satisfying all inequalities to reconstruct the yes-no ROC. The reconstruction shown in the right panel of Figure 7 suggests that the yes-no ROC takes on a somewhat concave shape. The ROC is also asymmetric relative to the negative diagonal (dashed line where H + FA = 1). Again, it should be highlighted that this ROC reconstruction sidesteps the need for an experimental design introducing response-bias manipulations or requesting confidence judgments, along with a set of auxiliary assumptions regarding selective influence or state-response mapping. On the other hand, this ROC was reconstructed on the basis of aggregated data, which could have somewhat distorted its shape (e.g., Morey, Pratte, & Rouder, 2008; Trippas et al., in press). Although we cannot dismiss this concern, the impact of aggregation is expected to be relatively minor and mostly affecting the symmetry of the function (see Pratte & Rouder, 2010). We will return to this issue later, when proposing and

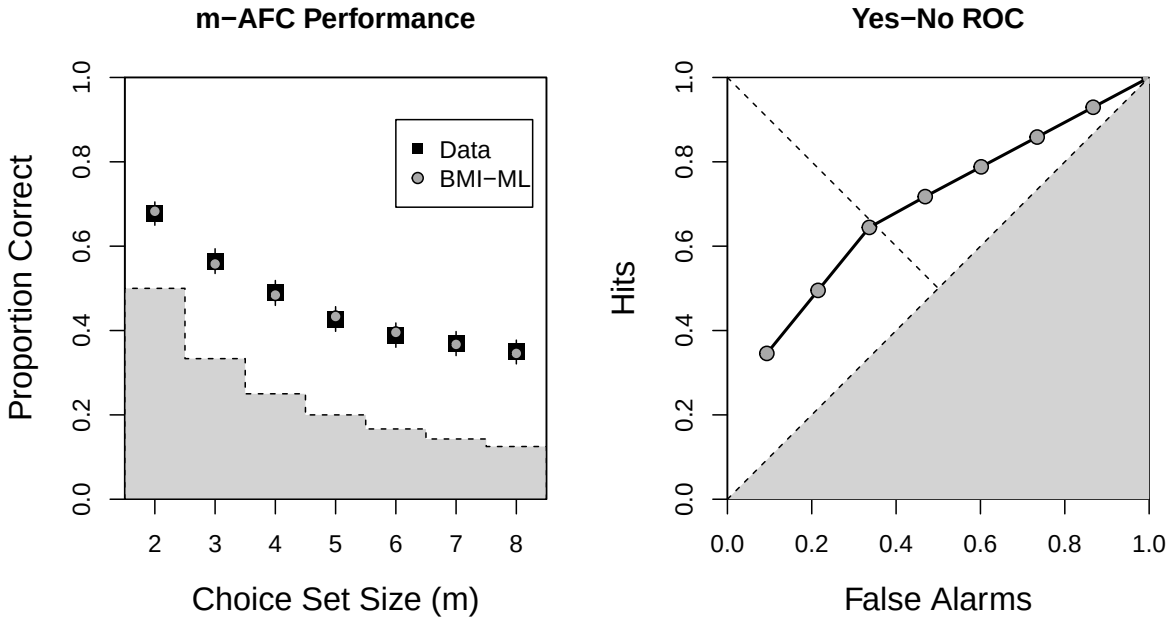

Figure 7. Experiment 1 Results. Observed and predicted performance in the m-AFC trials. Bars represent 95% confidence intervals. BMI-ML = Best-fitting estimates that respect the Block-Marschak and monotonic-likelihood inequalities. *Right Panel*: Reconstructed yes-no ROC using the predicted forced-choice performance. In both panels, the dashed lines delimiting the gray areas indicate chance-level performance.

implementing a direct test for ROC symmetry.

Experiment 2

Experiment 1 tested the Block-Marschak, monotonic likelihood, and independence inequalities, which were all found to hold. Based on the estimates obtained, we were able to reconstruct the yes-no ROC based on the forced-choice judgments. Experiment 2 replicates Experiment 1 with a slightly modified design in which yes-no judgments for single items are also requested. The rationale here is that the performance in the m-AFC trials will strongly constrain the expected hit and false-alarm rates. If the reconstructed ROC is accurately capturing the relationship between hits and false alarms, then it should be able to predict where the yes-no ROC point will lie.

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 the exception of two changes. First, we introduced twenty single-item trials (10 old and 10 new) 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.

Results and Discussion

The data are shown in Figure 8. As before, performance is above chance and decreasing as m increases. Note that performance was generally better than in Experiment 1, a result that can be attributed to the fewer number of m-AFC trials 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).

The data were again consistent with the Block-Marschak inequalities ($G^2 = 0.70$, $p = .91$). As before, introducing the monotonic likelihood constraints leads to virtually identical fit ($G^2 = 0.74$, $p = .95$). Also, the independence inequalities associated were once again perfectly fulfilled by these constrained estimates ($G^2 = 0$, $p = 1$). The right panel of Figure 8 shows the reconstructed yes-no ROC: Again, this function takes on a concave, asymmetric shape. Moreover, the yes-no data point is perfectly consistent with the ordered ROC points in the sense that altogether they can be captured by a single monotonically-increasing function with monotonically decreasing slope. This consistency between the reconstructed ROC and the observed yes-no point provides additional support to the different sets of constraints associated with SDT.

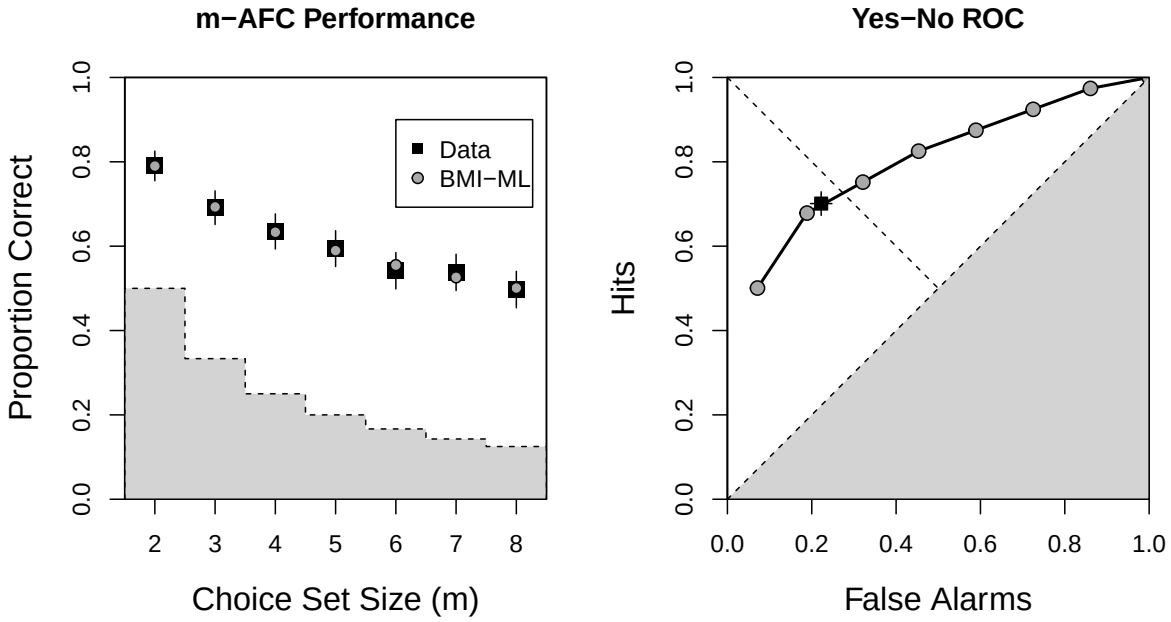

Figure 8. Experiment 2 Results. Observed and predicted performance in the m-AFC trials. Bars represent 95% confidence intervals. BMI-ML = Best-fitting estimates that respect the Block-Marschak and monotonic-likelihood inequalities. *Right Panel*: Reconstructed yes-no ROC using the predicted forced-choice performance. 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 Relationship Between Yes-No ROC Symmetry and Forced-Choice Accuracy

The reconstructed yes-no ROCs obtained in both Experiments 1 and 2 have a concave shape, and both appear to be *asymmetric* relative to the negative diagonal. We will now focus on the latter property, implementing a direct test of ROC symmetry that does not rely on the shape of the underlying strength distributions. Violations of ROC symmetry have been a major motivation for the development of extensions of the equal-variance Gaussian SDT model in recognition memory, introducing notions such as variable encoding, attentional failure, or additional retrieval processes (for a review of different models, see Yonelinas & Parks, 2007). Also, the assessment of ROC properties such as symmetry is often of critical importance given the practical implications it can have in the assessment of different decision makers (for a discussion, see Rotello, Heit,

& Dubé, 2015).

Current methods for assessing ROC symmetry raise some concerns. For instance, they often rely on confidence-rating data (e.g., Yonelinas & Parks, 2007). A recent critical test by Kellen and Klauer (2015) showed that confidence ratings do not behave as expected under a model such as the Gaussian SDT model (for a replication of these results, see McAdoo et al., in press). As an alternative, one could collect binary yes-no judgments across different response-bias conditions, but this approach often leads to extremely noisy data that can fail to meet some basic selective-influence assumptions (see Kellen, Klauer, & Bröder, 2013; Van Zandt, 2000). As a solution to these challenges, we will rely on a direct test of ROC symmetry that hinges on a simple equality prediction across response probabilities, and that does not require any kind of parametric assumptions.

Formally, a 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). This constraint can be expressed in terms of F_S and F_S^{-1}:

$$F_S(t) + F_S^{-1}(1-t) = 1. \tag{17}$$

Iverson and Bamber (1997) showed that ROC symmetry implies an equality between m-AFC judgments and judgments made in a 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.[12] To see this, let us denote the probability of a correct response by $P_C^{\langle m\star\rangle}$, which according to a latent-strength representation corresponds to the probability that the single value of ϵ_N is smaller than all the $m-1$ ϵ_S values. Again, using the universal SDT representation:

$$P_C^{\langle m\star\rangle} \quad = \quad P(\epsilon_N < \min(\epsilon_S^{\langle 1\rangle}, \epsilon_S^{\langle 2\rangle}, ..., \epsilon_S^{\langle m-1\rangle})),$$

[12] 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 this task participants are requested to choose the 'odd item' without knowing whether the m-alternative choice set includes $m-1$ signal or noise stimuli.

$$= \int_0^1 f_N(t)(1 - F_S(t))^{m-1}\, \mathrm{d}t. \tag{18}$$

With some rearrangement we can see that

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

If symmetry holds, then $F_S(t) = 1 - F_S^{-1}(1-t)$ (see Equation 17) and therefore $P_C^{\langle m\star\rangle} = P_C^{\langle m\rangle}$ for all m (compare the penultimate and last lines of Equations 12 and 18, respectively). The expected equality in choice accuracy between m-AFC and $m^\star$-AFC under ROC symmetry allows us to dismiss ROC symmetry if any $P_C^{\langle m\star\rangle}$ systematically differs from $P_C^{\langle m\rangle}$ for some m (for a recent implementation in syllogistic reasoning, see Trippas et al., in press). With the type of asymmetry typically observed in recognition-memory ROCs, including the one reconstructed in Experiment 1, we expect $P_C^{\langle m\rangle} > P_C^{\langle m\star\rangle}$. For example, an unequal-variance Gaussian SDT model with parameters $\mu_S = 1$ and $\sigma_S^2 = 1.3$, yielding an asymmetric ROC, expects the following forced-choice accuracy: $P_C^{\langle 4\rangle} = .55$, $P_C^{\langle 5\rangle} = .49$, and $P_C^{\langle 6\rangle} = .45.$, whereas $P_C^{\langle 4\star\rangle} = .51$, $P_C^{\langle 5\star\rangle} = .45$, and $P_C^{\langle 6\star\rangle} = .40$.

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 independence, the distributions underlying preferences must be double-exponential distributed, which in turn implies that choices 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 $P_C^{\langle 4\rangle} = .48$, $P_C^{\langle 5\rangle} = .40$, and $P_C^{\langle 6\rangle} = .35.$, which are all respectively smaller than $P_C^{\langle 4\star\rangle} = .55$, $P_C^{\langle 5\star\rangle} = .50$, and $P_C^{\langle 6\star\rangle} = .47$.

Experiment 3

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

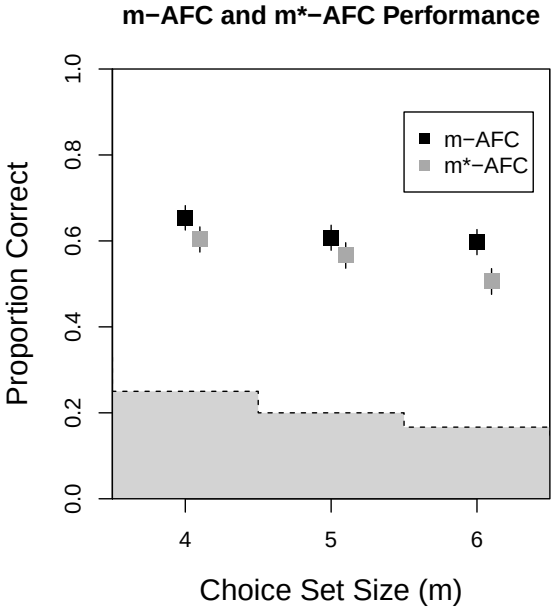

Figure 9. Experiment 3 results. The bars represent 95% confidence intervals.

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 trials per m). 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.[13] 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 9. These proportions are all above chance and respect regularity, but the $m^\star$-AFC judgments ended up violating the Block-Marschak inequalities, as $P_C^{\langle 4\star\rangle} - 2P_C^{\langle 5\star\rangle} + P_C^{\langle 6\star\rangle} = -0.03$. However, this violation 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 $P_C^{\langle 4\rangle} \geq P_C^{\langle 4\star\rangle}$, $P_C^{\langle 5\rangle} \geq P_C^{\langle 5\star\rangle}$, and $P_C^{\langle 6\rangle} \geq P_C^{\langle 6\star\rangle}$. We then compared this fit with the one from another model imposing the equality constraints $P_C^{\langle 4\rangle} = P_C^{\langle 4\star\rangle}$, $P_C^{\langle 5\rangle} = P_C^{\langle 5\star\rangle}$, and $P_C^{\langle 6\rangle} = P_C^{\langle 6\star\rangle}$. The difference in fit between the two models was $\Delta G^2 = 16.62$, $p < .001$, indicating that the equality constraint cannot provide a reasonable characterization of the data. This rejection in turn implies that the yes-no ROC is asymmetric, corroborating the visual inspection of the yes-no ROCs in Figures 7 and 8. Note that these results also imply the rejection of the latent distributions assumed by Luce's Choice Theory (Luce, 1959, 1977; Yellott, 1977) as they predict the opposite pattern, namely $P_C^{\langle m\rangle} \leq P_C^{\langle m\star\rangle}$.

General Discussion

Following the seminal work of Thurstone (1927), Signal Detection Theory postulates that judgments are based on an evaluation of sampled latent-strength values. This notion of a 'random-scale representation' is almost invariably found in the theoretical accounts proposed in cognitive psychology at large, a popularity that can be largely attributed to its empirical success and ability to connect with other theoretical notions (e.g., concerning forgetting, generalization, etc.) as well as more detailed process models (e.g., Criss & McClelland, 2006; Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997). But as important as these qualities might be, they do not constitute a

[13] O'Connor et al. (2011) had participants perform 3-AFC and 3*-AFC judgments. At face value, the results reported by O'Connor et al. suggest that there is no difference between both types of judgments. However, given the small number of participants and trials, and the use of $m = 3$ for which the predicted differences are often minute, it is likely that their data are simply not diagnostic for our present purposes.

strong empirical foundation to SDT, in the sense that they do not yield a set of behavioral constraints that are both sufficient and necessary for the existence of a random scale representation, in particular constraints that are extremely unlikely to hold by mere chance alone.

The present work discussed the basic constraints that SDT is subjected to in order for a random-scale representation to exist, namely the Block-Marschak inequalities (Block & Marschak, 1960; Falmagne, 1978). We also discussed additional constraints concerning independence and monotonic likelihood that altogether establish an intimate relation between different types of judgments – yes/no, forced choice, and ranking (Iverson & Bamber, 1997; Sattath & Tversky, 1976). The experiments reported here show that people's recognition-memory judgments are generally consistent with all of the aforementioned constraints. Given the extremely small set of empirical values that satisfy all constraints – even when plausible prior conditions are assumed to hold – these results provide a considerable support for SDT as a general framework (Roberts & Pashler, 2000). Moroever, these tests yielded predictions (yes-no ROCs) that were found to be extremely accurate when predicting new and unseen yes-no data. These predictions corroborated previous results in the literature indicating that ROCs are concave and asymmetric (e.g., Dube & Rotello, 2012). The results here have important implications for ongoing discussions in recognition memory, such as the comparison between SDT and threshold models, as discussed further below. But perhaps more importantly, they highlight the existence of approaches to SDT modeling that so far have not been part of researchers' toolboxes. We therefore dedicate most of our discussion to potential future directions and the different ways that these approaches can contribute to lines of research where SDT plays a role.

Boundary Conditions

The present work focused on recognition memory, an area where SDT has played a very important role. An obvious question is whether the same type of 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*. At the heart of the Block-Marschak inequalities is the notion that the relative rank ordering probabilities for any subset of options is independent from the larger choice set in which they are embedded. This assumption is violated in domains where there is a relationship between the ranking of options and the size of the choice set, such as when *capacity is limited*. For instance, the case of visual working memory, where one expects the fidelity of stimulus representations to decrease along with increases in the choice set size (e.g., van den Berg et al., 2012), is by definition at odds with the Block-Marschak inequalities. On the other hand, one could test the Block-Marschak inequalities (and perhaps reconstruct the yes-no ROC) in designs where the choice set size is fixed (e.g., Donkin et al., 2014).

Testing SDT in More Complex Designs

One important characteristic of the present tests is that they only considered two classes of stimuli – signal and noise. Future work could consider more complex designs in which multiple classes of stimuli, such as different types of distractors, are considered. Note that these 'enriched scenarios' are already covered in the original formulation of the Block-Marschak inequalities (see Equation 7). The introduction of multiple classes of stimuli raises the question of whether or not there are 'context effects' in which the inclusion of an item of a specific class affects performance in a non-trivial manner. Perhaps by encouraging a change in the way mnemonic information is considered by the decision maker. The context effects reported in many domains amount to violations of the assumption of regularity (e.g., Spektor et al., in press; Trueblood et al., 2013), which is part of the constraints imposed by the Block-Marschak inequalities.

The possibility of context dependency in recognition judgments was highlighted in a review by Malmberg (2008), who conjectured that retrieval processes are adjusted to the composition of test items, with more recollection-type processes being involved when most of the distractors are extremely familiar and/or similar to the studied items (for a similar point, see Heathcote, Raymond, & Dunn, 2006). Moreover, the testing of

context effects would be extremely relevant in applied topics such as eyewitness identification, where the introduction of certain types of alternatives (e.g., decoys similar to a suspect) can affect performance in rather nuanced ways (see Wixted et al., in press). Importantly, the possibility to test regularity and the Block-Marschak inequalities using aggregate data should not be overlooked given the fact that much research on eyewtiness testimony cannot obtain more than a single response per subject.

Further Testing of Threshold Accounts

In Experiments 1 and 2, the ROCs obtained took on a concave shape that is at odds with a high-threshold account and more in line with the expectations of the Gaussian SDT model (see Figure 2). One concern with these results is that they are based on aggregate data (e.g., Estes, 1956), despite the fact that comparisons between these two models based on individual and aggregate data generally tend to agree (e.g., Bröder & Schütz, 2009; Dube, Starns, Ratcliff, & Rotello, 2012; Kellen, Klauer, & Bröder, 2013; Province & Rouder, 2012; see also Pratte, Rouder, & Morey, 2010). Although the present results are relevant for the ongoing evaluation of parametric models such as the high-threshold model, it is important to keep in mind that the tests conducted were targeted at the general class of SDT models. They should not be seen as replacements for critical tests of core properties of threshold models.

But what are these core properties? Rouder and Morey (2009) provide a clear answer, namely that the probability distribution of responses is invariant within each of the postulated mental states – conditional independence. Take the case of errors in ranking judgments, discussed by Kellen and Klauer (2014). Whenever an individual fails to assign rank 1 to the old word, it means that the word was not retrieved and is therefore in a state in which its status is completely uncertain. Because the probability distribution over ranking judgments is invariant under this uncertainty state, there should be no difference between the ranking probabilities for weak and strong old items (e.g., probability of rank 2, conditional that rank is not 1). Data from two experiments showed this invariance to be false. More recently, Kellen et al. (2016) showed that these

data could be accounted for by a low-threshold model (Krantz, 1969; Luce, 1963), whose predictions are generally consistent with the reconstructed yes-no ROCs obtained in Experiments 1 and 2.

Other critical tests can be developed to test more complex threshold models that encompass the ones discussed so far. Let us sketch one test here that relies on the decision maker selecting *a subset of options* she believes to be new, from a choice set that includes both new and weak/strong old items (for examples of subsetting judgments, see Regenwetter & Grofman, 1998; Regenwetter & Marley, 1998). When a subset includes at least two options, with some probability two options are selected and the decision maker is requested to choose the one they judge most likely to be old (see Parks & Yonelinas, 2009; Starns, Dubé, & Frelinger, 2018). Now, consider the cases in which the selected pair is comprised of an old and a new item: $\{\epsilon_{\text{Weak}}, \epsilon_{\text{N}}\}$ and $\{\epsilon_{\text{Strong}}, \epsilon_{\text{N}}\}$. It can be shown that a general threshold model, which includes the high and low threshold models as special cases, expects the probability of the old item being chosen to be equal for both pairs. Other models like the Gaussian SDT model expect this probability to be larger in the case of the latter pair. Efforts to compare these two hypotheses are underway.

The Usefulness of Ranking Judgments

The ability to reconstruct ROCs in Experiments 1 and 2 was enabled by the relationship between yes-no, forced-choice, and ranking judgments. We did not collect ranking judgments directly, as they would not allow us to test the Block-Marschak inequalities (they would hold by definition). However, researchers should keep in mind the advantages of collecting ranking judgments, especially in domains where the collection of forced-choice judgments across multiple set sizes is impractical or unfeasible. 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 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., in press). 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 can prove to be essential for the theoretical characterization of eyewitness judgments.[14] Also relevant here is the fact that more theoretical work is needed in the study of scenarios in which the decision maker is free to not make a choice among the available options (see Corbin & Marley, 1974).

Testing Dual-Process Theories

The fact that SDT can accommodate any possible ROC compromises the use of ROCs to support a dual-process account. Whatever additional retrieval process is being postulated, there is no single ROC data that a single-process account cannot fit, and therefore the *necessity* for a dual-process account can never be established this way. But this is a two-way street: The dual process model includes the possibility of recognition by means of a latent-strength judgment, which means that in its broadest form, it always encompasses SDT. Therefore, one cannot *reject* dual-process accounts *while simultaneously accepting* SDT. This point is often overlooked because researchers tend to confuse the key notions in a theory with the parametric assumptions used in their implementation (Kellen & Singmann, 2016).

Against this backdrop, it is important to consider which alternative routes could be taken. First, one could attempt to validate the dual-process representation by conducting studies in which each of the postulated processes is *selectively influenced* (e.g., Pratte & Rouder, 2012). Although this approach would not be able to determine the necessity of a dual-process representation, a demonstration of selective influence

[14] This suggestion only applies to research settings. We are not suggesting that a ranking procedure should be adopted by police departments.

would allow one to argue for its sufficiency, but also for the fact that it can provide a more theoretically-sensible or interpretable account. But again, the success of dual-process accounts can be by and large determined by the auxiliary assumptions adopted (for a discussion, see Kellen & Singmann, 2016). The need for auxiliary assumptions could be mitigated through the development of critical tests. For instance, the critical tests developed by Kellen and Klauer (2014, 2015) could be used for testing whether recollection has been selectively influenced.

A related approach involves the search for ROC *crossovers* (Rouder et al., 2010; Rouder et al., 2014). The idea here is that the predictions of any reasonable single-process SDT account should be restricted to ordered ROCs: For instance, if decision maker i is better at discriminating between signal and noise stimuli than decision maker j, then the ROC of the former *dominates* the latter's, such that $\rho_i(\text{FA}) \geq \rho_j(\text{FA})$, for all $0 \leq \text{FA} \leq 1$. Therefore, any violation of this ROC dominance would be completely at odds with single-process accounts in general. Typically, ROC comparisons rely on confidence judgments, but the diagnostic value of such ROCs is questionable due to the need to impose auxiliary assumptions on the familiarity and recollection processes (see Kellen & Singmann, 2016). As discussed above, it is possible to use the ranking judgments to reconstruct ROCs. These reconstructed ROCs turn out to be more diagnostic due to the fact that the auxiliary assumptions made in the case of confidence ratings are not required in the case of ranking judgments (see Kellen & Klauer, 2011). Finally, researchers should keep in mind that one can use large choice set sizes (e.g., $m = 6$) without overwhelming participants by requesting partial ranks (e.g, rank the top three or four).

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 provides some support to the ROC characterizations obtained with confidence ratings. However, it would be unwise to

interpret our results as legitimizing the use confidence-rating ROCs to estimate yes-no ROCs. Our reluctance comes from the fact that confidence judgments often do not behave as a expected and/or can change the phenomena being studied (see Benjamin, Tullis, & Lee, 2013; Brainerd, Nakamura, Reyna, & Holliday, 2017; Kellen & Klauer, 2015; Miyoshi, Kuwahara, & Kawaguchi, 2018). Also relevant is the way confidence judgments are requested (e.g., one-step versus two-step procedure; see Moran, Teodorescu, & Usher, 2015). On a more technical side, the modeling of confidence-rating judgments often require auxiliary assumptions that can affect results (e.g., how response criteria can segment regions of latent-strength values; 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.

References

Avis, D. & Fukuda, K. (1992). A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry*, *8*, 295–313.

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

Barberá, S. & Pattanaik, P. K. (1986). Falmagne and the rationalizability of stochastic choices in terms of random orderings. *Econometrica*, *54*, 707–715.

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.

Berg, R. v. d., 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.

Birnbaum, M. H. (2011). 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: 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.

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.

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.

Corbin, R. & Marley, A. (1974). Random utility models with equality: An apparent, but not actual, generalization of random utility models. *Journal of Mathematical Psychology*, *11*, 274–293.

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. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–205.

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

DeCarlo, L. T. (2013). Signal detection models for the same–different task. *Journal of Mathematical Psychology*, *57*, 43–51.

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.

Dennis, S. & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 452–478.

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.

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

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

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

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*. New York: Oxford University Press.

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

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*. New York: Wiley.

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.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton: CRC Press.

Iverson, G. J. (2006). An essay on inequalities and order-restricted inference. *Journal of Mathematical Psychology*, *50*, 215–219.

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). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*, 291–306.

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., 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)*. New York: 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., Vogt, J., & Klauer, K. C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology*, *62*, 40–53.

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.

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)*. New York: Academic Press.

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

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*. New York: 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. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.

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

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

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, Massachusett: 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.

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., Key, K. N., & Gronlund, S. D. (in press). Stimulus effects and the mediation of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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

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. & 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.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376–388.

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). Hillsdale, NJ: Erlbaum.

Myung, J. I., Karabatsos, G., & Iverson, G. J. (2008). A statistician's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 309–327). Springer.

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.

Parks, C. M. & Yonelinas, A. P. (2009). Evidence for a memory threshold in second-choice recognition memory responses. *Proceedings of the National Academy of Sciences USA*, *106*, 11515–11519.

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.

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. & Grofman, B. (1998). Choosing subsets: A size-independent probabilistic model and the quest for a social welfare ordering. *Social Choice and Welfare*, *15*, 423–443.

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.

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

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)*. New York: 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.

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). Hillsdale, NJ: Erlbaum.

Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.

Silvapulle, M. J. & Sen, P. K. (2011). *Constrained statistical inference: Order, inequality, and shape constraints*. New Jersey: John Wiley & Sons.

Spektor, M. S., Kellen, D., & Hotaling, J. M. (in press). When the good looks bad: An experimental exploration of the repulsion effect. *Psychological Science*.

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.

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.

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

Swagman, A. R., Province, J. M., & Rouder, J. N. (2015). Performance on perceptual word identification is mediated by discrete states. *Psychonomic Bulletin & Review*, *22*, 265–273.

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.* New York: Wiley.

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

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.

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

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.

Wickens, T. D. (2002). *Elementary Signal Detection Theory.* Oxford: 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., Vul, E., Mickes, L., & Wilson, B. W. (in press). Models of lineup memory. *Cognitive Psychology.*

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.