Journal of Experimental Psychology: Learning, Memory, and Cognition

Critical Testing in Recognition Memory: Selective Influence, Single-Item Generalization, and the High-Threshold Hypothesis

David Kellen, Constantin G. Meyer-Grant, Henrik Singmann, and Karl Christoph Klauer

Online First Publication, February 13, 2025. https://dx.doi.org/10.1037/xlm0001434

CITATION

Kellen, D., Meyer-Grant, C. G., Singmann, H., & Klauer, K. C. (2025). Critical testing in recognition memory: Selective influence, single-item generalization, and the high-threshold hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://dx.doi.org/10.1037/xlm0001434

Journal of Experimental Psychology:
Learning, Memory, and Cognition

ISSN: 0278-7393

https://doi.org/10.1037/xlm0001434

Critical Testing in Recognition Memory: Selective Influence, Single-Item Generalization, and the High-Threshold Hypothesis

David Kellen[1], Constantin G. Meyer-Grant[2], Henrik Singmann[3], and Karl Christoph Klauer[2]
[1] Department of Psychology, Syracuse University
[2] Department of Psychology, University of Freiburg
[3] Department of Psychology, University College London

In recent years, discussions comparing high-threshold and continuous accounts of recognition-memory judgments have increasingly turned their attention toward critical testing. One of the defining features of this approach is its requirement for the relationship between theoretical assumptions and predictions to be laid out in a transparent and precise way. One of the (fortunate) consequences of this requirement is that it encourages researchers to debate the merits of the different assumptions at play. The present work addresses a recent attempt to overturn the dismissal of high-threshold models by getting rid of a background selective-influence assumption. However, it can be shown that the contrast process proposed to explain this violation undermines a more general assumption that we dubbed "single-item generalization." We argue that the case for the dismissal of these assumptions and the claimed support for the proposed high-threshold contrast account does not stand the scrutiny of their theoretical properties and empirical implications.

Keywords: critical testing, recognition memory, signal detection theory, high-threshold models, single-item generalization

A long-debated issue in memory research is how to best characterize recognition judgments. For instance, whether these judgments should be formalized as the outcome of comparing mnemonic strength values with a response threshold—"how familiar is this?"—or instead, the manifestation of a small number of discrete states standing for complete knowledge and pure ignorance—"you either remember it or you guess" (e.g., Bröder & Schütz, 2009; Dubé & Rotello, 2012; Egan, 1958; Province & Rouder, 2012).[1] These two alternative accounts have traditionally been instantiated by the Gaussian signal detection theory (SDT) model and the two high-threshold (2HT) model, which are illustrated in Figure 1 for the case of single-item yes/no judgments. Their dispute aside, both models have been applied to a wide variety of experimental and real-world scenarios, including eyewitness-identification judgments (e.g., Winter et al., 2022; Wixted et al., 2018).

The standard approach for comparing competing models like these two involves the use of *global performance indices* that take into account how well they can describe the data *as a whole*, relative each model's complexity or flexibility (e.g., Bröder & Schütz, 2009; Dubé et al., 2012; Dubé & Rotello, 2012; Kellen et al., 2013, 2015; Klauer & Kellen, 2015; Province & Rouder, 2012). But more recent years have seen a shift in the way model comparisons are conducted, with greater emphasis being placed on *critical tests* (Chechile & Dunn, 2021; Kellen et al., 2021; Kellen & Klauer, 2014, 2015; Ma et al., 2022; Meyer-Grant & Klauer, 2021; Starns et al., 2018). This approach focuses on *specific portions* of the data for which the competing models make qualitatively distinct predictions.

Compared to global model comparisons, critical testing offers considerable advantages both in terms of *transparency* and *generality*: They establish a clear connection between theory and data, such that one can clearly *see* in the data why one of the models fails while the other succeeds (for relevant discussions, see Birnbaum, 2011; Kellen, 2019; Kellen et al., 2021). They also forgo a number of auxiliary parametric assumptions. For example, they enable one to test an SDT account without having to commit to Gaussian mnemonic-strength distributions (see Figure 1). By dispensing with these kinds of assumptions, the scope of the test is greatly expanded. Instead of dismissing a single parametric instance (e.g., Gaussian SDT model), the verdict is extended to

[1] To be clear, not all discrete-state accounts subscribe to a complete knowledge versus pure-ignorance dichotomy (see Chechile, 2018). However, such accounts are beyond the scope of the present work.

Aaron S. Benjamin served as action editor.

David Kellen https://orcid.org/0000-0003-1483-7875

Constantin G. Meyer-Grant https://orcid.org/0000-0002-5991-6596

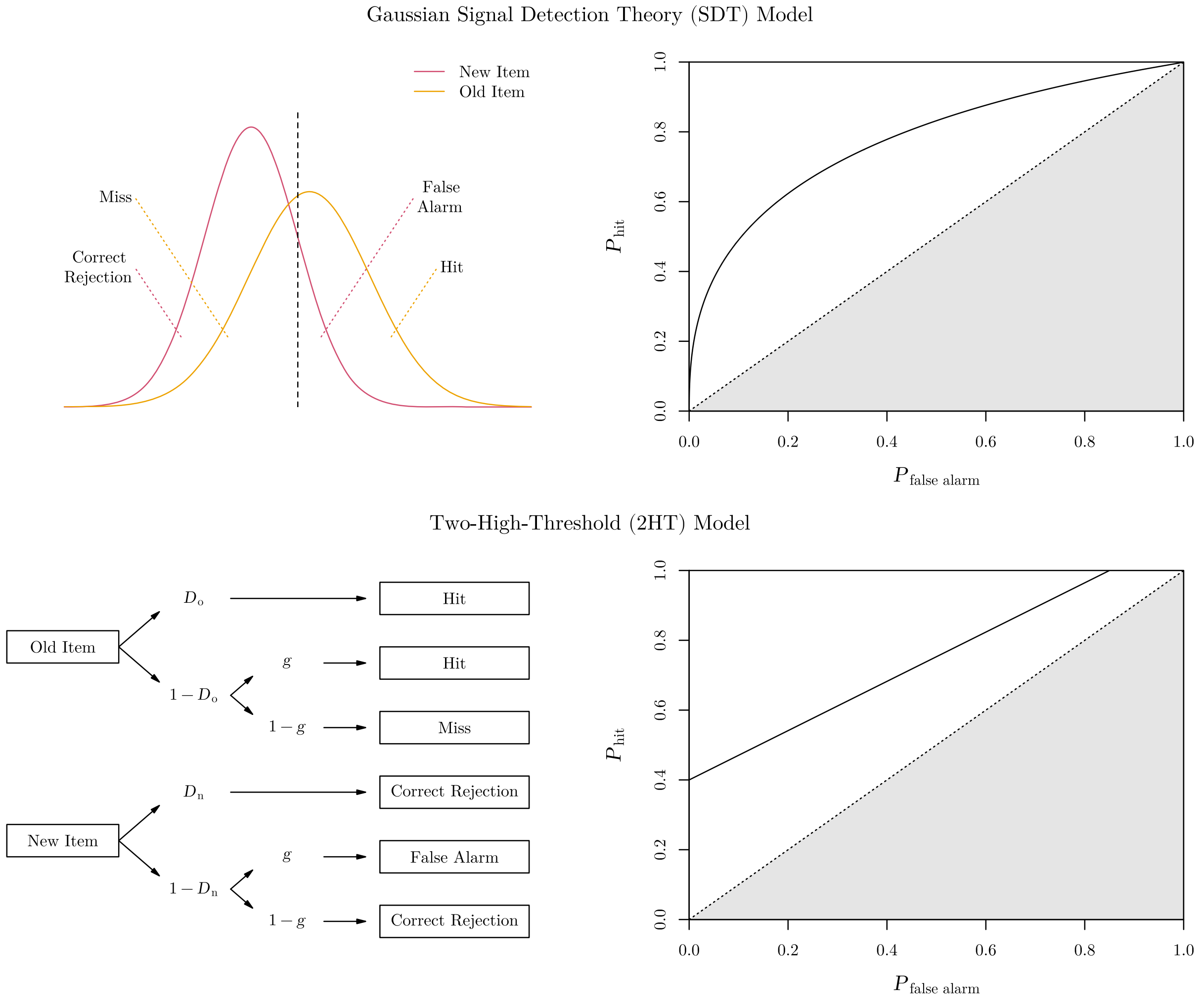
David Kellen and Constantin G. Meyer-Grant contributed equally to this work. For data and modeling scripts, see Kellen et al. (2024; https://osf.io/qz5re). David Kellen was supported by the National Science Foundation (NSF CAREER Award 2145308). The authors are grateful to Aaron Benjamin, Chad Dubé, John Wixted for their valuable comments.

David Kellen played a lead role in writing–original draft and an equal role in conceptualization and formal analysis. Constantin G. Meyer-Grant played a supporting role in writing–original draft and an equal role in conceptualization and formal analysis. Henrik Singmann played a supporting role in writing–review and editing. Karl Christoph Klauer played a supporting role in writing–review and editing.

Correspondence concerning this article should be addressed to David Kellen, Department of Psychology, Syracuse University, 765 Irving Avenue, Suite 486, Syracuse, NY 13210, United States. Email: davekellen@gmail.com

Figure 1

Illustration of Gaussian Signal Detection Model and Two High-Threshold Model



Note. Upper panels: The left panel illustrates the Gaussian SDT model, which assumes that old and new items are each associated with a Gaussian latent memory-strength distribution (with parameters μ_o, σ_o, $\mu_n = 0$, and $\sigma_n = 1$). During a recognition test, when the latent strength value of an item is greater than the response criterion (dashed line), then a “yes” recognition judgment is issued for that item. Otherwise, a “no” judgment is made. The right panel illustrates the predicted relationship between the hit and false-alarm rates (“yes” response rates to old and new items, respectively) when varying the response criterion. This relationship is commonly referred to as the receiver operating characteristic (ROC) function. Lower panels: The left panel illustrates the 2HT model, which assumes that during test, items are either in a detection state in which their true status is known. This state is reached by old and new items with probabilities D_o and D_n, respectively. With complementary probabilities $1 - D_o$ and $1 - D_n$, old and new items are said to be in an uncertainty state, such that the judgments on them are based on pure guessing, with “yes” and “no” responses being given with probability g and $1 - g$, respectively. The right panel illustrates a ROC function predicted by the 2HT model when varying guessing-probability g. See the online article for the color version of this figure.

a large family of models (e.g., SDT models with monotonic likelihood ratios; see Kellen et al., 2021).

By sharpening the comparison between competing accounts, critical tests bring to the fore the contributions of specific assumptions to the development of testable hypotheses. But they also highlight the permanent possibility of rescuing any theoretical account from recalcitrant data by adjusting said assumptions or doing away with them entirely (i.e., the Duhem–Quine thesis: see Duhem, 1954; Quine, 1951; see also Harding, 1976).

This possibility is at the center of the present work. Its goal is to evaluate a recent controversy surrounding a basal assumption in recognition-memory modeling and its role in the empirical comparison of competing 2HT and SDT accounts. The assumption at stake here is “single-item generalization.” It claims that the same latent representations (e.g., memory strength, discrete states) deployed in the characterization of single-item judgments (e.g., “Is this item old?”) can be generalized to judgments involving multiple items (e.g., “Which of these items is old?”).

Leveraging single-item generalization, D. M. Green (1960) famously proved a very general result, widely known as the *area theorem*, which shows that the accuracy rate in two-alternative forced choice (2AFC) judgments is equal to the area under the corresponding single-item ROC function (for a description of the latter, see Figure 1). In the context of recognition memory, the area theorem has been corroborated by empirical testing (Jang et al., 2009; but see Jou et al., 2016; Starns et al., 2017) and has been successfully extended to multiple-alternative forced-choice and ranking scenarios (see Kellen et al., 2012, 2021).

Single-item generalization was (indirectly) put into question by Malejka et al. (2022) when appealing the verdict of an earlier critical test by Kellen and Klauer (2014) pitting the 2HT and SDT models against each other. Specifically, Malejka et al. took issue with a *selective-influence assumption* adopted by Kellen and Klauer by arguing that it is inadequate in the context of recognition test trials where multiple items (or a tuple) are evaluated simultaneously (see also Chechile & Dunn, 2021).[2] According to Malejka et al., there is a comparative element in multiple-item judgments that fundamentally sets it apart from its single-item counterpart. When discussing the 2HT-based characterization of these judgments, they state that:

> While old–new recognition requires the evaluation of individuals items, the ranking task requires comparing multiple items within one trial (i.e., the items in the current test display). Hence, the *two high-thresholds must operate on the item tuple's familiarity contrast* [emphasis added] and not on an individual item's memory strength. In our opinion, this assumption is more in line with multiple-item discrimination than assuming that items are processed in isolation. Hence, single-item and multiple-item recognition tasks are quite different, and thus can and should require different process (and measurement) models. (p. 18)

These theoretical considerations have important implications that are not limited to the specific dispute between 2HT and SDT accounts. Beyond selective influence in the context of Kellen and Klauer (2014), they ultimately entail a wholesale rejection of single-item generalization. More specifically, selective influence is presumed to be violated on account of single-item generalization not holding true. This state of affairs effectively blocks a main theoretical avenue for developing joint models for closely related tasks (cf. Cox et al., 2018; Schurgin et al., 2020).

The goal of the present work is to provide a critical analysis of the empirical and theoretical arguments put forth by Malejka et al. (2022). After providing the necessary background, we begin by re-evaluating their experimental and modeling results and show that, in their attempt to reappraise the 2HT model, Malejka et al. did not adequately consider the predictions of the rival SDT model. But this omission prevents them from realizing that their 2HT modelling results are to be expected when the data are generated by an SDT model (Malejka et al., 2022, Experiment 1). Moreover, a closer examination of their novel experimental designs leads to a surprising and quite interesting discovery: When 2HT model parameters are estimated from SDT-generated data, impossible values, namely *negative probabilities*, become possible (Experiment 2) or even highly likely a priori (Experiment 3). These new and remarkable predictions coming out of SDT are corroborated by these studies' results. In a perhaps unexpected development, results calling for a number of controvertible, post hoc assumptions in order to uphold the 2HT model turn out to be in line with SDT predictions derived from first principles.

We will then turn our attention to the theoretical "familiarity-contrast" high-threshold model proposed by Malejka et al. (2022) as an alternative explanation for Kellen and Klauer's (2014) results and as more plausible account of multiple-item judgments. Our analyses show that this model allows for gross violations of empirically corroborated predictions obtained from single-item generalization.[3] It is also rejected by a critical test targeting one of its basic predictions. All things considered, we find it difficult to identify compelling grounds for treating the proposed contrast mechanism as a viable alternative explanation of the results originally reported by Kellen and Klauer.

SDT and 2HT Models for Ranking Judgments

Kellen and Klauer (2014) reported two experiments implementing a critical test contrasting the SDT and 2HT models. These tests involved a ranking task alongside a manipulation of how items were encountered during the initial study phase. Specifically, during the study phase, participants studied a single list of words, with some being presented once (*weak items*) and others thrice (*strong items*). Later in the test phase, each trial was comprised of four (three) words—one old and three (two) new. Participants were made aware of this composition and were instructed to rank the words according to their belief that they were previously studied (for an illustration, see Figure 2). The dependent variable of interest here was the ranking distribution of old words, and how this distribution differs between sets including weak and strong words.

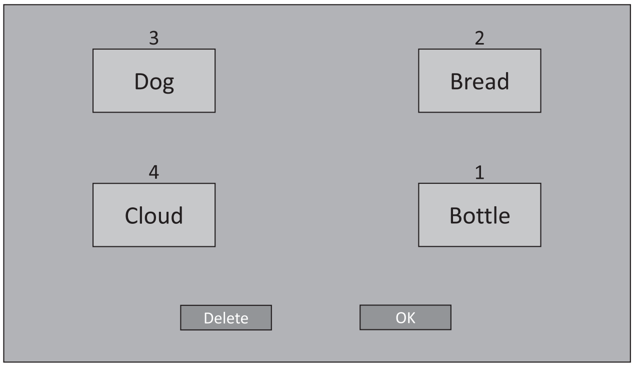
In typical studies on "study-strength" effects (e.g., Aytaç et al., 2024; Ratcliff et al., 1990; Stretch & Wixted, 1998), weak and strong items are often studied and tested separately, or are associated with conspicuous perceptual features such as color (e.g., weak green items, red strong items; e.g., Stretch & Wixted, 1998). An important goal of these studies is to create the conditions for participants' engagement with test items to vary as a function of how said item was presumably studied (e.g., "is this a strong item or is it new?"). Kellen and Klauer (2014) purposely deviated from these experimental designs due to their ability to introduce systematic changes in the recognition of new items; their critical test hinged on the study-repetition manipulation *selectively influencing* the recognition of old items. Accordingly, their experimental designs were intended to maximize the plausibility of said assumption and rule out well-known alternative accounts such as "differentiation" that would be relevant if weak and strong items were studied in different lists or explicitly labeled as weak or strong at test (e.g., Criss, 2006). This is why previously published data, such as Chechile et al. (2012), were not subjected to reanalysis—their experimental designs rendered the assumption of selective influence implausible. That being said, these experimental-design considerations were *not* spelled out by Kellen and Klauer, although they were discussed in later work, when

[2] Malejka et al. (2022) referred to it as the "lure-detection invariance" assumption, in direct reference to parameter D_n of the 2HT model. We prefer using the term *selective influence* as it can be used more generally. One should note, however, that the precise technical meaning of the latter term ultimately depends on the model that is being referred to.

[3] To be absolutely clear about this, Malejka et al. (2022) themselves do not question or even directly refer to single-item generalization. Nevertheless, its rejection is a logical consequence of their theoretical proposal. One of the goals of the present work is to unveil this issue.

Figure 2

Illustration of a Four-Alternative Ranking Trial in Kellen and Klauer's (2014) Experiment 1



Note. In each ranking trial, one of the presented items is old (e.g., Bread) and three are new. Ranks are assigned to items by clicking on them (i.e., the first item clicked on receives Rank 1, etc.). In this illustration, ranks have been assigned to all items as indicated by the numbers (i.e., ranks) shown above each item.

identifying the requirements for implementing a related critical test using published data (e.g., see Kellen & Klauer, 2015, pp. 547).

The SDT model assumes that the probability $R_{i,K}$ that an old item among K alternatives is assigned rank $i = 1, \ldots, K$, corresponds to the probability of the item's latent-strength value being the ith largest one (without the possibility of ties). The latent strength associated with each option is assumed to be an independent sample from each one's respective latent distribution, with all $K - 1$ new items being associated with the same distribution. Formally,

$$R_{i,K} = \binom{K-1}{i-1} \int_{-\infty}^{\infty} f_{\mathrm{o}}(x) F_{\mathrm{n}}(x)^{K-i} (1 - F_{\mathrm{n}}(x))^{i-1} \mathrm{d}x, \quad (1)$$

where f_{o} is the probability density function of old-item latent strength and F_{n} is the cumulative distribution function of new-item latent strength.

In turn, the 2HT model assumes that the ranking of the old item is a function of whether it is detected or not, and in the latter case, how many new items were detected. Specifically, it is assumed that, if the old item is detected with probability D_{o}, then it is assigned Rank 1. If the old item is not detected, with complementary probability $1 - D_{\mathrm{o}}$, then it is said to be in an uncertainty state, along with any new item that was not actively rejected (i.e., not detected as new). In these circumstances, the p out of K test items that happen to be in this uncertainty state are randomly assigned the top p out of K possible rank positions, whereas any new items that were actively rejected are randomly assigned to the remaining bottom ranks.

The rank probabilities $R_{i,K}$ predicted by the 2HT model considered here can be conveniently formalized as follows:

$$R_{i,K} = \begin{cases} D_{\mathrm{o}} + (1 - D_{\mathrm{o}})\xi(i), & \text{if } i = 1 \\ (1 - D_{\mathrm{o}})\xi(i), & \text{if } 2 \leq i \leq K \end{cases}, \quad (2)$$

where $\xi(i)$ denotes the probability of a nondetected old item being assigned rank i, which is a function of the number of items of uncertain status. The version of the 2HT model considered by Kellen and Klauer (2014) does not make an explicit reference to D_{n}. The reason being that it is possible—under selective-influence assumptions—to derive testable predictions without making any claims about the stochastic relationship between new-item detections; for example, that they are independent (for a similar approach, see Chechile & Dunn, 2021). That being said, one could have assumed that new-item detections are independent and identically distributed with probability D_{n}. In that case (see also Malejka et al., 2022, Equation 3)[4]

$$\xi(i, D_{\mathrm{n}}) = \sum_{j=1}^{K} \binom{K-1}{j-1} D_{\mathrm{n}}^{K-j} (1 - D_{\mathrm{n}})^{j-1} \frac{1}{j}. \quad (3)$$

As previously discussed, Kellen and Klauer (2014) assumed that the study-repetition manipulation utilized in their studies selectively influenced the remembering of old items. This assumption was deemed plausible in the specific context of their experimental designs. In terms of both models, this assumption means that parameters μ_{o} and D_{o} were presumably influenced by the manipulation, whereas other parameters such as D_{n} remained invariant.

Under the assumption that selective influence holds, both the SDT and 2HT models can be shown to make distinct predictions. Let $c_{2,K}$ denote the conditional probability that the old item is assigned Rank 2 out of K, given that it was *not* assigned Rank 1,

$$c_{2,K} = \frac{R_{2,K}}{1 - R_{1,K}}. \quad (4)$$

The SDT model then, under a wide range of parametric assumptions, expects c_2 to be greater for strong items than for weak items; that is, $c^{\mathrm{w}}_{2,K} \leq c^{\mathrm{s}}_{2,K}$ (for a proof, see Kellen & Klauer, 2014, supplemental materials). In words, the SDT model expects the latent strengths of strong items to be somewhat larger than their weak counterparts, even when not assigned Rank 1.

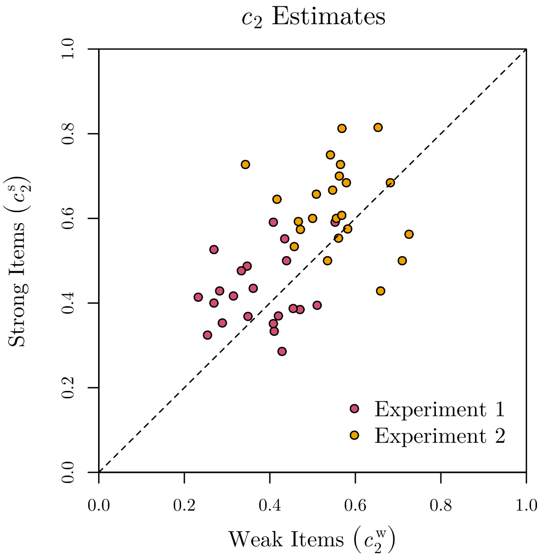
In contrast, the 2HT model with selective influence in force expects $c^{\mathrm{w}}_{2,K} = c^{\mathrm{s}}_{2,K}$. It is easy to see why these two models make distinct predictions: The 2HT model assumes that items, when their status is uncertain, are all alike in the sense that the response distributions associated to them are one and the same (for a discussion, see Rouder & Morey, 2009). In this case, the ranking of the old item, regardless of its study conditions, will be determined by a *pure guessing* process that is applied to all the test options that were not actively rejected.

Across two experiments, Kellen and Klauer (2014) found $c^{\mathrm{s}}_{2,K}$ estimates to be greater than their $c^{\mathrm{w}}_{2,K}$ counterparts, a result that speaks for the SDT model and against the complete information loss—guessing for nondetected items—postulated by the 2HT model (see Figure 3). This result has since then been replicated by McAdoo and Gronlund (2016), McAdoo et al. (2019), and more recently by Malejka et al. (2022).

[4] Note that it is presumed that guessing-based rankings are equiprobable among the j items in an uncertainty state; for each such item, the ranking-assignment probability is $\frac{1}{j}$. This follows from the assumption that recognition judgments for nondetected items are based on pure guesses.

Figure 3

Individual c_2 Estimates Obtained by Kellen and Klauer (2014) for Trials in Which a Weak (c_2^{w}) or Strong Old Item (c_2^{s}) Were Included



Note. Experiment 1 used four alternatives (for an illustration, see Figure 2), whereas Experiment 2 used three alternatives instead of four. See the online article for the color version of this figure.

Comparing Models With Forced-Choice and Ranking Judgments

In an attempt to put selective influence in the 2HT model to the test, Malejka et al. (2022) conducted a number of studies in which they found D_{n} estimates to be affected by the kind of study-strength manipulation used by Kellen and Klauer (2014). Specifically, they estimated D_{n} to be greater among ranking trials that include strong items compared to those that included weak items; that is, $D_{\mathrm{n}}^{\mathrm{w}} \leq D_{\mathrm{n}}^{\mathrm{s}}$. Their first study was basically a replication of Kellen and Klauer's (2014) Experiment 1, with participants encountering four-alternative ranking judgment trials that included either weak or strong items. Two follow-up studies implemented more complex experiment designs in which participants encountered *sequences of forced-choice and ranking judgment trials*. Across all three studies, the testing of selective influence followed the same strategy: 2HT models were fit to ranking or ranking and forced-choice judgments, and the restriction $D_{\mathrm{n}}^{\mathrm{w}} = D_{\mathrm{n}}^{\mathrm{s}}$ was evaluated.

Before going into the details of each test conducted by Malejka et al. (2022), some general considerations are in order. First, let us take a moment to consider the point of contention that motivated their investigation: The possibility that Kellen and Klauer (2014) mistook a violation of selective influence in the 2HT model for evidence for the rival SDT model. This possibility stems from the fact that a 2HT model violating the selective-influence assumption, such that $D_{\mathrm{n}}^{\mathrm{w}} \leq D_{\mathrm{n}}^{\mathrm{s}}$, is able to accommodate the observed pattern of $c_{2,K}^{\mathrm{w}} \leq c_{2,K}^{\mathrm{s}}$. But this is only one side of the coin: There is also the possibility that ranking data generated by SDT, when fitted by the 2HT model, will look like a violation of selective influence, with $D_{\mathrm{n}}^{\mathrm{w}} \leq D_{\mathrm{n}}^{\mathrm{s}}$. Coherence therefore demands that the concerns raised about Kellen and Klauer (2014) also extend to follow-up investigations attempting to test selective influence in 2HT model empirically. More specifically, any such investigation must come to terms with the existing ambiguity and provide assurances that it does not confuse evidence for a given model-based hypothesis (D_{n} varies with study strength) with the success of a rival model (SDT is the data-generating account). To help the reader understand what is at stake here, Appendix A provides an analogous but more familiar scenario involving ROC data. In this scenario, we show how the observation of nonlinear ROCs, when seen through a 2HT lens, are to be interpreted as evidence against the selective influence of response-bias manipulations, putting into question any model comparison based on ROC data (e.g., Kellen et al., 2013; Malejka & Bröder, 2019).

As will become clear below, the empirical investigation conducted by Malejka et al. (2022) does not live up to the very standard that motivated it in the first place. This issue is manifested in the fact that all of their results supporting the claim that $D_{\mathrm{n}}^{\mathrm{w}} \leq D_{\mathrm{n}}^{\mathrm{s}}$ *are expected when fitting 2HT models to SDT-generated data*. In other words, there is no privileged model-based characterization; they are ambiguous or nondiagnostic in that sense.

When faced with this kind of ambiguity, researchers are expected to provide additional arguments supporting their favored interpretation. As discussed earlier, in their critical tests, Kellen and colleagues appealed to the characteristics of the experimental design to make a case for selective influence (see Kellen et al., 2021; Kellen & Klauer, 2014, 2015). Similar appeals have been deployed when discussing ROC data and the presumption of selective influence in response-bias manipulations (see Bröder & Malejka, 2017; Malejka & Bröder, 2019). But for the present case of ranking judgments, Malejka et al.'s (2022) criticism is *not* driven by some perceived experimental oversight on Kellen and Klauer's (2014) part but rather by a *theoretical claim* about the way people make multiple-item judgments (see their quote earlier). This means that, in the face of ambiguous or nondiagnostic results, one should carefully examine the virtues of the theoretical account being offered (e.g., Kellen et al., 2018). We will pursue this route later on by conducting a thorough evaluation of the "contrast 2HT modeling" account proposed by Malejka et al. We will show that it makes a number of implausible predictions and fails a critical test. Altogether, we see no good reason to find this proposed account viable, let alone grant it some kind of privileged status over any other.

Moreover, in the context of the new experimental paradigm introduced by Malejka et al. (2022), it is not always the case that SDT-generated data are consistent with the 2HT model in its principal variant. As we will discuss in detail below, in the experimental designs combining forced-choice and ranking judgments (Experiments 2 and 3), there is the possibility of obtaining *negative* D_{n} estimates, which are obviously nonpermissible as these parameters refer to probabilities. In one particular experimental design (Experiment 3), negative estimates are even expected under many circumstances, and this is exactly what is found. Rather than vindicating the viability of the 2HT model (or at least the archetypal 2HT model without post hoc extensions), the experimental designs developed by Malejka et al. (2022) turn out to reveal a novel critical test that corroborates SDT predictions derived from first principles.

Ranking Judgements

Malejka et al.'s (2022) first approach to testing selective influence consisted of estimating D_{n} from four-alternative ranking judgments, separately for cases with weak and strong items. The parameter estimates from their first experiment showed that, indeed, D_n^s was generally greater than D_n^w. However, these estimates by themselves cannot be taken as evidence against selective influence. After all, they are driven by the observation that $c_{2,4}^w \leq c_{2,4}^s$; that is, the empirical challenge faced by the 2HT model in the first place. And a result that is expected under the rival SDT account.

To express this issue in the simplest possible manner, it is useful to start by framing our discussion in the context of a *three-alternative* ranking task. According to the 2HT model, the corresponding ranking probabilities are:

$$\begin{aligned} R_{1,3} &= D_o + (1 - D_o)D_n^2 + (1 - D_o)D_n(1 - D_n) \\ &\quad + \frac{(1 - D_o)(1 - D_n)^2}{3}, \\ R_{2,3} &= (1 - D_o)D_n(1 - D_n) + \frac{(1 - D_o)(1 - D_n)^2}{3}, \\ R_{3,3} &= \frac{(1 - D_o)(1 - D_n)^2}{3}, \\ &\text{with} \\ c_{2,3} &= \frac{R_{2,3}}{R_{2,3} + R_{3,3}}. \end{aligned} \tag{5}$$

With some rearrangement, we can see that D_n is directly determined by $c_{2,3}$,

$$D_n = \frac{1 - 2c_{2,3}}{c_{2,3} - 2}, \tag{6}$$

which can then be used to determine D_o. These results show that ranking probabilities—provided that certain accuracy constraints are satisfied—can be directly translated into D_o and D_n values. These constraints are simply that ranking probabilities for the old item cannot drop below chance nor reach ceiling; that is, $1/3 \leq R_{1,3} < 1$ and $c_{2,3} \geq 1/2$. It follows then that any two conditions yielding different $c_{2,3}$ values necessarily imply different D_n values. For as shown in Equations 5 and 6, *according to the 2HT model*, $c_{2,3}$ *is simply a function of* D_n *and vice versa* (see also Kellen & Klauer, 2011).

Now, consider once again the critical-test results originally reported by Kellen and Klauer (2014). Their report that $c_{2,K}^s$ is generally greater than $c_{2,K}^w$ can be trivially handled by simply allowing D_n to differ between the two conditions. What this means is that observing differences in D_n across conditions does not provide any kind of novel insight, regardless of how many times these results are replicated and the 2HT model is fit to it. All that this course of action achieves is reiterate the known formal relationship between model parameters and data. And reiterating this relationship does not affect the plausibility of selective influence given that $c_{2,K}^w \leq c_{2,K}^s$ is natively expected by the rival SDT.[5]

To illustrate this point, consider an artificial scenario where data are generated from an unequal-variance Gaussian SDT model with $\sigma_o^2 = 1.3$. As shown in Table 1, changes in the mean of the latent old-item distribution (μ_o) affect both $R_{1,3}$ and $c_{2,3}$. According to the 2HT model, these effects directly map onto D_o and D_n, both of which increase alongside μ_o (see "First Method" column). In this toy example, we know for a fact that only the old-item distribution is changing—a simple story of selective influence. But according to Malejka et al.'s (2022) reasoning, the empirical outcomes of this scenario are to be interpreted as speaking against selective influence in the 2HT model and therefore against the validity of the original critical test (see also Appendix A).

Table 1

SDT-Generated Predictions and Resulting 2HT Parameters Estimates

SDT generation						2HT estimation			
Parameter		Prediction				First method		Second method	
μ_o	σ_o^2	$R_{1,3}$	$c_{2,3}$	$P_{n,on}^{\ominus A}$	$P'^{\ominus B}_{n,on}$	D_o	D_n	A: D_n	B: D_n
0.2	1.3	.40	.51	.39	.25	.09	.01	−.22	−.56
0.4	1.3	.46	.55	.42	.27	.15	.06	−.15	−.51
0.6	1.3	.51	.58	.46	.29	.22	.11	−.08	−.47
0.8	1.3	.57	.62	.49	.31	.28	.17	−.02	−.43
1.0	1.3	.62	.65	.53	.32	.35	.22	.05	−.39
1.2	1.3	.68	.68	.56	.34	.41	.27	.12	−.35
1.4	1.3	.72	.71	.59	.35	.48	.32	.18	−.32
1.6	1.3	.77	.73	.62	.37	.54	.37	.24	−.29
1.8	1.3	.81	.76	.65	.38	.60	.42	.30	−.26
2.0	1.3	.85	.78	.68	.39	.65	.46	.36	−.24
2.2	1.3	.88	.80	.71	.40	.70	.51	.41	−.22
2.4	1.3	.90	.82	.73	.41	.75	.55	.46	−.19

Note. SDT parameters (μ, σ_o^2) used to generate predictions $R_{1,3}$, $c_{2,3}$, $P_{n,on}^{\ominus A}$, and $P'^{\ominus B}_{n,on}$. The 2HT estimates obtained from the ranking (First Method) and forced-choice predictions (D_n; with Second Method Versions A and B). $R_{1,3}$ = probability of old item being assigned Rank 1 given three alternatives. $c_{2,3}$ = probability of old item being assigned Rank 2, given that it was not assigned Rank 1. $P_{n,on}^{\ominus A}$ and $P'^{\ominus B}_{n,on}$ = correct forced-choice probabilities in cases where the old item was not Ranked 1 (Versions A and B of the Second Method, respectively). SDT = signal detection theory; 2HT = two high-threshold.

The running three-alternative ranking task example is useful because it renders the relationship between parameters and data as transparent as it could be. That being said, Malejka et al.'s (2022) Experiment 1 rankings did not involve three alternatives *but four*. Their experimental design with four-alternative ranking trials provides 3 + 3 = 6 degrees of freedom, more than enough to estimate the four 2HT parameters D_o^w, D_o^s, D_n^w, and D_n^s (see Equation 3). The remaining two degrees of freedom were used to test the model's goodness of fit, which was generally found to be satisfactory. Model misfits were statistically significant ($p < .05$) for only 9% of the participants. In contrast, the three-alternative ranking example only provides the four degrees of freedom necessary for parameter estimation, which means that it does not offer the same opportunities for model corroboration. The question then is whether going from three to four alternatives is of any consequence for the point that we are trying to make here—*no, not really*. First, note that it is still the case that $c_{2,4}$ (or $c_{3,4}$ for that matter) is solely a (strictly increasing) function of new-item detection (see Equations 2 and 3; see also Equation 4 in Malejka et al., 2022):

$$c_{2,4} = \frac{\xi(2,D_n)}{\xi(2,D_n) + \xi(3,D_n) + \xi(4,D_n)}. \tag{7}$$

This relationship implies that, as in the three-alternative case, the mere observation that c_2 is greater for strong than weak items already calls for an estimate of D_n^s that is greater than its D_n^w counterpart.

[5] Confusions between logical/conceptual relations and empirical evidence are quite common in psychology, often taking very subtle forms (see M. A. Wallach & Wallach, 1994, 1998; L. Wallach & Wallach, 2010).

But what about the goodness-of-fit tests of the 2HT model as a whole? Could they help overcome the ambiguity surrounding the $c_{2,4}$ differences? Even though the four-alternative ranking design provides enough degrees of freedom to conduct a test of the 2HT model, its power is too low for the test results to have any real diagnostic value, as data coming from SDT can be accommodated by the 2HT model. To see this, let us simulate data from a unequal-variance Gaussian SDT model, using the same number of test trials (75 per condition) conducted in Experiment 1 of Malejka et al. (2022). Individual μ_o^w and μ_o^s parameters were sampled from a uniform distribution ranging between 0.25 and 2 (these samples were ordered to ensure that $\mu_o^w < \mu_o^s$). In turn, a common σ_o parameter was sampled from a uniform distribution ranging between 1 and 1.5. A total of 5,000 individual samples were taken. Out of these, the 2HT model only failed to fit 13% of them (with $p <$ .05). For 81% of the individual cases, D_o^s was estimated to be greater than D_o^w, with median estimates of .45 and .27, respectively. In turn, D_n^s was estimated to be larger than D_n^w in 75% of the individual cases, with medians .29 and .12, respectively. Altogether, these simulation results show that the basic pattern of results put forth by Malejka et al. as vindicating the viability of the 2HT model is essentially what you would expect to see if the data originated from the rival SDT model.

At this point, one could object that our analysis is one-sided by always taking SDT to be the true data-generating model. Why not also consider an alternative scenario where a 2HT model (sans selective influence) takes on that role? Simply because the question here is whether we might be fooling ourselves when drawing conclusions specific to 2HT model (does selective influence hold?) from these studies. And you address that question by considering cases where the 2HT model is *not* generating the data.

Forced-Choice-Then-Ranking Judgements

In Experiments 2 and 3, Malejka et al. (2022) relied on a somewhat more involved method for estimating D_n, relying on a selection of forced-choice judgments and subsequent ranking judgments that included the same test items. Let $P_{n,on}$ denote the probability of *correctly choosing the new item* in a pair that includes one old item and one new item. According to the 2HT model

$$\begin{aligned} P_{n,on} &= D_o + (1 - D_o)D_n + (1 - D_o)(1 - D_n)\tfrac{1}{2} \\ &= 1 - \tfrac{(1-D_o)(1-D_n)}{2}. \end{aligned} \tag{8}$$

Now, consider the same correct choice probability, only that this time we are restricting ourselves to the cases in which the old item was *not* detected,

$$P_{n,on}^{\ominus} = 1 - \frac{(1 - D_n)}{2}. \tag{9}$$

This choice probability, if available to the researcher, would provide a direct estimate of parameter D_n,

$$D_n = 2P_{n,on}^{\ominus} - 1. \tag{10}$$

But how can the researcher identify trials in which old-item detection failed? Consider follow-up four-alternative ranking trials that include the old items previously encountered. Under the assumption that the detection status of items is carried over across tests (which we will take for granted), every instance in which the old item is not assigned Rank 1 is also one in which it was not detected. Hence, these cases can be used to estimate $P_{n,on}^{\ominus}$ and thus D_n.

Malejka et al. (2022) implemented two versions of this method in their Experiments 2 and 3. We will refer to these versions as *A* and *B*. In Version *A* (Experiment 2), participants first engaged in the aforementioned 2AFC judgments. In a subsequent test block, participants were presented with four-alternative ranking trials. Each ranking trial was comprised of an old item encountered in the previous test block *alongside three novel new items*. In Version *B*, each forced-choice trial was immediately followed by a four-alternative ranking trial that included *both items* from the previous trial (i.e., only two of the three new items were novel). The two experimental designs are illustrated in Figure 4.

The D_n estimates obtained with version *A* replicated the basic pattern found in Experiment 1, with D_n^s being generally greater than D_n^w. At first glance, this result can be seen as providing corroborating evidence for the violation of selective influence. But as before, one needs to provide assurances that this result is not ambiguous in the sense of not being obtained if the data were generated by a SDT model. In short, *is SDT expected to produce these results? Once again, the answer is yes*.

Taking for granted the assumption that the familiarity values are precisely the same across forced-choice and ranking trials, SDT establishes $P_{n,on}^{\ominus A}$ as the probability that the old item, when not assigned Rank 1 in a *K*-alternative forced choice trial, has a greater familiarity than the new item included in the forced-choice trial. Formally, this corresponds to

$$P_{n,on}^{\ominus A} = \frac{\int_{-\infty}^{\infty} f_o(x)F_n(x)(1 - F_n(x)^{K-1})\mathrm{d}x}{\int_{-\infty}^{\infty} f_o(x)(1 - F_n(x)^{K-1})\mathrm{d}x}, \tag{11}$$

with $K/2(K + 1) \leq P_{n,on}^{\ominus A} \leq 1$.[6]

As shown in Table 1, for the case of three-alternative rankings, $P_{n,on}^{\ominus A}$ increases as a function of μ_o, which in turn entails that $D_n^w \leq D_n^s$ when data are generated by an SDT model. From this relationship, which also holds for four alternatives, it follows that the observation that $D_n^w \leq D_n^s$ is not diagnostic, when trying to overturn the dismissal of the 2HT model over SDT reported by Kellen and Klauer (2014).

More interestingly, Table 1 also shows how SDT can predict $P_{n,on}^{\ominus A}$ values *below* 1/2. These below-chance values, when plugged into Equation 10, result in *negative* D_n values. At first glance, it might seem strange that the two models differ on the lower bound of $P_{n,on}^{\ominus A}$ and that SDT can in fact predict below-chance performance. But a closer look shows how this difference stems from the unique ways in which the two models characterize recognition judgments.[7]

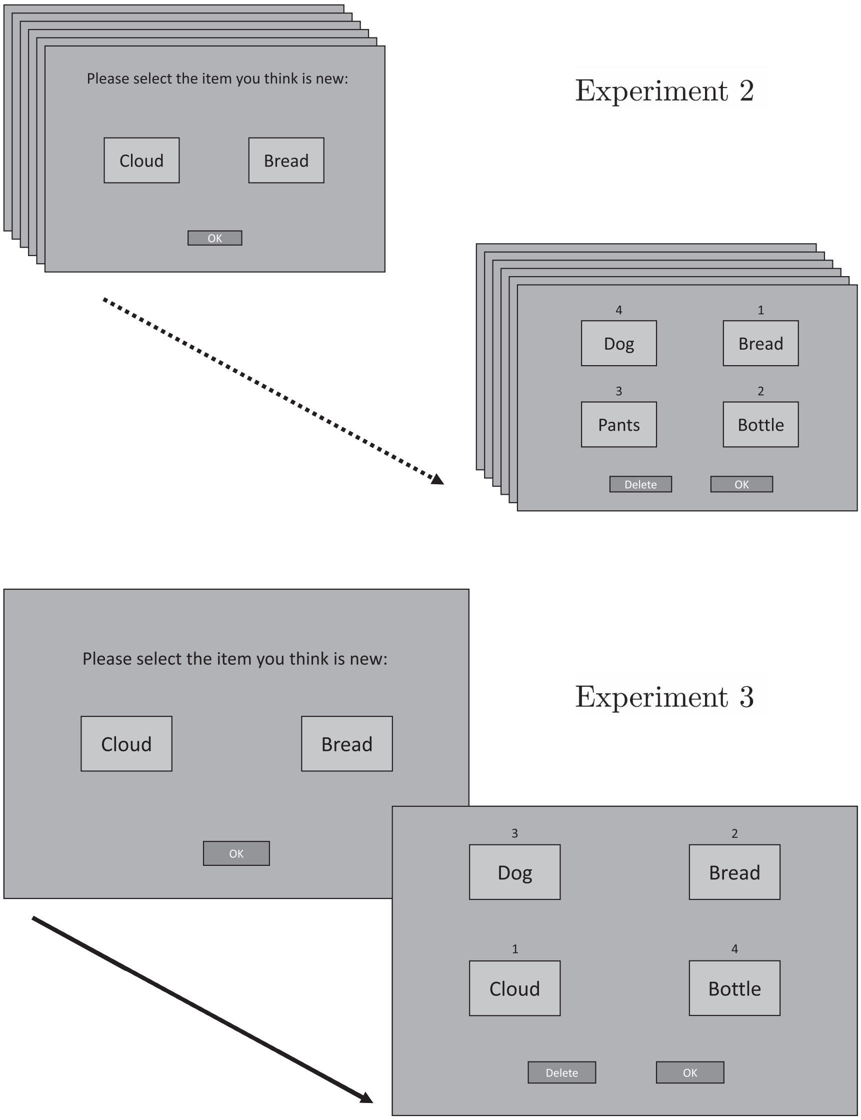
The 2HT model postulates that all nondetected items—old and new alike—are subjected to the same guessing processes without

[6] For the lower bound, we assume that the latent-strength distribution of old items is identical to the new items'. That is, we are purposely excluding cases found in the unequal-variance Gaussian SDT model where the model makes unreasonable below-chance predictions (e.g., $\mu_o = 0$ and $\sigma_o > 1$).

[7] Contrary to many other circumstances where it is the culprit, this prediction of below-chance accuracy by SDT is *not* caused by the use of an unequal-variance Gaussian parametrization (see also Footnote 6; cf. Kellen & Klauer, 2011).

Figure 4

Illustration of the Sequences of Forced-Choice and Ranking Trials Used by Malejka et al. (2022) Experiments 2 and 3, Implementing Versions A and B of Their Second Method



distinction. This assumption is commonly referred to as *conditional independence* (e.g., Kellen et al., 2013; Province & Rouder, 2012). In a ranking task, this assumption establishes that there is no difference between a nondetected item that is assigned Rank 2 and another one assigned Rank 3, for instance—they are all guesses (see Equation 2). In the forced-choice trial, the worst-case scenario is when neither item is detected, which means that the response will be a guess, with 1/2 probability of being correct (see Equation 9).

For SDT, both ranking and forced-choice judgments are based on the latent strength of each item and how they compare to each other. When the old items are assigned Ranks 2 and 3, this means that, in their respective ranking trials, they had the second and third highest latent strength, respectively. In contrast to the 2HT model, these items are *not* treated alike in the case of ranking. In fact, the expected latent strength of the former item is greater than the latter's. To see this, consider the example in the first row of Table 1, in which the mean or expected latent strength of old items (μ_o) is 0.2. Conditional on rank *k*, what is the expected latent strength? For Ranks 1–3 out of three, the means are 1.09, 0.08, and −0.89, respectively. Aside from their decreasing order, note how the latter value is negative, far below the expected latent strength of new items, which is set to zero by convention (but without loss of generality). In these instances where the old item is assigned Rank 3, the probability of its latent strength being greater than a novel new item's latent strength—the new item paired with it in the forced-choice trial—is just .25.

Figure 5 illustrates the results obtained in Malejka et al.'s (2022) Experiment 2, which implemented Version *A* of their method. Most of the $P^{\ominus A}_{\mathrm{n,on}}$ estimates obtained are above 1/2 (left panel), which correspond to positive D_n estimates (right panel). These results are not that diagnostic given that the range of outcomes permitted by the 2HT model are nested within SDT's, with only a small region of disagreement (see left panel). Regardless, it is by now clear that the combination of forced-choice and ranking judgments is able to lay bare the unique aspects of the two models. The question is whether there is an experimental design that can amplify them so that divergent predictions are made in almost all possible circumstances. It turns out that Version B, implemented in Experiment 3, delivers on this promise.

In Version B, the old–new item pairs presented in the forced-choice trials appear together once again in the ranking trials (for an illustration, see Figure 4). Again, the 2HT model predicts $P^{\ominus B}_{\text{n, on}}$ to be determined by new-item detection and guessing, with a lower bound of 1/2. Malejka et al. (2022) used Equation 10 to estimate D_{n} from the rate of correct responses in the 2AFC task, given the old item was not ranked first in the subsequent ranking task. However, for Version B, this rate—which we here denote by $P'^{\ominus B}_{\text{n, on}}$—is *not* identical to $P^{\ominus B}_{\text{n, on}}$. The reason for this is that the probability of the target being ranked first depends on whether the new item that appears in both tasks was correctly rejected (detected as new) or not. If the new item was correctly rejected by the participant, for example, it is more likely that the target will be ranked first than if the new item was not rejected. At the same time, rejecting this new item guarantees a correct response in the corresponding 2AFC task. As a result, the relationship between $P'^{\ominus B}_{\text{n, on}}$ and D_{n} is more complex than between $P^{\ominus B}_{\text{n, on}}$ and D_{n}. Unfortunately, there is no way to directly estimate $P^{\ominus B}_{\text{n, on}}$ from the data of Malejka et al.'s (2022) Experiment 3. But presupposing the possibility of non-detected items being freely rearranged between the forced-choice and ranking judgments and $K = 4$, it can be shown that the 2HT model predicts

$$P'^{\ominus B}_{\text{n, on}} = \frac{7}{6} - \frac{2}{(D_{\text{n}})^2 + 2D_{\text{n}} + 3}. \tag{12}$$

The lower bound of $P'^{\ominus B}_{\text{n, on}}$ is still 1/2.[8] This lower bound holds irrespective of the number of alternatives in the ranking task since there is no other option but to guess in the 2AFC task when neither the old item nor the new item are correctly detected.[9] As per Equation 12, an estimate of the parameter D_{n} based on $P'^{\ominus B}_{\text{n, on}}$ (with $K = 4$) is given by

$$D_n = \frac{\sqrt{-72(P'^{\ominus B}_{\text{n, on}})^2 + 96P'^{\ominus B}_{\text{n, on}} - 14}}{7 - 6P'^{\ominus B}_{\text{n, on}}} - 1. \tag{13}$$

In turn, according to SDT, $P'^{\ominus B}_{\text{n, on}}$ corresponds to the probability that the latent strength of the old item is above the new item from the forced-choice trial included in the forced-choice trial *but not all* of the remaining $K - 2$ new items included in the K-alternative ranking task. Formally:

$$P'^{\ominus B}_{\text{n, on}} = \frac{\int_{-\infty}^{\infty} f_{\text{o}}(x) F_{\text{n}}(x)(1 - F_{\text{n}}(x)^{K-2})\text{d}x}{\int_{-\infty}^{\infty} f_{\text{o}}(x)\left(1 - F_{\text{n}}(x)^{K-1}\right)\text{d}x}, \tag{14}$$

with $K - 2/2(K - 1) \leq P'^{\ominus B}_{\text{n, on}} \leq K - 2/K - 1$.[10]

Looking back at Table 1 and its three-alternative ranking example, we see once again how SDT-generated data, when fed through Equation 13, replicates the $D^{\text{w}}_{\text{n}} \leq D^{\text{s}}_{\text{n}}$ pattern found by Malejka et al. (2022). But more importantly, we see that SDT yields $P'^{\ominus B}_{\text{n, on}}$ smaller than 1/2 for the entire range of performance levels considered there. That is, for the three-alternative ranking case, all estimated $D_{\text{n}} \leq 0$. These predictions are not specific to the Gaussian parametrization of SDT, as 1/2 *is the upper bound of* $P'^{\ominus B}_{\text{n, on}}$. The reason for this upper bound is straightforward: Given that the old item was not assigned Rank 1, *at least* one of the new items must have surpassed the old item in terms of latent-strength values. In the case of three alternatives, it follows that the probability that the old item is ranked higher than the new item that it is paired against in the forced-choice trial is 1/2 at best.

Given that there is no overlap between the two models' permissible outcomes, aside from the single point $P'^{\ominus B}_{\text{n, on}} = 1/2$, Version B with three-alternative rankings can be said to be a maximally diagnostic experimental design. In comparison, an experimental design with four-alternative rankings, as implemented by Malejka et al. (2022) in their Experiment 3, is slightly less diagnostic: As illustrated in the left panel of Figure 6, SDT postulates an upper bound of $P'^{\ominus B}_{\text{n, on}} = 2/3$, which leads to a small overlap with the 2HT model's permissible outcomes. That being said, SDT still makes a clear prediction for the experimental design with four-alternative rankings: frequent instances where $P'^{\ominus B}_{\text{n, on}} \leq 1/2$. It is also worth noting that the study-strength manipulation does not play a role in these predictions, although it can be useful as a way to expand the range of performance observed.

The $P'^{\ominus B}_{\text{n, on}}$ estimates coming from Malejka et al.'s (2022) Experiment 3 are shown in the left panel of Figure 6. For weak items, 92% of the $P'^{\ominus B}_{\text{n, on}}$ estimates are below 1/2. For strong items, 75%. Per Equation 13, all of these cases correspond to negative D_{n} estimates. Many of these estimates are well captured by the range of permissible outcomes allowed by SDT, whereas the 2HT model's respective range misses almost all of them (see Figure 6, left panel). Relative model performance, quantified via Bayes factors (see Heck & Davis-Stober, 2019), generally favored SDT's range of permissible outcomes over 2HT's, with a median individual-level Bayes factor of 17, and with values over 10 (strong relative support) obtained for 61% of the individuals.

Altogether, the results from Malejka et al.'s (2022) Experiments 2 and 3 appear to be more in line with SDT. Its basic tenets are shown to predict $P^{\ominus A}_{\text{n, on}} \leq 1/2$ within a limited range of circumstances, and that is what is found. SDT is also shown to predict $P'^{\ominus B}_{\text{n, on}}$ below 1/2 under many circumstances and, once again, that is what is found. According to SDT, it could not have been any other way. The 2HT model, on the other hand, has no native ability to accommodate these differential results.[11]

[8] In Version B, the 2HT model now also predicts an upper bound for $P^{\ominus B}_{\text{n, on}}$, which for $K = 4$ turns out to be located at 5/6.

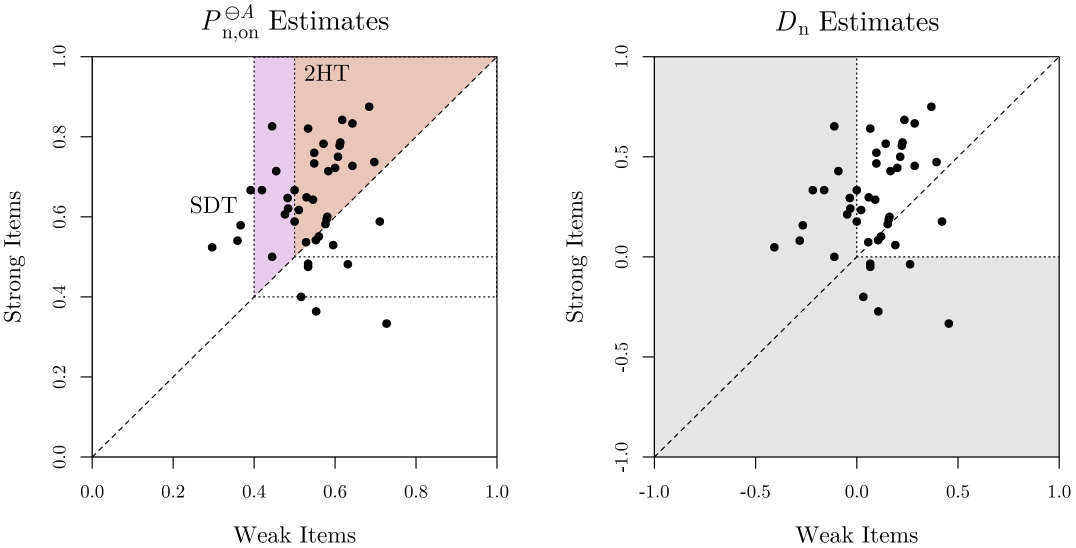
[9] Again, this presupposes that the order in the ranking task is determined by the memory states and—when conditioning on the memory states—is independent of the actual 2AFC response.

[10] For the lower bound, we again assume that the latent-strength distribution of old items is identical to the new items' (see Footnote 6).

[11] As previously stated, these predictions presume that the memory states are carried over from the forced-choice trial to the ranking trial. We considered a number of different ways in which this assumption could be violated; for example, for the SDT model, introduce noise to each item's latent strength between the two trials. Overall, the kinds of violations considered led to negligible changes in the predictions made by both models.

Figure 5

$P^{\ominus A}_{\mathrm{n,on}}$ and D_n Estimates Based on the Data From Malejka et al. (2022) Experiment 2



Note. Left panel: $P^{\ominus A}_{\mathrm{n,on}}$ estimates obtained by Malejka et al. (2022, Experiment 2). The shaded regions correspond to the values permitted by SDT (assuming $\mu^{\mathrm{w}}_{\mathrm{o}} \leq \mu^{\mathrm{s}}_{\mathrm{o}}$) and the 2HT model (assuming that $D^{\mathrm{w}}_{\mathrm{o}} \leq D^{\mathrm{s}}_{\mathrm{o}}$ and $D^{\mathrm{w}}_{\mathrm{n}} \leq D^{\mathrm{s}}_{\mathrm{n}}$). Note that the region delineated by the 2HT model *is nested inside* SDT's. Right panel: D_{n} estimates obtained in the same experiment, using Version *A* of Malejka et al.'s second estimation method. The shaded area covers nonpermissible value pairs, as D_{n} is a probability and therefore must range between 0 and 1. SDT = signal detection theory; 2HT = two high-threshold. See the online article for the color version of this figure.

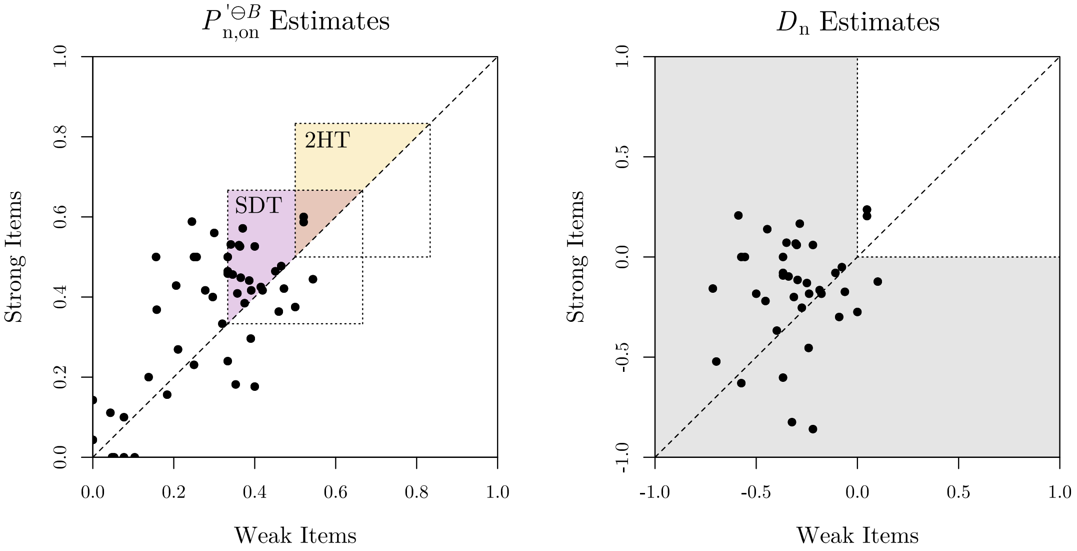
The challenges faced by the 2HT model here were acknowledged by Malejka et al. (2022). In response to the negative D_{n} estimates obtained in their Experiment 3, they proposed a post hoc revision of the 2HT model that attributes the negative D_{n} estimates illustrated in Figure 6 to an unaccounted manifestation of "response consistency." That is, participants taking for certain that the nonchosen item in a forced-choice trial is old and therefore automatically assigning it Rank 1 in the ranking trial that immediately follows—perhaps because they are trying to save time or complete the task with as little effort as possible. For instance, if participants selected the word Cloud over Bread as the new item in the forced-choice trial, as in the example illustrated in the bottom half of Figure 4, "response consistency" would entail that the latter is ranked 1 in the ranking trial that comes immediately after. To explore this possibility, Malejka et al. fitted an extended 2HT model that included the possibility of consistent responding whenever the old item was not detected, which can decrease the predicted $P'^{\ominus B}_{\mathrm{n,on}}$ below 1/2. The probability of this occurrence was quantified by a parameter *c*, which was estimated to be .46 on average.

Besides its post hoc status, the "response consistency" hypothesis as an explanation for $P'^{\ominus B}_{\mathrm{n,on}} \leq 1/2$ also has the notable feature of presuming that the kind of responding that the experiment was explicitly designed to prevent is in fact a major driving force. To wit, the forced-choice trials conducted in Malejka et al.'s (2022) Experiment 3 purposely (a) requested participants to choose the *new* item in the pair presented to them, (b) provided instructions that explicitly informed participants that new–new pairs are also possible, and (c) randomized item positions when transitioning from a forced-choice trial to a ranking trial. The whole point behind these experimental-design choices was to prevent participants from inferring that the nonchosen item in the forced-choice trial is necessarily old and expressing it in the follow-up ranking trial or simply carrying on their previous responses. Future work attempting to validate the presence of this strategy (which could be implemented in both 2HT and SDT models) should consider developing new ways to suppress it.

Interestingly, there is also an alternative interpretation for the relative success of introducing "response consistency" into the 2HT model. A core assumption of the 2HT model is that new items cannot be *misremembered* as old. This assumption has led to conspicuous problems, as documented in a number of recent studies (Meyer-Grant & Klauer, 2021; Starns, 2021; Starns et al., 2018; Starns & Ma, 2018; Voormann et al., in press). According to SDT, misremembering is not only possible but can be common under the right experimental conditions. The present case can be seen as just another instance of the same problem. In fact, if one considers what "response consistency" actually does for the 2HT model, it is easy to see that it introduces the possibility of new items being treated *as if* they were incorrectly detected (i.e., misremembered) as old. In other words, it offers a remedy for a

Figure 6

$P'^{\ominus B}_{\text{n, on}}$ and D_n Estimates Based on the Data From Malejka et al. (2022) Experiment 3



Note. Left panel: $P'^{\ominus B}_{\text{n, on}}$ estimates obtained by Malejka et al. (2022, Experiment 3). The shaded regions correspond to the values permitted by SDT (assuming μ_o^w) and the 2HT model (assuming that D_o^w and D_n^w). Right panel: D_n estimates obtained in the same experiment, using Version *B* of Malejka et al.'s second estimation method (via Equation 13). The shaded area covers nonpermissible value pairs, as D_n is a probability and therefore must range between 0 and 1. SDT = signal detection theory; 2HT = two high-threshold. See the online article for the color version of this figure.

model shortcoming that researchers working with the 2HT model have repeatedly encountered before.[12]

To evaluate the tenability of this alternative explanation for the success of the extended 2HT model, we simulated 5,000 individual data sets from a Gaussian SDT model and fitted them with the extended 2HT model. Individual μ_o^w and μ_o^s parameters were sampled from a uniform distribution ranging between 0.25 and 2 (these samples were ordered to ensure that $\mu_\text{o}^\text{w} < \mu_\text{o}^\text{s}$), and a common σ_o parameter was sampled from a uniform distribution ranging between 1 and 1.5.[13] Out of these 5,000 individual data sets, the extended 2HT model produced statistically significant misfits ($p < .05$) for 9% of them. Moreover, D_n^s was estimated to be greater than D_n^w in 69% of the cases, and imposing the restriction $c = 0$ led to statistically significant increases in misfit in 55% of the times. Once again, the general pattern of 2HT modeling results observed by Malejka et al. (2022) is found when generating data from SDT.

Discussion

The reanalysis of Malejka et al. (2022) produced two main insights. First, the interpretation of their results as evidence pointing toward a violation of selective influence presupposes that the core assumptions of the 2HT model are simply taken as given. However, when adopting a broader perspective by entertaining the idea that the 2HT model could potentially be invalid, the reported results turn out to be nondiagnostic. That is, they are to be expected when the data come from the rival SDT model. What this means is that none of Malejka et al. (2022) results makes a compelling case against Kellen and Klauer's (2014) original selective-influence assumption per se, nor against the conclusions that they originally drew.

Second, the results turn out to manifest a signature prediction of SDT models at large that cannot be accounted for by the 2HT model without post hoc modifications in need of follow-up experimental

[12] Another—perhaps more direct—remedy is to abandon the notion that thresholds are "high," which leads to the class of *low-threshold models* (Luce, 1963; see also Kellen et al., 2016; McAdoo & Gronlund, 2020; Starns, 2021; Starns & Ma, 2018).

[13] The data fitted by Malejka et al. (2022) and the one simulated here differ in one important aspect. The former included two distinct categories for *logically incongruent* responses in the forced-choice and ranking trials (e.g., response category "LL": Choose the new item in the forced-choice trial and assign it Rank 1 immediately afterward). These categories were handled by three nuisance parameters (a_1, a_2, and a_3). Because the SDT model implemented here presumes a perfect carryover of familiarity values between judgments, it assigns probability zero to these incongruent response categories. This difference effectively reduces the number of degrees of freedom per individual data set by four, providing a total of 2 × 3 = 6 degrees of freedom. This reduction is compensated by the fact that the extended model no longer requires the three aforementioned nuisance parameters, leaving it with five parameters: D_o^w, D_n^w, D_o^s, D_n^s, and c. One attractive feature of the model fits obtained in this simulation is that they are not distorted by the ability of the nuisance parameters to account for the incongruent responses that neither model has a satisfactory explanation for at the moment. One possible explanation for these responses is that, contrary to what has been assumed so far, familiarity values or detection states are *not* carried over across trials, perhaps due to a second retrieval attempt in the ranking trials (the fact that attempts to relax this assumption do not explain the observed $P'^{\ominus B}_{\text{n, on}}$, see Footnote 11, does not imply that it could not play a role here). Another possibility is that participants are sometimes careless or sloppy. However, evaluating the merits of these and other possibilities is beyond the scope of the present work.

verification. Notably, the modifications introduced to accommodate the results from this single study affect a 2HT model feature that has been highlighted by previous critical reports. Namely, the presumption that new items cannot be misremembered (Meyer-Grant & Klauer, 2021; Starns, 2021; Starns et al., 2018; Starns & Ma, 2018; Voormann et al., in press). That said, further research is needed to investigate the manifestation of "response consistency," which can be incorporated into both 2HT and SDT models.

At this point, the viability of the 2HT model depends on two novel assumptions introduced by Malejka et al. (2022). One of them is the "response consistency" assumption just discussed, which enables it to account for the results from Experiment 3. The other one is the "contrast-processing" assumption that invalidates Kellen and Klauer's (2014) conclusions but also boils down to an abandonment of the selective-influence assumption on account of an unqualified dismissal of single-item generalization. We turn our attention to this second proposal and resultant issues in the section below.

Evaluating the Familiarity-Contrast Account

Malejka et al. (2022) showed how embedding a "contrast process" in the 2HT model yields the prediction that $D_n^w \leq D_n^s$. They also appeal to this account when describing some of their results (e.g., D_n estimates from ranking judgments and their respective predictions for forced-choice judgments; see their p. 10).

The purpose of this section is to provide a thorough reevaluation of this contrast account, which takes into consideration its broader implications for recognition-memory judgments at large. First, we will discuss how numerous studies have shown that multiple-item judgments can be successfully predicted from single-item judgments. These results corroborate the general assumption of *single-item generalization* that can be traced back to D. M. Green's (1960) area theorem. We will then turn our attention to the model proposed by Malejka et al. (2022). Beyond its impact on selective influence, we will show that this model produces gross violations of single-item generalization and other implausible results. Last, we will discuss a critical test that directly targets the contrast process and show how the latter entails a prediction that runs counter to previously published data.

Single-Item Generalization

In the context of recognition memory, single-item generalization is the assumption that the mnemonic status of items in single-item judgments also apply to other testing contexts, such as multiple-item tests. To give a simple example couched in SDT, consider the respective latent memory strengths of an old and a new item, denoted here by X_o and X_n. For a given response criterion κ, the probabilities of a "yes" judgment to these items are assumed to correspond to the probabilities that their respective latent strengths are greater than the response criterion (i.e., $X_o > \kappa$ or $X_n > \kappa$). As illustrated in Figure 1, the relationship between these two probabilities as a function of response criterion is known as the single-item ROC function. In turn, in the context of a 2AFC trial, in which one of the two items is to be chosen as the old one, the probability of a correct response is assumed to correspond to the probability that X_o is greater than X_n.

D. M. Green (1960) showed that, under these assumptions, the area under the single-item ROC function corresponds to the probability of a correct response in a forced-choice trial. This result, known as the area theorem, does not depend on the parametric distributions illustrated in Figure 1. According to D. M. Green (2020), the area theorem "is *the* critical contribution of the theory" (p. 222). A lesser-known result by Iverson and Bamber (1997) generalizes the area theorem to forced-choice judgments involving more than two options (for a discussion, see Kellen et al., 2021).

Putting aside its formal elegance or historical relevance, in the context of recognition memory we find ample empirical evidence supporting the validity of the area theorem and its generalization. Consider the following examples illustrated in Figure 7:

- Jang et al. (2009) evaluated participants' confidence judgments in single-item and 2AFC trials. Fitting a single Gaussian SDT model to both types of trials (thereby enforcing the Area Theorem, albeit with auxiliary parametric assumptions) results in acceptable fits for 90% of the 63 participants (only six individual data sets with $p <$ .05). One way to gauge this success is to consider how the same Gaussian SDT model fares when *predicting* accuracy in 2AFC judgments solely on the basis of single-item yes/no judgments: As shown in the top panel of Figure 7, there is considerable agreement between the predicted and observed accuracy, with a rank-correlation of 0.90, with no indication of systematic under- or over-prediction (median difference = 0.006; Wilcoxon W = 1,235, p = .121).

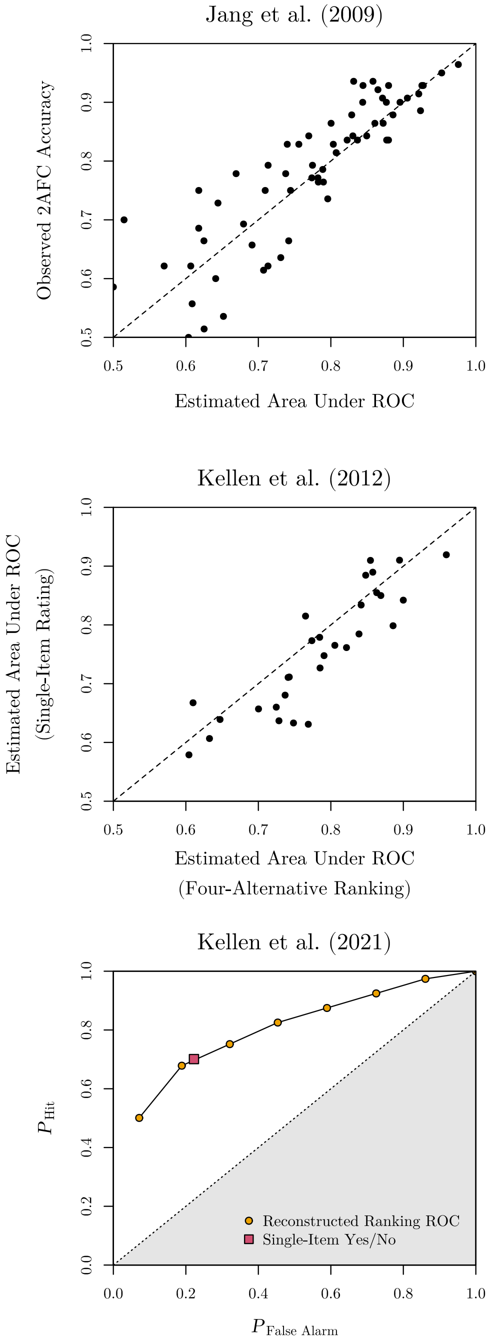
- Kellen et al. (2012) reported a study in which thirty participants engaged in a recognition test that intermixed single-item and four-alternative ranking trials. Again, fitting a single Gaussian SDT model to both types of trials resulted in acceptable fits for 87% of the participants (only four individual data sets with $p < .05$). These successful fits can be qualified in terms of the convergence between the predicted areas under the yes/no ROC (or 2AFC accuracy) obtained when fitting single-item ratings and four-alternative rankings separately. As shown in the middle panel of Figure 7, the rank correlation between these predictions is 0.89, with a small deviation in favor of ratings (median difference = 0.03; Wilcoxon W = 374, p = .003).

- Kellen et al. (2021, Experiment 2) intermixed single-item yes/no recognition trials with multiple-alternative forced-choice trials, with the number of alternatives ranging from two to eight. The judgments obtained from these forced-choice trials were used in a nonparametric reconstruction of the yes/no ROC (for details, see Kellen et al., 2021). As shown in the bottom panel of Figure 7, the data point for the single-item hit and false rates falls right on top of the reconstructed yes/no ROC function.

Ideally, any attempt to dispense with single-item generalization needs to come to terms with these empirical successes, not recast them as puzzles.

Figure 7

Illustration of Different Results Consistent With Single-Item Generalization



Note. Top panel: Estimates of the area under the yes/no ROC obtained with confidence ratings and the observed 2AFC accuracy. Data from Jang et al. (2009). Middle panel: Estimates of the area under the yes/no ROC obtained with four-alternative ranking judgments and single-item ratings. Data from Kellen et al. (2012). Bottom panel: Reconstructed yes/no ROC obtained from ranking judgments, which were in turn derived from forced/choice judgments (for details, see Kellen et al., 2021). Single yes/no ROC point collected in the same study. Data from Kellen et al. (2021, Experiment 2). ROC = receiver operating characteristic; 2AFC = two-alternative forced choice. See the online article for the color version of this figure.

Two-High Threshold Contrast Model

The two-high threshold contrast (2HTC) model sketched out by Malejka et al. (2022) postulates that the detection probabilities emerge from a contrast between the different alternatives presented at test (for technical details, see their Appendix A). More specifically, it assumes that each test item has a latent-strength value ψ'. These latent-strength values are independent random samples from different latent distributions associated with old and new items. When K test items are encountered simultaneously, as in the case of the $K = 4$ alternative ranking task illustrated in Figure 2, the latent *contrast-strength* value of each item k, denoted by ψ_k, is computed by subtracting the weighted average of all the other $K - 1$ items,

$$\psi_k = \psi'_k - \sum_{j \neq k} w_j \psi'_j, \tag{15}$$

with $w_j \geq 0$ and $\sum w_j = 1$. For simplicity, it can be assumed that $w_j = 1/K - 1$. It is worth noting that this type of contrast finds precedent in the "eyewitness-identification" literature, where an identical proposal is presumed to operate in lineup procedures (see Wixted et al., 2018). There is also precedent in the "multi-alternative decision making" literature, where identical or very similar contrast processes have been proposed (e.g., see Palminteri et al., 2015, p. 11; Roe et al., 2001, p. 374; Trueblood et al., 2014, p. 186).

According to the 2HTC model, the probability of an *old* item being correctly detected as *old* corresponds to the probability that its contrast-strength value ψ_o surpasses an *upper* threshold h_u,

$$D_o = P(\psi_o \geq h_u). \tag{16}$$

The probability of a *new* item being correctly detected as *new* corresponds to the probability of ψ_n falling below a *lower* threshold h_l,

$$D_n = P(\psi_n \leq h_l). \tag{17}$$

Test items for which $h_l \leq \psi \leq h_u$ are said to be in an uncertainty state. Notably, it is assumed that both thresholds are "high," in the sense that only old items can surpass h_u, and only new items can fall below h_l. In other words, $P(\psi_n \geq h_u) = P(\psi_o \leq h_l) = 0$.

Since a study-strength manipulation is assumed to affect the latent strength values of old items, their expected value should be larger for strong items than for weak items. This effect is expected to *affect the detection of both old and new items*, such that $D_o^w \leq D_o^s$ and $D_n^w \leq D_n^s$, which in turn implies the empirically corroborated prediction that $c_2^w \leq c_2^s$ (see Figure 3).

For a proper assessment of the relationship between the familiarity-contrast process and single-item generalization, it is crucial to understand how the model applies to single-item judgments. This is easily achieved by appealing to existing precedent in casting the 2HT model in terms of latent strengths (for a recent example, see Malejka & Bröder, 2019). If only one item is encountered per test trial, it follows that

$$\begin{aligned} D'_o &= P(\psi'_o \geq h_u), \\ D'_n &= P(\psi'_n \leq h_l). \end{aligned} \tag{18}$$

The similarities between this latent-variable and the continuous SDT models found in the literature at large are obvious (see Kellen et al., 2021; Malejka & Bröder, 2019; Meyer-Grant & Klauer, 2021;

Rouder & Morey, 2009). An important distinction, however, is that the SDT models provide researchers with the means to identify a *common latent scale* to the point that one can pin down the relative distance between latent distributions, as in the case of the SDT index *d′* (see, e.g., Macmillan & Creelman, 2005). But this scaling is only possible when allowing the latent distributions to cross the imposed thresholds *with nonzero probability*.[14] But by definition, these crossovers are *not* allowed in the case of the 2HTC model, as both of its h_{l} and h_{u} thresholds are stipulated to be "high."

The ambiguity surrounding the scaling of the latent variables under high-threshold assumptions introduces all kinds of strange predictions when moving from a single item to two items, as in the classic case of a 2AFC task considered by the area theorem. To see this, consider the case where $\psi'_{\text{o}}\tilde{U}(0,1)$ and $\psi'_{\text{n}}\tilde{U}(-4,-1)$, with $h_{\text{l}} = -3.5$ and $h_{\text{u}} = 0.75$. For reference, "~" stands for "distributed as" and $U(\min, \max)$ denotes a uniform distribution with a given lower (min) and upper bound (max). Based on these distributions and high thresholds, it follows that

- Single item: $D'_{\text{o}} = .25$ and $D'_{\text{n}} = .17$,
- 2AFC: $D_{\text{o}} = 1$ and $D_{\text{n}} = .33$.

In this specific example, we see a modest performance in single-item judgments, with a predicted area under the ROC of .69. For reference, the area entailed by chance performance is .50. But when presented with two items at a time (one old and one new), performance skyrockets to perfect accuracy by virtue of $D_{\text{o}} = 1$. This discrepancy is a gross violation of the area theorem, according to which both values, the area under the single-item ROC and the 2AFC accuracy, should coincide.[15]

Alternatively, let $\psi'_{\text{o}}\tilde{U}(6,10)$ and $\psi'_{\text{n}}\tilde{U}(0,2)$, with $h_{\text{l}} = 0.5$ and $h_{\text{u}} = 8$,

- Single item: $D'_{\text{o}} = .50$ and $D'_{\text{n}} = .25$,
- 2AFC: $D_{\text{o}} = .25$ and $D_{\text{n}} = 1$.

Here, we see that presenting a new item alongside an old one halves the amount of times that the latter are expected to be detected, while also rendering the correct detection of new items a virtual certainty. Again, these predictions violate the area theorem (predicted area is .81, predicted accuracy is 1).[16]

In reaction, one could object to these examples by arguing that the latent variables (i.e., ψ'_{o} and ψ'_{n}) that govern recognition in these situations are assumed to be different across tasks. Taking this one step further, one might also be inclined to argue that by stipulating task-invariant thresholds, our approach constitutes an illegitimate extension of the 2HTC to the single-item case. But for all its ostensible plausibility, this line of reasoning ultimately proves specious. Not only would it mean that the 2HTC model has an undisclosed (and unexplained) limit condition, but it does also not preclude potential violations of the area theorem. In fact, violations are expected under most circumstances, as the area under the single-item ROC function according to the 2HTC model can be shown to equal $(1 + D'_{\text{o}} + D'_{\text{n}} - D'_{\text{o}}D'_{\text{n}})/2$ and the probability for a correct response in the 2AFC task is given by $(1 + D_{\text{o}} + D_{\text{n}} - D_{\text{o}}D_{\text{n}})/2$. The theorem is thus only satisfied when $D'_{\text{n}} = D_{\text{n}}$ and $D'_{\text{o}} = D_{\text{o}}$, as well as in the rather implausible cases where the thresholds and/or distributions in both tasks are somehow precisely coordinated such that changes between D'_{o} and D_{o} are exactly offset by changes between D'_{n} and D_{n}.

An objection against the notion that the distributions of ψ'_{o} and ψ'_{n} are invariant between single-item and multiple-item tests would also fail to address the issues that arise when *simply varying the number of alternatives*. Let $\psi'_{\text{o}}\tilde{U}(1.5, 2.5)$ and $\psi'_{\text{n}}\tilde{U}(0,1)$, with $h_{\text{l}} = -1.5$ and $h_{\text{u}} = 1.5$. Under these parameter values, the expected detection probabilities for 2AFC trials take on reasonable values but the detection of new items drops to near zero *as soon as a single additional alternative is introduced*,

- 2AFC: $D_{\text{o}} = .50$ and $D_{\text{n}} = .50$,
- 3AFC: $D_{\text{o}} = .50$ and $D_{\text{n}} = .01$.

Aside from the drastic drop in D_{n}, these expected detection probabilities imply accuracy rates of $P_{2\text{AFC}} = .875$ and $P_{3\text{AFC}} = .668$ in 2AFC and 3AFC trials, respectively. These rates violate a *multiplicative inequality* that is implied by the assumption that the latent variables underlying decisions are independent (Sattath & Tversky, 1976; see also Kellen et al., 2021),

$$P_{3\text{AFC}} \geq (P_{2\text{AFC}})^2. \tag{19}$$

However, $.875^2 \approx .766$, which is noticeably greater than .668. This violation is particularly interesting for two reasons: First, ROC reconstructions, such as the one illustrated in the bottom panel of Figure 7, are predicated on this and similar multiplicative inequalities holding true (Kellen et al., 2021; Sattath & Tversky, 1976). Second, to the best of our knowledge, there is no empirical evidence even suggesting that these inequalities are violated. The only studies directly testing them, namely the first two experiments reported by Kellen et al. (2021), show them passing with flying colors.

As before, one could attempt to brush aside this issue by dismissing the assumption that the same "high" thresholds apply across different numbers of alternatives. But such a maneuver would only further undermine the prospects of the 2HTC model ever linking different types of judgments—links that are established under the rival SDT model. One would be essentially committing high-threshold modeling to a *strong operationist* stance that has long been dismissed as unworkable for any substantive research program (see C. D. Green, 1992; Koch, 1992; Leahey, 1980).[17]

[14] The inability to compute indices such as SDT's *d′* with hit or false-alarm rates of 0 or 1 is a reflection of this requirement.

[15] As pointed out by a reviewer, the 2HTC model was developed with ranking judgments in mind, not forced-choice judgments. That much granted, we still believe that it is acceptable to assume that the processes involved in assigning an item Rank 1 are the same as those involved in selecting that item as old.

[16] These examples may appear to present fairly extreme cases with nonoverlapping signal and noise distributions. However, when distributions are overlapping, such as when we let $\psi'_{\text{n}}\tilde{U}(0,1)$ and $\psi'_{\text{o}}\tilde{U}(l,u)$, where $l \in (0,1)$ and $u > 1$, it follows that the smallest possible value that the contrast $\psi_{\text{o}} = \psi'_{\text{o}} - \psi'_{\text{n}}$ can take is $l - 1 < 0$. But this implies that the high-threshold assumption postulated for the detection probabilities in Equations 16 and 17 cannot be upheld under stable thresholds for the two tasks since positive values of D_{n} in the single-item task necessitate that $h_{\text{l}} > 0$ and therefore $P(\psi_{\text{o}} < h_{\text{l}}) > 0$.

[17] Strong operationism is not necessarily problematic in the context of a *cognitive-psychometric research program*, where the focus is on a characterization that is tailored to a specific experimental paradigm (e.g., Batchelder, 2010; Riefer et al., 2002). But one should not confuse the application of threshold models in this context with their analysis as *substantive hypotheses*, as in the case of Kellen and Klauer (2014) or Province and Rouder (2012).

All things considered, the cost of taking the 2HTC model seriously appears to be too high. The model does succeed in providing a high-threshold answer to Kellen and Klauer's (2014) and Kellen et al.'s (2021) critical-test results. But this is achieved at the expense of more basal relationships between single-item and multiple-item judgments.

Perhaps as a measure of last resort, one could argue that the present analysis is predicated on an overly strong reading of the 2HTC model, which was only meant as an illustration of how violations of selective influence *could* occur. We do not see how this defensive position could be tenable. As discussed earlier, Malejka et al. (2022) questioned Kellen and Klauer's (2014) assumption of selective influence in part based on theoretical considerations materialized in the 2HTC model. But if not this theoretical account, which one then? Without proposing one, the renunciation of selective influence on account of some undisclosed process could be easily mistaken for a whim. And Malejka et al. themselves appeal to the 2HTC's predictions for different numbers of alternatives (e.g., see their p. 10); so it is unclear to us on what grounds one could reasonably accept these appeals while finding the present analysis to be beyond the pale. Finally, we note that the cases provided here are not unflattering exceptions but representative examples of the predictions obtained when exploring the 2HTC model (e.g., see Footnote 16).

A New Critical Test

Fortunately, it is possible to evaluate a contrast account while leaving the technical issues found in the 2HTC model aside. At the heart of this account is a *mutual dependency* among the multiple test items. That is, experimental manipulations affecting the recognition of old items are bound to affect the recognition of new items and vice versa.

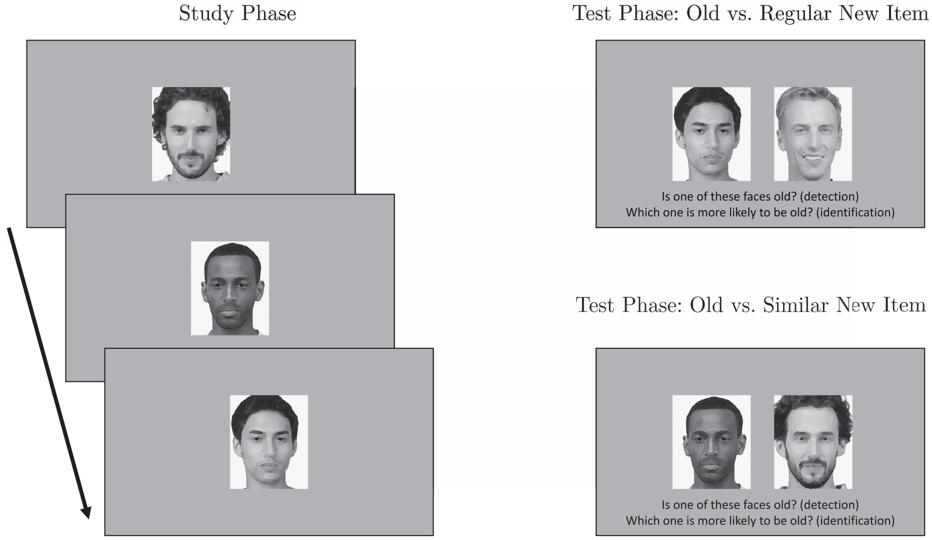
The predicted increase in D_n as a result of the strengthening of old items is therefore only one side of the coin. On the other side is the predicted influence of new-item manipulations on D_o. Equation 15 can be used to illustrate this: Manipulating the ψ' values associated to new items will affect the ψ values for old items and therefore D_o. For instance, increasing ψ'_n values will result in lower D_o values.

Meyer-Grant and Klauer (2021) reported a study that speaks directly to this mutual dependency and the effect of manipulating new items on old-item detection D_o. Their study involved a *simultaneous detection and identification* paradigm (see, e.g., Macmillan & Creelman, 2005) in which participants, after studying a list of faces, encountered test trials comprised of pairs of faces. In one half of the trials, both faces were new. In the other half, one of the faces was old. Upon encountering each face pair at test, participants were first requested to make a *detection judgment*, indicating whether they believed that one of the faces was old. A "detection hit" was said to occur whenever a participant correctly indicated that an old–new pair included an old face. This detection judgment was followed by an *identification judgment* that pinpointed the face found to be old. Figure 8 sketches out this experimental design.

A crucial element of Meyer-Grant and Klauer's (2021) experiment was the manipulation of *new-item similarity*: In some of the test trials, the new faces encountered were similar to studied but untested faces (similar new items), whereas in the rest of the trials the new faces were not similar to any of the studied ones in any

Figure 8

Illustration of the Experimental Design Used by Meyer-Grant and Klauer (2021)


Study Phase
Test Phase: Old vs. Regular New Item
Is one of these faces old? (detection)
Which one is more likely to be old? (identification)
Test Phase: Old vs. Similar New Item
Is one of these faces old? (detection)
Which one is more likely to be old? (identification)

Note. This illustration does not depict all possible test trials (i.e., "Regular New vs. Similar New Item" and "Regular New vs. Regular New Item" were omitted). Furthermore, it is important to point out that participants first had to complete the detection task (in form of a four-level confidence rating ranging from "sure that no old face is present" to "sure that an old face is present") and only afterward had to complete the identification task (i.e., they were asked to select one of the two faces as being more likely old).

systematic way (regular new items). The goal of this manipulation was to implement a critical test for the 2HT model. Meyer-Grant and Klauer showed that the 2HT model must predict that detection hit-rates (P_{hit}) *are smaller* for pairs including a similar new face compared to pairs including a new face that bore no such similarity ($P_{\text{hit, similar new}} < P_{\text{hit, regular new}}$; for a detailed proof, see their Proposition 10). This prediction turns out to be grossly incorrect—detection hit-rates turn out to be *greater* when the pair includes a new face that is similar to a studied one (see Figure 9, bottom panel). In contrast, this result is easily accommodated by a Gaussian SDT model presuming single-item generalization. All that it takes is assuming that the latent-strength distribution associated with similar new items falls in between the distributions for regular new items and old items (see both panels of Figure 9; for a similar characterization, see Starns et al., 2007).

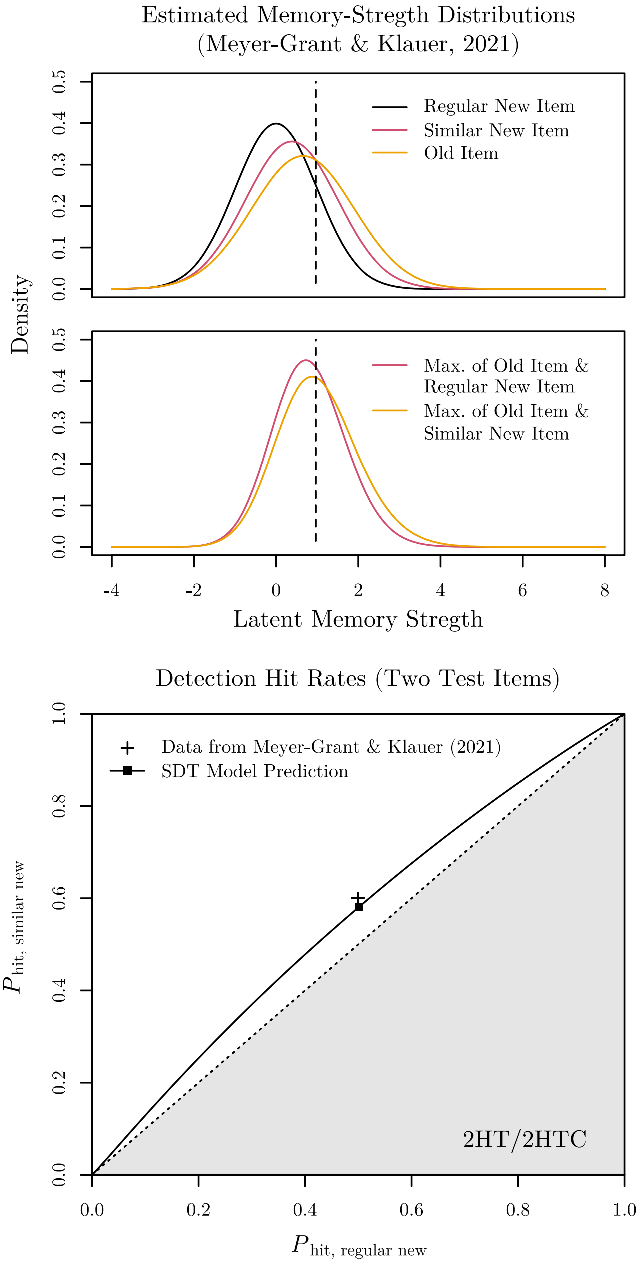
Introducing a contrast process to either the 2HT or SDT model brings no benefit whatsoever. In fact, it undermines both models. For the 2HT model, it exacerbates its incorrect prediction. For the SDT model, it forces the model to shift from a correct prediction to an incorrect one (see Figure 9, bottom panel). Formal proofs of the 2HT and SDT predictions are provided in Appendix B and in Meyer-Grant and Klauer (2022), respectively.

In response, one could argue that the simultaneous detection and identification paradigm utilized by Meyer-Grant and Klauer (2021) may elicit cognitive processes that differ from those involved in the ranking tasks specifically targeted by Malejka et al. (2022). But given that in Meyer-Grant and Klauer's experiment, two test items were always presented simultaneously, this would stand in direct contradiction to Malejka et al.'s own reasoning (see the earlier quote) as to why ranking judgments supposedly involve a contrast process; namely because a ranking task requires the assessment of multiple items per trial, as opposed to tasks requiring the assessment of only a single item. Of course, one could try to dismiss the contradiction by simply declaring that the contrast process exclusively operates on ranking judgments, and that it does not extend to simultaneous detection and identification. The reason for this exception would be a theoretical puzzle for someone to figure out in the future. SDT modelers presuming single-item generalization, on the other hand, will find themselves in very different circumstances, as they are able to apply a single, consistent account, without having to quarantine every possible experimental-task variant away from each other, nor having to issue any kind of promissory notes.

Furthermore, this caveat does also not apply to similar results of a recently conducted experiment implementing a ranking task that varied the number of alternatives (i.e., $K \in \{3, 4, 5\}$; Meyer-Grant & Jakob, in press). It can be shown that the contrast process outlined by Malejka et al. (2022) should lead to a decrease in D_{n} when K increases under fairly general conditions. According to this process, D_{n} is determined by comparing a new item's latent-strength value with the weighted average of the remaining $K - 1$ values (see Equation 15). Since it only includes a single value coming from the old-item distribution, this average will tend to decrease with an increase in K. This results in a less pronounced difference compared to the single new-item value and, consequently, a decrease in D_{n}. Contrary to this prediction, however, Meyer-Grant and Jakob (in press) observed data that call for an *increase* of D_{n} with K. What is more, these results are once again to be expected if the data had been generated by an SDT model.

Figure 9

Illustration of the SDT Model and Its Predictions for the Data From Meyer-Grant and Klauer (2021)



Note. Top panel: Latent-strength distributions estimated by Meyer-Grant and Klauer (2021). The top density plot illustrates the latent-strength distributions for old, regular new, and similar new items. The bottom density plot illustrates the distributions of the maximum latent strength when pairing an old item with a regular new item or a similar new item. Bottom panel: Observed detection hit rates for pairing of an old item with a regular new item or a similar new item. These rates are accompanied by their respective 95% bootstrap confidence intervals. The SDT model predictions obtained from the latent-strength distributions illustrated above are given by the solid black line. The range of predictions allowed by the 2HT and 2HTC models is given by the shaded area. SDT = signal detection theory; 2HT = two high-threshold; 2HTC = two high-threshold contrast. See the online article for the color version of this figure.

General Discussion

As the saying goes, there is no free lunch. The same can be said of models and theories; their predictions must be paid upfront in the form of auxiliary assumptions. But whenever possible, researchers should attempt to scrutinize the latter. The work of Malejka et al. (2022) is an attempt to do so. In it, they correctly identified that the dismissal of the 2HT model reported by Kellen and Klauer (2014) depended on selective influence being satisfied. In hindsight, Kellen and Klauer should have done a better job laying out the rationale for this assumption. But as pointed out earlier, Kellen and Klauer's experiments were designed with the plausibility of this assumption in mind. Regardless, selective influence could be false for theoretical reasons, such as the operation of a contrast process. The problem with Malejka et al.'s conclusions is that what they took to be evidence against selective influence turns out to be expected in data generated from SDT. Rather than reclaiming the viability of the 2HT model, their results could very well be reiterating the alignment of people's judgments with SDT. Their observation of negative D_{n} estimates in one of their studies is consistent with this possibility. To make matters worse, the theoretical motivation offered by their contrast account yields implausible predictions, including a critical-test prediction that is empirically rejected (for similar results, see Meyer-Grant & Jakob, in press). As it stands, we see no empirical or theoretical grounds to renounce Kellen and Klauer's (2014) dismissal of the 2HT model.

However, the present lack of evidence against selective influence and single-item generalization does not mean that there are no good reasons to doubt them *elsewhere*. For instance, it seems plausible that single-item judgments do not generalize well to circumstances where multiple *similar* items (i.e., items that are similar to each other) are presented simultaneously, as in the case of eyewitness-identification lineups (see Wixted et al., 2018).[18] Indeed, the finding that performance in different kinds of lineups is superior to performance in single-item "showups" speaks directly against the assumption (e.g., Kellen & McAdoo, 2022; Wixted et al., 2018). That being said, these circumstances are markedly different from the ones considered in the present work. Therefore, such findings do not necessarily speak against single-item generalization at large. Rather, they contribute toward a better delineation of its scope.

Preoccupations surrounding scope bring us to another important issue, namely the importance of distinguishing between cases in which *a prediction fails* from those in which *no prediction can be made* (for a relevant discussion, see Kellen et al., 2023). Again, consider the case of D. M. Green's (1960) area theorem. The theorem rests on the assumption that, when shown two items in a forced-choice trial, participants will select the one that maximizes a given latent-strength variable (i.e., select the most familiar item). In other words, it is presumed that both items are in fact evaluated and compared. However, it is easy to think of circumstances where this is *not* the case; for example, when making the inspection of individual items costly. Any claim that the area theorem is violated under such circumstances is *logically invalid* given that one of its key premises fails to hold. For example, Starns et al. (2017) reported a study showing that participants often failed to make a comparison in 2AFC trials (ca. 20%–33%), basing their choices on the evaluation of a single item. However, it is worth noting that items in these 2AFC trials were kept wide apart on the screen, which could have encouraged participants to adopt such a strategy. Researchers should be aware of this risk when developing and implementing new tests (for relevant discussions, see Chechile & Dunn, 2021; Szollosi et al., 2023; Szollosi & Newell, 2020).

An interesting development coming out of the present work is the discovery of signature predictions on $P'^{\ominus B}_{\text{n, on}}$ that do not require the deployment of study-strength or response-bias manipulations and their respective selective-influence assumptions. Coincidentally, another signature prediction operating under similar circumstances was recently discovered by Chechile and Dunn (2021). They showed that the distributions of *conditional* old-item ranking probabilities ($c_{2,K}$, $c_{3,K}$, $c_{4,K}$, …) predicted by the 2HT model are *monotonically increasing*, whereas a large family of SDT models predicts *nonmonotonic* conditional ranking probabilities. McCormick and Semmler (2023) tested these diverging predictions and reported results in favor of nonmonotonic conditional ranking probabilities, as predicted by SDT.[19] As alternative testing methods in researchers' toolboxes, the application of Malejka et al.'s Versions *A* and *B* as methods for evaluating high-threshold accounts strike us as quite promising.

It should also be noted that not all critical-test results reported so far speak in favor of the SDT account and against the rival 2HT model. For instance, Kellen and Klauer (2015) reported a critical test of *confidence-rating judgments for unrecognized items* that turned out to be in line with the 2HT model. In short, they found that the conditional distribution of confidence-ratings for unrecognized items was unaffected by study-repetition manipulations. This result runs counter to the SDT prediction that for strong old items, errors should be rarer but also less extreme (less confident). However, we do not think that this discrepant result undermines the overall support reviewed so far. Looking at the SDT model, the culprit is its assumption that confidence ratings result from direct overlaying of confidence criteria on the latent-strength scale. Specifically, this assumption appears to be at fault whenever participants judge studied items as "new." It is possible that the problem lies in the way participants engage with the recognition task, especially when there are no ostensive task demands (e.g., Delay & Wixted, 2021; Klauer & Kellen, 2010). If this turns out to be the case, then researchers should be able to mitigate it by tinkering with their experimental designs.[20]

Given its track record and outlook, what is to be made of the 2HT model? Well, that depends on what you are trying to do with it. A common issue in debates between proponents of continuous and

[18] In this particular case, the new items are similar to the old item *included in that trial* (in the context of a lineup, a guilty suspect paired with filler faces matching the suspect description). Speaking of similarity in this sense should be distinguished from a new item's similarity with another old item not presented at test (e.g., Meyer-Grant & Klauer, 2021). For a discussion on these different senses, see Meyer-Grant and Klauer (2022; see also Heathcote et al., 2009; Tulving, 1981).

[19] These results are consistent with our own indirect assessments. Using formal results described by Kellen et al. (2021), it is possible to derive ranking probabilities from forced-choice judgments. These derivations are part of the process through which the ROC found in the bottom panel of Figure 7 is obtained. The conditional ranking probabilities obtained through this process, using the data from Kellen et al., turn out to be nonmonotonic.

[20] In our view, this revision of the SDT model is less demanding than Malejka et al.'s (2022) introduction of "response consistency." The reason is simply that the explanation given for the recalcitrant confidence judgments does not imply that participants somehow acted against the instructions given by the experimenter. Whereas in the case of "response consistency," one must buy into the idea that participants often refused to engage with the novel test items presented to them in the ranking trials.

discrete-state accounts is a confusion regarding the role played by the latter. One thing is for the 2HT model to stand as a proxy for a *substantive hypothesis*—in that case, we believe that there are clear grounds for its dismissal. Another is for 2HT model to serve as a *pragmatic option in a cognitive-psychometric enterprise* (cf. Batchelder, 2010; Chechile, 2018; Riefer et al., 2002). Depending on the application, a case could be made that specific failures are negligible or perhaps even irrelevant (see Batchelder & Alexander, 2013; but see also Brady et al., 2023; Dubé et al., 2013; Pazzaglia et al., 2013; Williams et al., 2023). Please note that SDT can also be deployed as a proxy for a substantive hypothesis or as a cognitive-psychometric tool.

In closing, it is important to dispel a negative reading of the present 2HT/SDT debate. Given that the 2HT model has long been seen by many as empirically inadequate (e.g., Egan, 1958), one could perceive the present discussion as an example of how research in psychology is intellectually bankrupt, with failed theories not being given a proper funeral as long as enough people are interested in them (see Meehl, 1978). The problem with such a reading is that it ignores the fact that the core principles of the 2HT model have for the longest time been dismissed on *invalid* grounds (see Bröder & Schütz, 2009; Klauer & Kellen, 2010; Malmberg, 2002). Recent efforts to address this problem have led to many new tests, such as the ones discussed here, ultimately producing a stronger case that rests on a much deeper understanding of how different memory judgments relate to each other. From where we stand, it looks like progress.

References

Aytaç, S., Kılıç, A., Criss, A. H., & Kellen, D. (2024). Retrieving effectively from source memory: Evidence for differentiation and local matching processes. *Cognitive Psychology*, *149*, 1–24. https://doi.org/10.1016/j.cogpsych.2023.101617

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71–93). American Psychological Association.

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dubé, and Rotello (2013). *Psychological Bulletin*, *139*(6), 1204–1212. https://doi.org/10.1037/a0033894

Birnbaum, M. H. (2011). Testing mixture models of transitive preference: Comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*(4), 675–683. https://doi.org/10.1037/a0023852

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2023). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review*, *30*(2), 421–449. https://doi.org/10.3758/s13423-022-02179-w

Bröder, A., & Malejka, S. (2017). On a problematic procedure to manipulate response biases in recognition experiments: The case of "implied" base rates. *Memory*, *25*(6), 736–743. https://doi.org/10.1080/09658211.2016.1214735

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*(3), 587–606. https://doi.org/10.1037/a0015279

Chechile, R. A. (2018). *Analyzing memory: The formation retention and measurement of memory*. MIT Press.

Chechile, R. A., & Dunn, J. C. (2021). Critical tests of the two high-threshold model of recognition via analyses of hazard functions. *Journal of Mathematical Psychology*, *105*, Article 102600. https://doi.org/10.1016/j.jmp.2021.102600

Chechile, R. A., Sloboda, L. N., & Chamberland, J. R. (2012). Obtaining separate measures for implicit and explicit memory. *Journal of Mathematical Psychology*, *56*(1), 35–53. https://doi.org/10.1016/j.jmp.2012.01.002

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, *147*(4), 545–590. https://doi.org/10.1037/xge0000407

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language*, *55*(4), 461–478. https://doi.org/10.1016/j.jml.2006.08.003

Delay, C. G., & Wixted, J. T. (2021). Discrete-state versus continuous models of the confidence–accuracy relationship in recognition memory. *Psychonomic Bulletin & Review*, *28*(3), 556–564. https://doi.org/10.3758/s13423-020-01831-7

Dubé, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 130–151. https://doi.org/10.1037/a0024957

Dubé, C., Rotello, C. M., & Pazzaglia, A. (2013). The statistical accuracy and theoretical status of discrete-state MPT models: Reply to Batchelder and Alexander (2013). *Psychological Bulletin*, *139*(6), 1213–1220. https://doi.org/10.1037/a0034453

Dubé, 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*(3), 389–406. https://doi.org/10.1016/j.jml.2012.06.002

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

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Rep.). Indiana University Hearing and Communication Laboratory.

Green, C. D. (1992). Of immortal mythological beasts: Operationism in psychology. *Theory & Psychology*, *2*(3), 291–320. https://doi.org/10.1177/0959354392023003

Green, D. M. (1960). Psychoacoustics and detection theory. *The Journal of the Acoustical Society of America*, *32*, 1189–1203. https://doi.org/10.1121/1.1907882

Green, D. M. (2020). A homily on signal detection theory. *The Journal of the Acoustical Society of America*, *148*, 222–225. https://doi.org/10.1121/10.0001525

Harding, S. G. (1976). *Can theories be refuted? Essays on the Duhem–Quine thesis*. D. Reidel Publishing.

Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review*, *16*(5), 824–831. https://doi.org/10.3758/PBR.16.5.824

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87. https://doi.org/10.1016/j.jmp.2019.03.004

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

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*(2), 291–306. https://doi.org/10.1037/a0015525

Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in two-alternative-forced-choice recognition: Is the test criterion-free? *Acta Psychologica*, *167*, 30–44. https://doi.org/10.1016/j.actpsy.2016.03.014

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, *2*, 160–165. https://doi.org/10.1007/s42113-019-00037-y

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. https://doi.org/10.1016/j.jmp.2016.03.001

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology*, *55*(3), 251–266. https://doi.org/10.1016/j.jmp.2010.11.004

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*(6), 1795–1804. https://doi.org/10.1037/xlm0000016

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, *122*(3), 542–557. https://doi.org/10.1037/a0039251

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*(4), 693–719. https://doi.org/10.3758/s13423-013-0407-2

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review*, *119*(3), 457–479. https://doi.org/10.1037/a0027727

Kellen, D., & McAdoo, R. M. (2022). Toward a more comprehensive modeling of sequential lineups. *Cognitive Research: Principles and Implications*, *7*, Article 65. https://doi.org/10.1186/s41235-022-00397-3

Kellen, D., Meyer-Grant, C. G., & Klauer, K. C. (2024, July). *Critical testing in recognition memory: Selective influence, single-item generalization, and the high-threshold hypothesis*. Open Science Framework. https://osf.io/qz5re

Kellen, D., Pedersen, A. P., & Klauer, K. C. (2023). *On piecemeal testing and the possibility of psychological science*. PsyArXiv. https://osf.io/preprints/psyarxiv/89bkp

Kellen, D., Singmann, H., & Batchelder, W. H. (2018). Classic-probability accounts of mirrored (quantum-like) order effects in human judgments. *Decision*, *5*(4), 323–338. https://doi.org/10.1037/dec0000080

Kellen, D., Singmann, H., Vogt, J., & Klauer, K. C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology*, *62*(1), 40–53. https://doi.org/10.1027/1618-3169/a000272

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review*, *128*(6), 1022–1050. https://doi.org/10.1037/rev0000288

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*(4), 465–478. https://doi.org/10.3758/pbr.17.4.465

Klauer, K. C., & Kellen, D. (2015). The flexibility of models of recognition memory: The case of confidence ratings. *Journal of Mathematical Psychology*, *67*, 8–25. https://doi.org/10.1016/j.jmp.2015.05.002

Koch, S. (1992). Psychology's Bridgman vs Bridgman's Bridgman: An essay in reconstruction. *Theory & Psychology*, *2*(3), 261–290. https://doi.org/10.1177/0959354392023002

Leahey, T. H. (1980). The myth of operationism. *Journal of Mind and Behavior*, *1*(2), 127–143. https://www.jstor.org/stable/43852818

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61–79. https://doi.org/10.1037/h0039723

Ma, Q., Starns, J. J., & Kellen, D. (2022). Bias effects in a two-stage recognition paradigm: A challenge for "pure" threshold and signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(10), 1484–1506. https://doi.org/10.1037/xlm0001107

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

Malejka, S., & Bröder, A. (2019). Exploring the shape of signal-detection distributions in individual recognition ROC data. *Journal of Memory and Language*, *104*, 83–107. https://doi.org/10.1016/j.jml.2018.09.001

Malejka, S., Heck, D. W., & Erdfelder, E. (2022). Recognition-memory models and ranking tasks: The importance of auxiliary assumptions for tests of the two-high-threshold model. *Journal of Memory and Language*, *127*, Article 104356. https://doi.org/10.1016/j.jml.2022.104356

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 380–387. https://doi.org/10.1037/0278-7393.28.2.380

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*(1), Article 11. https://doi.org/10.1186/s41235-016-0014-7

McAdoo, R. M., & Gronlund, S. D. (2020). Theoretical note: Exploring Luce's (1963) low-threshold model applied to recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 247–256. https://doi.org/10.1037/xlm0000731

McAdoo, R. M., Key, K. N., & Gronlund, S. D. (2019). Task effects determine whether recognition memory is mediated discretely or continuously. *Memory & Cognition*, *47*(4), 683–695. https://doi.org/10.3758/s13421-019-00894-9

McCormick, K. M., & Semmler, C. (2023). *Letting go of the grail: Falsifying the theory of 'true' eyewitness identifications*. PsyArXiv. https://osf.io/preprints/psyarxiv/dszqw

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meyer-Grant, C. G., & Jakob, M. (in press). Ranking tasks in recognition memory: A direct test of the two-high-threshold contrast model. *Journal of Experimental Psychology: General*.

Meyer-Grant, C. G., & Klauer, K. C. (2021). Monotonicity of rank order probabilities in signal detection models of simultaneous detection and identification. *Journal of Mathematical Psychology*, *105*, Article 102615. https://doi.org/10.1016/j.jmp.2021.102615

Meyer-Grant, C. G., & Klauer, K. C. (2022). Disentangling different aspects of between-item similarity unveils evidence against the ensemble model of lineup memory. *Computational Brain & Behavior*, *5*(2), 160–174. https://doi.org/10.1007/s42113-022-00135-4

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*, Article 8096. https://doi.org/10.1038/ncomms9096

Pazzaglia, A. M., Dubé, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–1203. https://doi.org/10.1037/a0033044

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), 14357–14362. https://doi.org/10.1073/pnas.1103880109

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60*, 20–43. https://doi.org/10.2307/2181906

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163–178. https://doi.org/10.1037/0278-7393.16.2.163

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*(2), 184–201. https://doi.org/10.1037//1040-3590.14.2.184

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370–392. https://doi.org/10.1037/0033-295X.108.2.370

Rouder, J. N., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review*, *116*(3), 655–660. https://doi.org/10.1037/a0016413

Sattath, S., & Tversky, A. (1976). Unite and conquer: A multiplicative inequality for choice probabilities. *Econometrica*, *44*(1), 79–89. https://doi.org/10.2307/1911382

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Starns, J. J. (2021). High- and low-threshold models of the relationship between response time and confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(4), 671–684. https://doi.org/10.1037/xlm0000960

Starns, J. J., Chen, T., & Staub, A. (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language*, *93*, 55–66. https://doi.org/10.1016/j.jml.2016.09.001

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

Starns, J. J., Lane, S. M., Alonzo, J. D., & Roussel, C. C. (2007). Metamnemonic control over the discriminability of memory evidence: A signal detection analysis of warning effects in the associative list paradigm. *Journal of Memory and Language*, *56*(4), 592–607. https://doi.org/10.1016/j.jml.2006.08.013

Starns, J. J., & Ma, Q. (2018). Guessing versus misremembering in recognition: A comparison of continuous, two-high-threshold, and low-threshold models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 527–539. https://doi.org/10.1037/xlm0000461

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1379–1396. https://doi.org/10.1037/0278-7393.24.6.1379

Szollosi, A., Donkin, C., & Newell, B. R. (2023). Toward nonprobabilistic explanations of learning and decision-making. *Psychological Review*, *130*(2), 546–568. https://doi.org/10.1037/rev0000355

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*, *24*(12), 1008–1018. https://doi.org/10.1016/j.tics.2020.09.005

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(2), 179–205. https://doi.org/10.1037/a0036137

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 479–496. https://doi.org/10.1016/S0022-5371(81)90129-8

Voormann, A., Rothe-Wulf, A., Meyer-Grant, C. G., & Klauer, K. C. (in press). Sometimes memory misleads: Variants of the error-speed effect strengthen the evidence for systematically misleading memory signals in recognition memory. *Psychonomic Bulletin & Review*.

Wallach, L., & Wallach, M. A. (2010). Some theories are unfalsifiable: A comment on Trafimow. *Theory & Psychology*, *20*(5), 703–706. https://doi.org/10.1177/0959354310373676

Wallach, M. A., & Wallach, L. (1994). Gergen versus the mainstream: Are hypotheses in social psychology subject to empirical test? *Journal of Personality and Social Psychology*, *67*(2), 233–242. https://doi.org/10.1037/0022-3514.67.2.233

Wallach, M. A., & Wallach, L. (1998). When experiments serve little purposes: Misguided research in mainstream psychology. *Theory & Psychology*, *8*(2), 183–194. https://doi.org/10.1177/0959354398082005

Williams, J. R., Robinson, M. M., & Brady, T. F. (2023). There is no theory-free measure of "swaps" in visual working memory experiments. *Computational Brain & Behavior*, *6*(3), 159–171. https://doi.org/10.1007/s42113-022-00150-5

Winter, K., Menne, N. M., Bell, R., & Buchner, A. (2022). Experimental validation of a multinomial processing tree model for analyzing eyewitness identification decisions. *Scientific Reports*, *12*(1), Article 15571. https://doi.org/10.1038/s41598-022-19513-w

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology*, *105*, 81–114. https://doi.org/10.1016/j.cogpsych.2018.06.001

Appendix A

Example of Ambiguous Evidence Against Selective Influence

Imagine a researcher who intends to utilize a 2HT model for quantifying latent memory and guessing processes in an experiment implementing a 2AFC paradigm that includes a three-level "response-bias" and a two-level memory-strength manipulation. The "response-bias" manipulation is presumed to selectively influence the 2HT model's guessing parameter g (i.e., the probability of selecting the left item given that none of the items was detected). The model only includes a single detection parameter D: $= D_{\text{n}}(1 - D_{\text{o}}) + (1 - D_{\text{n}})D_{\text{o}} + D_{\text{n}}D_{\text{o}}$, which corresponds to the probability that at least one of the two items will be correctly detected. In addition, D is allowed to differ depending on the memory-strength manipulation (i.e., $D^{\text{w}} < D^{\text{s}}$). Furthermore, the researcher defines a "hit" as participants correctly selecting the left item of a pair with the old item on the left, whereas if participants incorrectly select the left item of a pair with the old item on the right, this is considered a "false alarm."

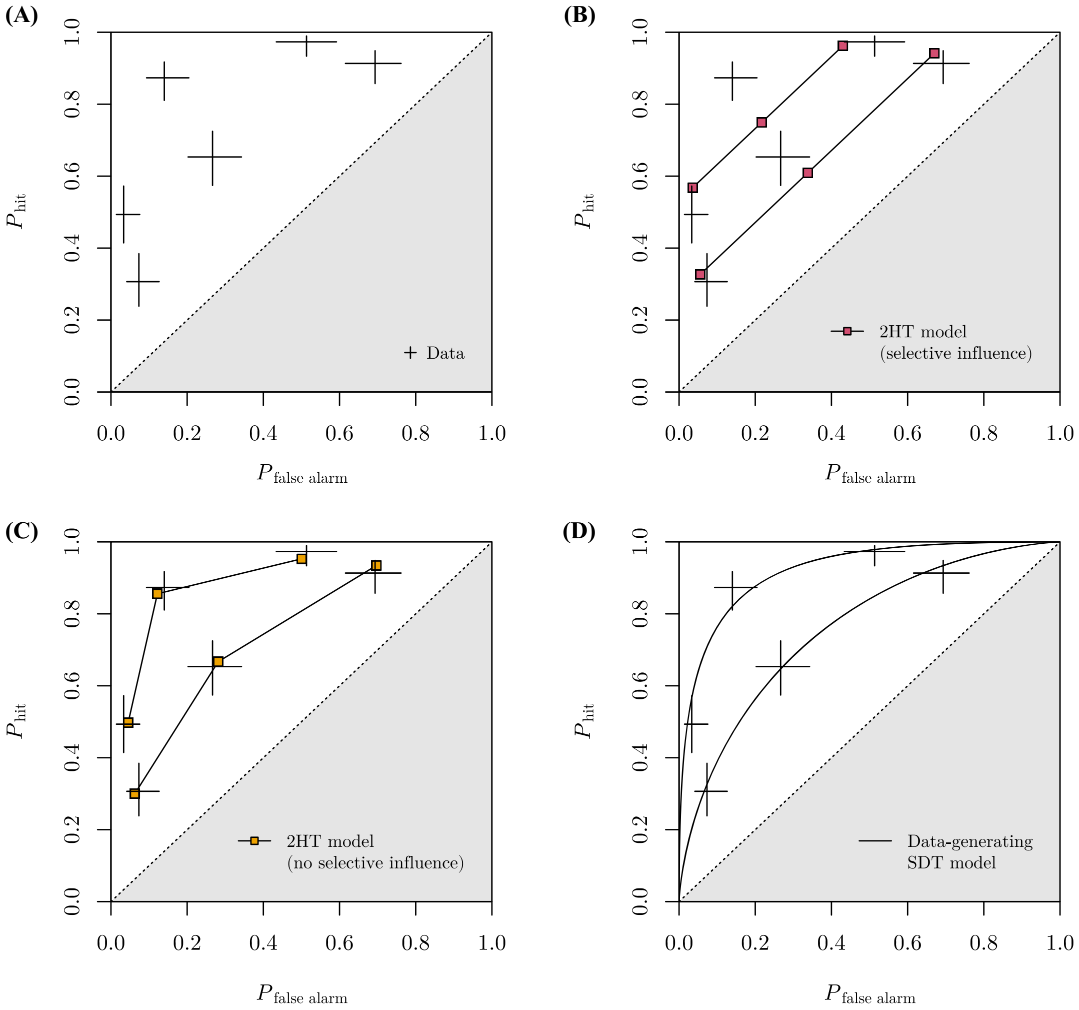
Panel A of Figure A1 depicts hypothetical data obtained from such an experiment in ROC space. As can be seen, the ROC data clearly exhibit nonlinearity. This, however, should lead the researcher to question the empirical adequacy of the 2HT model, as it is bound to predict strictly linear ROCs in the present case (Figure A1, Panel B). Indeed, fitting a 2HT model to the data reveals a clear inability to accurately predict the observed patterns ($G^2(7) = 36.28$, $p < .001$; with $\widehat{D^{\text{w}}} = .27$, $\widehat{D^{\text{s}}} = .53$, and $\hat{g} = (.92, .46, .08)^{\text{T}}$).

But after some contemplation, the researcher might realize that this conclusion hinges on the tacit assumption that the alleged "response-bias" manipulation selectively influences the 2HT model's guessing parameter. In other words, if one would allow for a violation of this assumption, one could reconcile the 2HT account with the recalcitrant data by arguing that the detection probability D is reduced when participants are put in biased conditions. In this case, the ROC data can be well described by multiplying parameter D with a reduction factor

(*Appendices continue*)

Figure A1

Illustration of How Incorrect Model Assumptions Can Lead to Illusory Evidence Against Selective Influence



Note. Panel A depicts hypothetical data of an experiment implementing a two-alternative forced-choice paradigm that includes a three-level "response-bias" and a two-level memory-strength manipulation. In each condition, there were 300 observations. For half of those, the old item was presented on the left, and for the other half, it was presented on the right. The crossed lines correspond to 95% bootstrap confidence intervals. Panel B depicts a fitted 2HT model that enforces selective influence of the "response-bias" manipulation. Panel C depicts a fitted 2HT model that relaxes this assumption by including a reduction factor $0 < \alpha < 1$ that decreases detection performance in biased conditions. Panel D depicts the true data-generating Gaussian SDT model. SDT = signal detection theory; 2HT = two high-threshold. See the online article for the color version of this figure.

$0 \le \alpha \le 1$ (see Figure A1, Panel C): Fitting this modified 2HT model suggests that it can successfully account for the present data ($G^2(6) = 4.70$, $p = .58$; with $\hat{D}^{\text{w}} = .39$, $\hat{D}^{\text{s}} = .73$, $\hat{g} = (.91, .46, .08)^{\text{T}}$, and $\hat{\alpha} = .62$).

Following the line of reasoning advocated by Malejka et al. (2022), the researcher should therefore conclude that selective influence is violated based on the above-outlined analyses. Put differently, the data should apparently be taken to suggest that the "response-bias" manipulation also affects detection accuracy. But as it turns out, this conclusion must not necessarily be correct—in fact, in the present example, *it is not*.

To see why this is the case, consider a reality in which the data were generated by a Gaussian SDT model. Moreover, suppose that selective influence holds true; that is, the "response-bias" manipulation only affects the position of the SDT model's response criterion κ. Such a model will predict curvilinear ROC shapes that match the ROC data depicted in Figure A1. By now, it might come as a no surprise that the hypothetical data considered thus far were indeed generated by that very model (see Figure A1, Panel D; with $d'^{\text{w}} = 0.5$, $d'^{\text{s}} = 1.0$, and $\kappa = (-1, 0, 1)^{\text{T}}$). This (artificial) reality entails that the researcher's conclusion regarding selective influence is factually incorrect, despite the supporting test results obtained.

The problem encountered here, which is the same found in the case of Malejka et al. (2022), is the treatment of the core assumptions of the 2HT model (such as the high-threshold

(*Appendices continue*)

assumption) as axioms; that is, these assumptions are not scrutinized and simply taken at face value. In the present example, however, they happen to be violated. As a consequence, the researcher erroneously clings to an untenable model (viz. the 2HT model) by mistaking evidence against it for evidence against selective influence. Of course, no modeler is exempt from this problem; after all, models are always approximations of an unknown but in all likelihood more complex reality. This is why the comparison with known competitors is so important; it allows one to scrutinize the relative merits of the unique accounts that they offer.

Appendix B

Proof of Familiarity-Contrast Predictions

First, we recall that—according to both the 2HT and the 2HTC model—the probability of a detection hit (i.e., a correct "old item present" response) in a situation where two test items (an old and a new one) are simultaneously presented is given by

$$P_{\text{hit}} = D_{\text{o}} + (1 - D_{\text{o}})D_{\text{n}}g_{\{0,1\}} + (1 - D_{\text{o}})(1 - D_{\text{n}})g_{\{0,0\}}, \quad \text{(B1)}$$

where $g_{\{0,0\}}$ is the guessing parameter (i.e., the conditional probability of giving a "old item present" response) if both items entered an uncertainty state and $g_{\{0,1\}}$ is the guessing parameters for the situation in which only one of the items (i.e., the old one) entered an uncertainty state while the new item was correctly detected as such (Meyer-Grant & Klauer, 2021). Furthermore, Meyer-Grant and Klauer (2021, Proposition 10) established that $g_{\{0,0\}} < g_{\{0,1\}}$ must hold in order to account for apparent patterns in their data.

As already discussed earlier, it follows from the contrast mechanism at the heart of the 2HTC model that a manipulation of new item similarity should not only affect D_{n}, but also D_{o}. More specifically, decreasing D_{n} by making a new item systematically similar to an old one (Meyer-Grant & Klauer, 2021) should also decrease D_{o} (Malejka et al., 2022, Appendix A).

However, it turns out that both a decrease in D_{n} *and* a decrease in D_{o} can each only reduce the hit probability predicted by the 2HTC model for the detection tasks conducted by Meyer-Grant and Klauer (2021). To see why this is the case, we first take the partial derivative of Equation B1 with respect to D_{n} which yields

$$\frac{\partial P_{\text{hit}}}{\partial D_{\text{n}}} = (1 - D_{\text{o}})(g_{\{0,1\}} - g_{\{0,0\}}). \quad \text{(B2)}$$

Clearly, the right-hand side of Equation B2 is always positive since $g_{\{0,0\}} < g_{\{0,1\}}$ and $D_{\text{o}} < 1$. Thus, Equation B1 is strictly increasing in D_{n}. Next we also take the partial derivative of Equation B1 with respect to D_{o} which yields

$$\frac{\partial P_{\text{hit}}}{\partial D_{\text{o}}} = 1 - D_{\text{n}}g_{\{0,1\}} - (1 - D_{\text{n}})g_{\{0,0\}}. \quad \text{(B3)}$$

In order for Equation B1 to be strictly increasing in D_{o}, it must consequently always hold that

$$1 - D_{\text{n}}g_{\{0,1\}} - (1 - D_{\text{n}})g_{\{0,0\}}) > 0. \quad \text{(B4)}$$

After some algebraic manipulations where we take into account that $g_{\{0,0\}} - g_{\{0,1\}} < 0$, we find that Equation B4 is equivalent to

$$\frac{g_{\{0,0\}} - 1}{g_{\{0,0\}} - g_{\{0,1\}}} > D_{\text{n}}. \quad \text{(B5)}$$

But this is clearly always true since the left-hand side of Equation B5 must always be greater than one and $D_{\text{n}} < 1$.

Received December 11, 2023
Revision received October 4, 2024
Accepted October 9, 2024 ■