Journal of Experimental Psychology: General

Ranking Tasks in Recognition Memory: A Direct Test of the Two-High-Threshold Contrast Model

Constantin G. Meyer-Grant and Marie Jakob

Online First Publication, January 27, 2025. https://dx.doi.org/10.1037/xge0001700

CITATION

Meyer-Grant, C. G., & Jakob, M. (2025). Ranking tasks in recognition memory: A direct test of the two-high-threshold contrast model. *Journal of Experimental Psychology: General*. Advance online publication. https://dx.doi.org/10.1037/xge0001700

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2025 American Psychological Association
ISSN: 0096-3445

https://doi.org/10.1037/xge0001700

BRIEF REPORT

Ranking Tasks in Recognition Memory: A Direct Test of the Two-High-Threshold Contrast Model

Constantin G. Meyer-Grant and Marie Jakob
Department of Psychology, University of Freiburg

It has long been debated whether latent memory signals determine recognition judgments directly or through a small number of discrete states. Often, signal detection theory (SDT) models instantiate the former perspective, whereas the two-high-threshold (2HT) model instantiates the latter. Kellen and Klauer (2014) conducted a critical test using a ranking paradigm that yielded results in line with common SDT models and incompatible with the 2HT model. However, Malejka et al. (2022) recently challenged their conclusion. They argued that the 2HT model can account for the critical effect if detection probabilities were determined by a memory-signal contrast between simultaneously presented stimuli. Here, we test this contrast mechanism directly. We show that when only a single old item is presented, such a contrast mechanism entails a decrease in the probability of correctly rejecting the accompanying new items as their number increases. SDT models, on the other hand, predict the opposite pattern. Results of an empirical investigation were in agreement with SDT and inconsistent with the 2HT contrast model. Thus, our findings strengthen the conclusions of Kellen and Klauer (2014) and provide further evidence for SDT models of recognition memory.

Public Significance Statement

This study investigates whether recognition involves a comparison between jointly encountered items. In particular, we test a theoretical proposal according to which a person only recognizes a known item when it is sufficiently more familiar than the other items and, analogously, only realizes that an item is unknown when it is sufficiently less familiar. Otherwise, people are assumed to be completely unaware of whether an item is known or unknown. However, the findings of our study speak against these ideas. Instead, they support the view that people can access more fine-grained memory information. Since people are often required to recognize multiple items simultaneously in real life (e.g., when making eyewitness-identification judgments), this has various practical implications.

Keywords: recognition memory, signal detection theory, high-threshold model, contrast mechanism, ranking

Ashleigh Maxcey served as action editor.

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

Marie Jakob https://orcid.org/0009-0001-6116-9702

Constantin G. Meyer-Grant and Marie Jakob contributed equally to this work. All data collected for this project, analysis scripts, and materials are available at https://osf.io/mce6n/. The reported results have previously been presented at the 57th Annual Meeting of the Society for Mathematical Psychology and the 22nd International Conference on Cognitive Modeling.

The authors have no conflicts of interest to disclose. Marie Jakob received support from the research training group Statistical Modeling in Psychology (Grant 310365261), which is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG). The authors thank Michel Regenwetter, Clintin Davis-Stober, David Kellen, and Karl Christoph Klauer for helpful discussions surrounding the present work.

Constantin G. Meyer-Grant played a lead role in conceptualization, supervision, visualization, and writing–original draft, a supporting role in software, and an equal role in formal analysis, methodology, project administration, validation, and writing–review and editing. Marie Jakob played a lead role in data curation, funding acquisition, investigation, and software, a supporting role in writing–original draft, and an equal role in formal analysis, methodology, project administration, validation, and writing–review and editing.

Correspondence concerning this article should be addressed to Constantin G. Meyer-Grant, Department of Psychology, University of Freiburg, Engelbergerstr. 41, 79085 Freiburg, Germany. Email: constantin.meyer-grant@psychologie.uni-freiburg.de

How latent memory information influences recognition judgments is the subject of a long-standing debate in the field (see, e.g., Batchelder & Alexander, 2013; Bröder & Schütz, 2009; Dubé et al., 2012, 2013; Dubé & Rotello, 2012; Province & Rouder, 2012). According to one viewpoint, these judgments are based on a direct assessment of latent memory signals, indicating that decision makers have immediate access to continuous memory representations. Proponents of an alternative position postulate that latent memory information is mediated by a small number of discrete mental states, so that a decision maker either correctly realizes that an item is (un)known or is completely oblivious as to whether it has previously been encountered.

These two divergent perspectives are commonly represented by the unequal-variance Gaussian (UVG) model (Egan, 1958)—a specific parametrization of the signal detection theory (SDT) model framework (Green & Swets, 1966)—and the two-high-threshold (2HT) model (Snodgrass & Corwin, 1988), respectively. Thus, researchers have often defaulted to a juxtaposition of UVG and 2HT models in terms of various goodness-of-fit measures (e.g., Bröder & Schütz, 2009; Dubé & Rotello, 2012; Kellen et al., 2013; Klauer & Kellen, 2015) as a means to compare the two competing accounts. In recent years, however, this approach has been repeatedly criticized (see Birnbaum, 2011; Kellen, 2019; Pitt & Myung, 2002; Roberts & Pashler, 2000), primarily for its inability to disentangle a model's core theoretical principles from its auxiliary assumptions. This, in turn, has caused a shift toward *critical testing* (e.g., Chechile & Dunn, 2021; Kellen et al., 2021; Kellen & Klauer, 2015; Ma et al., 2022; Meyer-Grant & Klauer, 2021; Starns et al., 2018; for a relevant discussion, see also Kellen et al., in press).

In 2014, for instance, Kellen and Klauer implemented a critical test contrasting SDT and 2HT models by conducting two experiments that utilized a ranking paradigm in combination with a manipulation of memory strength. In the initial study phase of their experiments, participants were asked to memorize a set of words, some of which were presented once (also termed "weak items") and others were presented three times (also termed "strong items"). In the subsequent test phase, each trial comprised multiple words presented simultaneously (i.e., four and three words in Experiments 1 and 2, respectively). Of these, one word was an old item from the initial study list, and the remaining words were new items. Participants were informed of this setup and instructed to rank the words according to their subjective assessment of how likely it was that they had encountered the respective words before.

In this context, SDT and 2HT models make contradictory predictions regarding the probability of the old item being assigned rank two, conditional on it not being assigned rank one (henceforth denoted by c_2). In the SDT framework, c_2 equals the probability of the old-item memory signal being surpassed by one and only one new-item memory signal, given that it is not the largest one. Thus, this probability is directly influenced by old-item memory strength, which typically results in SDT models predicting that $c_2^{\text{weak}} < c_2^{\text{strong}}$.

According to the 2HT model, on the other hand, c_2 only depends on the parameter D_n (see Kellen et al., in press; Kellen & Klauer, 2014; Malejka et al., 2022; see also Appendix A), the probability of correctly rejecting a new item. However, a manipulation of old-item memory strength, as implemented in Kellen and Klauer (2014), is typically assumed to selectively influence the probability of recognizing an old item (i.e., the parameter D_o) while leaving D_n unaffected (Kellen et al., 2015). Building on this assumption, Kellen and Klauer (2014) pointed out that the 2HT model must predict that $c_2^{\text{weak}} = c_2^{\text{strong}}$. Their empirical results, however, turned out to be inconsistent with this prediction but agreed with the prediction of SDT, which Kellen and Klauer (2014) interpreted as evidence against the 2HT model.

The 2HT Contrast Model

Malejka et al. (2022) recently challenged this line of argument. They conjectured that, in the presence of multiple test items, a contrast mechanism determines the probabilities with which old and new items enter the correct detection state. More precisely, they proposed that when $K \geq 2$ items are presented simultaneously, decision-makers first contrast the latent memory signal of the kth item (i.e., ψ'_k, where $k \in \{1, \ldots, K\}$) with a weighted average of the remaining $K - 1$ latent memory signals. This yields the memory-signal contrast

$$\psi_k = \psi'_k - \sum_{j \neq k} w_j \psi'_j, \tag{1}$$

where $w_j \geq 0$ and $\sum w_j = 1$ with $w_j = 1/(K-1)$ being a natural choice. According to Malejka et al. (2022), an old item is correctly detected whenever its memory-signal contrast ψ_o exceeds an upper threshold h_u, such that

$$D_\text{o} = P(\psi_\text{o} > h_\text{u}). \tag{2}$$

Analogously, a new item is rejected whenever its memory-signal contrast ψ_n falls below a lower threshold h_l, such that

$$D_\text{n} = P(\psi_\text{n} < h_\text{l}). \tag{3}$$

As a consequence, the conclusion of Kellen and Klauer (2014) is questioned, insofar as a manipulation of old-item memory strength would no longer exclusively affect D_o, but also D_n and, by extension, c_2 (Malejka et al., 2022).

In response to this, Kellen et al. (in press) pointed out that although such a modification saves the model from the critical test of Kellen and Klauer (2014), it also leads to several undesirable properties when it comes to establishing a unified theory for different recognition memory tasks. Furthermore, these authors criticized Malejka et al. (2022) for not directly testing the proposed contrast mechanism.[1] The present work aims to address this issue by implementing such a test with the goal of providing a more definitive answer as to whether a contrast mechanism in ranking tasks is empirically warranted.

A Direct Test of the Contrast Mechanism

As we will demonstrate in the following, the novel mechanism at the heart of Malejka et al.'s (2022) 2HT contrast (2HTC) model implies a characteristic change in certain model parameters when varying K (i.e., the number of test items presented simultaneously)—given that the set always contains only a single old item. In particular, the probability of

[1] Malejka et al. (2022) verified that the D_n parameter behaves as predicted by the proposed contrast mechanism when old-item memory strength was manipulated. However, as argued by Kellen et al. (in press), they did not test critical predictions beyond the effect that the contrast mechanism was specifically designed to account for.

correctly rejecting a new item (i.e., D_n) should tend to decline the more new items are presented per set.

To elucidate the effect of K on D_n, suppose that ψ_k is the memory-signal contrast of a new item. Hence, $\psi'_k \sim \psi'_\text{n}$ in the right-hand side of Equation 1, where ψ'_n is the latent strength of a new item. In that case, the weighted sum of the remaining memory signals (i.e., $\sum_{j \neq k} w_j \psi'_j$) includes exactly one memory signal belonging to an old item (henceforth also denoted by ψ'_o) and $K - 2$ new-item memory signals. By implication, increasing K diminishes the influence of the old item on the average memory signal, as it is progressively outnumbered by new items. But this results in an increase of the expected memory-signal contrast for a new item (ψ_n), which in turn leads to a decrease in D_n.

For illustrative purposes, let us consider a concrete example in which $\psi'_\text{n} \sim \mathcal{U}(0, 1)$ and $\psi'_\text{o} \sim \mathcal{U}(1/2, 2)$; that is, they are both uniformly distributed (see Figure 1, top-left panel).[2] As depicted in the top-right panel of Figure 1, the average memory signal will tend to decrease as K increases, resulting in a decrease in D_n (see Figure 1, bottom panel).

To test this prediction, we conducted a recognition experiment where we combined a ranking paradigm with a manipulation of the number of simultaneously presented test items, implementing three different conditions (i.e., $K \in \{3, 4, 5\}$). We then leveraged the fact that c_2 is a one-to-one function of D_n, irrespective of the value of K. Thus, inverse functions exist that uniquely map permissible values of c_2 to D_n, which we will denote as $t^{K=3}: [1/2, 1] \mapsto [0, 1]$, $t^{K=4}: [1/3, 1] \mapsto [0, 1]$, and $t^{K=5}: [1/4, 1] \mapsto [0, 1]$. When $K = 3$, for example, it holds that

$$t^{K=3}(c_2^{K=3}) = \frac{1 - 2c_2^{K=3}}{c_2^{K=3} - 2}, \tag{4}$$

for which we provide a detailed derivation in Appendix A alongside closed-form expressions for $t^{K=4}(c_2^{K=4})$ and $t^{K=5}(c_2^{K=5})$ (see Equations A8 and A9). The functions $t^{K=3}$, $t^{K=4}$, and $t^{K=5}$ are depicted in Figure 2. Critically, if the contrast mechanism suggested by Malejka et al. (2022) determines D_n, we should expect to observe $t^{K=3}(\hat{c}_2^{K=3}) > t^{K=4}(\hat{c}_2^{K=4}) > t^{K=5}(\hat{c}_2^{K=5})$, where $\hat{c}_2^{K=3}$, $\hat{c}_2^{K=4}$, and $\hat{c}_2^{K=5}$ denote the relative frequencies of assigning rank two to the old item conditional on it not being assigned rank one and given that K equals 3, 4, and 5, respectively.

Method

Participants

Sixty native German speakers (45 female, 14 male, one diverse) between the ages of 19 and 31 ($M_\text{age} = 21.45$, $SD_\text{age} = 2.86$) participated in the experiment. Participants either received partial course credit or a base payment of € 6, as well as an additional performance-dependent reward of up to € 4.

Materials

The stimulus pool consisted of 930 images generated by a generative adversarial network (Karras et al., 2019), half of which showed male faces and the other half female faces. Images were presented on a black background.

Procedure

The experiment began with a study phase followed by a test phase. For every participant, 168 images (comprising an equal number of images showing male and female faces) were randomly drawn from the stimulus pool to be presented in the study phase. Of the remaining images, 504 were randomly drawn to serve as new items during test. In the study phase, 168 images were sequentially presented for 2,000 ms each, separated by an interstimulus interval of 800 ms. The presentation order was randomized.

In each trial of the subsequent test phase, a single item from the study phase was presented together with two, three, or four accompanying new items (see Figure 3). The position of the old item was randomized, and all shown faces were of the same gender. Participants were asked to rank the images according to how much they believed they had seen them during study (i.e., they were to assign rank one to the image they considered most likely to have been studied, and so forth). The test phase consisted of four blocks that each comprised 42 trials. Between the blocks, participants could take a self-paced break. The presentation order in the test phase was also randomized with the restriction that the number of trials showing men and women and trials containing three, four, and five images was balanced across blocks. Participants received points depending on the rank they assigned to the old image. The total number of points determined their performance-dependent reward.

Transparency and Openness

All data, analysis scripts, and materials are available at https://osf.io/mce6n/. This study was not preregistered.

Results

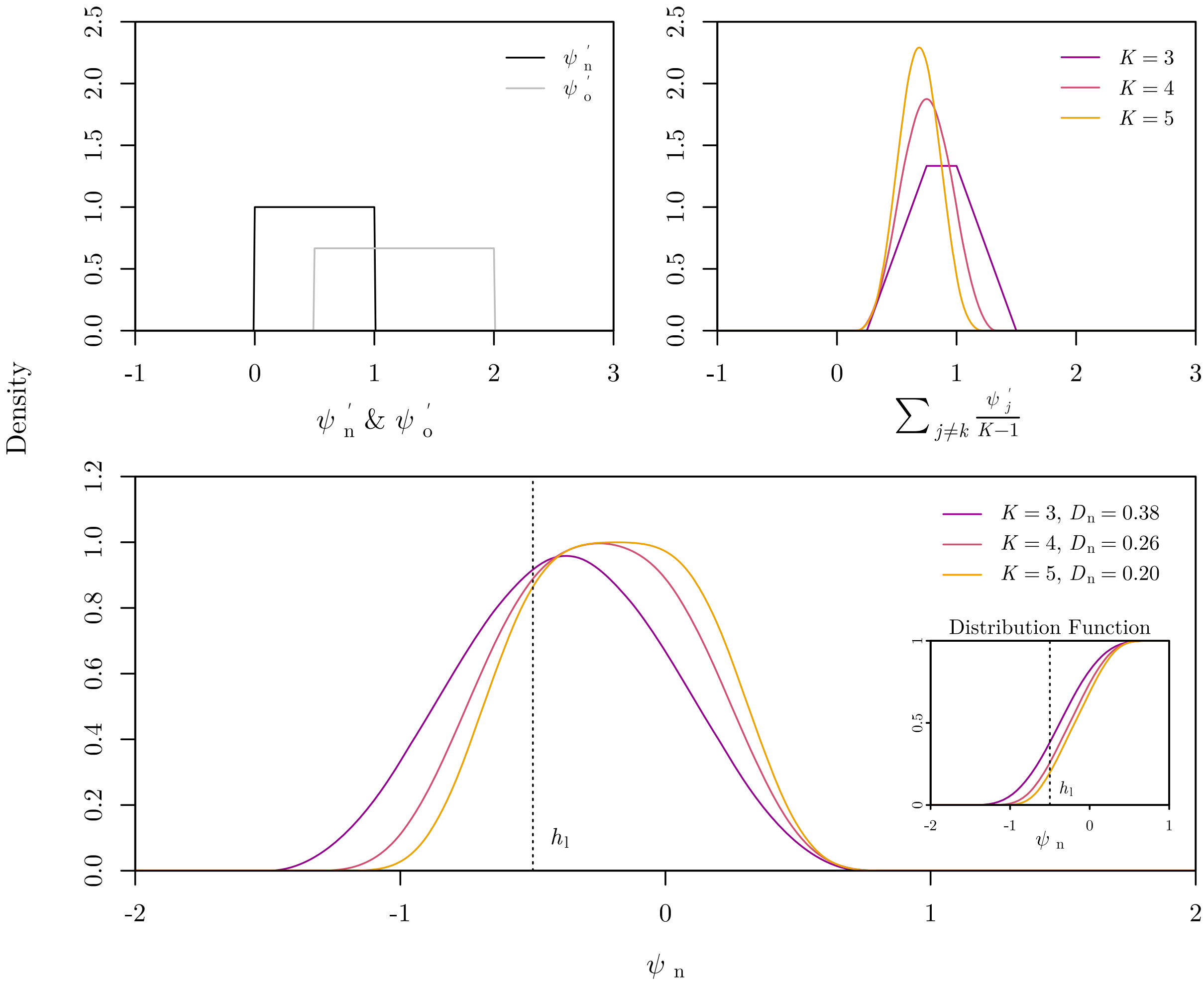
In order to analyze the empirical relationship of $t^{K=3}(\hat{c}_2^{K=3})$, $t^{K=4}(\hat{c}_2^{K=4})$, and $t^{K=5}(\hat{c}_2^{K=5})$, we first used a generalized linear mixed model with participants as random-effect factor and with the number of simultaneously presented test items as within-subjects fixed-effect factor.[3] There was strong evidence for the presence of an effect, $\chi^2_\text{LR}(2) = 46.25$, $p < .001$. Contrasts comparing $t^{K=3}(\hat{c}_2^{K=3})$ to

[2] Taken by itself, the above-introduced contrast mechanism allows for old items being mistaken for new and new items being mistaken for old (i.e., $P(\psi_\text{o} < h_\text{l}) \geq 0$ and $P(\psi_\text{n} > h_\text{u}) \geq 0$). This is because the probability of entering one or the other state depends only on the memory-signal contrast; otherwise, the system would have to know whether an item is old or new independent of the available memory information. However, Malejka et al. (2022) stipulated that in "line with 2HT theory, both thresholds are assumed to be 'high' in the sense that only [memory-signal contrasts of old items] may exceed h_u and only [memory-signal contrasts of new items] may fall below h_l" (Malejka et al., 2022, p. 18). One can attempt to preserve this property by assuming that the distributions of old-item and new-item memory signals (i.e., their supports) are, at least partially, nonoverlapping. In our example (see also Figure 1), this ensures that the memory-signal contrast of an old item can never fall below $h_\text{l} = -0.5$. While the somewhat arbitrary nature of such a construction is evident, the present work is not the place to scrutinize its theoretical merits (for an in-depth discussion, see Kellen et al., in press). Here, we will simply take this premise at face value.

[3] Some of the observed $\hat{c}_2$ values were below chance level, which would translate into negative (or sometimes even complex) values for D_n. Leaving aside the fact that this observation alone is already inconsistent with the 2HT/2HTC model, we have treated these cases by setting D_n to the most likely permissible value given the observed value of $\hat{c}_2$ (viz., $D_\text{n} = 0$).

Figure 1

Latent Memory-Signal and Contrast Distributions According to the 2HTC Model



Note. Top left panel: Latent memory-signal distributions of new (ψ'_n) and old items (ψ'_o) with $\psi'_n \sim \mathcal{U}(0,1)$ and $\psi'_o \sim \mathcal{U}(1/2,2)$. Top right panel: Distributions of the average memory signals $\sum_{j \neq k} \frac{\psi'_j}{K-1}$ that is subtracted from a new item's memory signal ψ'_n in the contrast calculation (see Equation 1) depending on $K \in \{3,4,5\}$. Bottom panel: Contrast densities and distribution functions (subplot) for a new item ψ_n with $h_l = -0.5$ for $K \in \{3,4,5\}$. 2HTC = two-high-threshold contrast. See the online article for the color version of this figure.

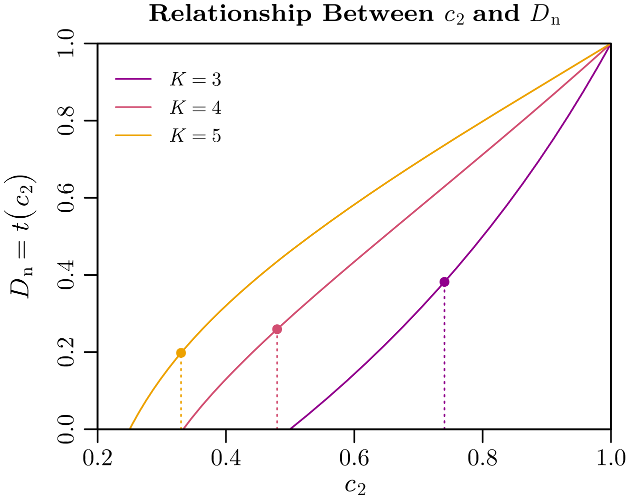
$t^{K=4}(\hat{c}_2^{K=4})$—$\log(OR) = 0.53$, 95% CI [0.31, 0.74], $\chi^2_{LR}(1) = 23.25$, $p < .001$—and $t^{K=4}(\hat{c}_2^{K=4})$ to $t^{K=5}(\hat{c}_2^{K=5})$—$\log(OR) = 0.18$, 95% CI [0.00, 0.35], $\chi^2_{LR}(1) = 3.69$, $p = .055$—suggest, however, that the pattern is exactly opposite to what is predicted by the 2HTC model (see Figure 4).

We additionally tested the order constraint $t^{K=3}(\hat{c}_2^{K=3}) \geq t^{K=4}(\hat{c}_2^{K=4}) \geq t^{K=5}(\hat{c}_2^{K=5})$ using a Bayesian-hierarchical modeling approach (see Appendix B for details). Comparing the model that enforces the order constraint with an unconstrained model revealed strong evidence against the order-constrained model, as indicated by a Bayes factor of 0.007 (a Bayes factor below one implying a preference for the unconstrained model).

Finally, we also tested the critical constraint on c_2 directly. This was done by calculating a Bayes factor for each participant, which compared a model that allowed for all theoretically plausible values of c_2 (i.e., values that were above guessing level; see also Footnote 3) to one that additionally required the order constraint $t^{K=3}(c_2^{K=3}) \geq t^{K=4}(c_2^{K=4}) \geq t^{K=5}(c_2^{K=5})$ to be satisfied (see Appendix C for details). Results clearly suggest that the order constraint is violated in the data: The product of individual Bayes factors amounted to roughly 3.30×10^{-34}, which decisively rejects the hypothesis that the order constraint holds universally. Of the 60 Bayes factors (with an average Bayes factor of 0.71), 46 provided some evidence against the order-constrained model, and for 24 of them, the evidence was at least moderate (i.e., a Bayes factor smaller than 1/3; see Lee & Wagenmakers, 2014). By contrast, only one of the individual Bayes factors provided more than anecdotal evidence in favor of the order-constrained model.

Figure 2

Inverse Functions $t^{K=3}$, $t^{K=4}$, and $t^{K=5}$, Mapping Permissible Values of c_2 to D_n According to the 2HT/2HTC Model



Note. Increasing K may result in a decrease in c_2, even in the absence of a decrease in D_n. In other words, $c_2^{K=3} > c_2^{K=4} > c_2^{K=5}$ does not guarantee that $D_\mathrm{n}^{K=3} > D_\mathrm{n}^{K=4} > D_\mathrm{n}^{K=5}$. Indeed, the 2HTC model not only predicts a decrease in c_2 but additionally expects this decrease to be quite pronounced. For example, under the values for D_n derived in Figure 1 (i.e., 0.38, 0.26, and 0.20 for K being equal to 3, 4, and 5, respectively), one would predict $c_2^{K=3} = 0.74$, $c_2^{K=4} = 0.48$, and $c_2^{K=5} = 0.33$ (as indicated by the points and dashed lines in the present figure). 2HT = two-high-threshold; 2HTC = two-high-threshold contrast. See the online article for the color version of this figure.

Discussion

The results of our analyses indicate a clear qualitative divergence between the predictions of the 2HTC model and the empirical data patterns. That is, the observed values of $\hat{c}_2^{K=3}$, $\hat{c}_2^{K=4}$, and $\hat{c}_2^{K=5}$ seem to conflict with the idea that the above-described contrast mechanism determines the detection parameters of a 2HT model in ranking tasks. This, however, introduces a dilemma: If we follow Malejka et al.'s (2022) theoretical proposal and modify the model so that it can account for Kellen and Klauer's (2014) data, it is coerced into making predictions that run counter to the effects observed in the present study. Thus, our findings not only cast doubt on the contrast mechanism proposed by Malejka et al. (2022) but—when combined with the results reported by Kellen and Klauer (2014)—on the 2HT model in general. At a minimum, this calls for an alternative causal mechanism that explains why manipulating memory strength of an old item should affect the probability of correctly rejecting a new item.

Comparison Between the 2HTC and UVG Model

In light of these considerations, one might ask how the competing SDT model framework fares in comparison. It turns out that if ranking data were generated by a UVG model, for example, we would typically expect to find a relationship between $\hat{c}_2^{K=3}$, $\hat{c}_2^{K=4}$, and $\hat{c}_2^{K=5}$ such that $t^{K=3}(\hat{c}_2^{K=3}) < t^{K=4}(\hat{c}_2^{K=4}) < t^{K=5}(\hat{c}_2^{K=5})$ (see Figure 5). Interestingly, this is opposite to the predictions of the 2HTC model and precisely the kind of relationship that our data seem to suggest. Simply put, the empirical patterns that are at best perplexing under the 2HTC model are naturally predicted by the rival SDT account. To quantify this evident disparity in their ability to account for the present data, we turned to model-based analyses of the aggregated data, juxtaposing the 2HTC and UVG model.

The UVG model has only two parameters. This allows us to fit the model to the frequencies with which the targets were assigned the first rank versus the frequencies with which they were not. Importantly, these quantities are orthogonal to $c_2^{K=3}$, $c_2^{K=4}$, and $c_2^{K=5}$ because of conditioning. In other words, the UVG model can make a generalizing prediction for $c_2^{K=3}$, $c_2^{K=4}$, and $c_2^{K=5}$ based on independent data; that is, without being directly informed by the corresponding observations (Busemeyer & Wang, 2000). Results are illustrated in Figure 5, which reveals that the predictions of the UVG accurately capture the observed data patterns. This not only underscores the model's ability to adequately describe the empirical findings but also shows that the model is highly consistent in its ability to "accurately interpolate and extrapolate" (Busemeyer & Wang, 2000, p. 179).

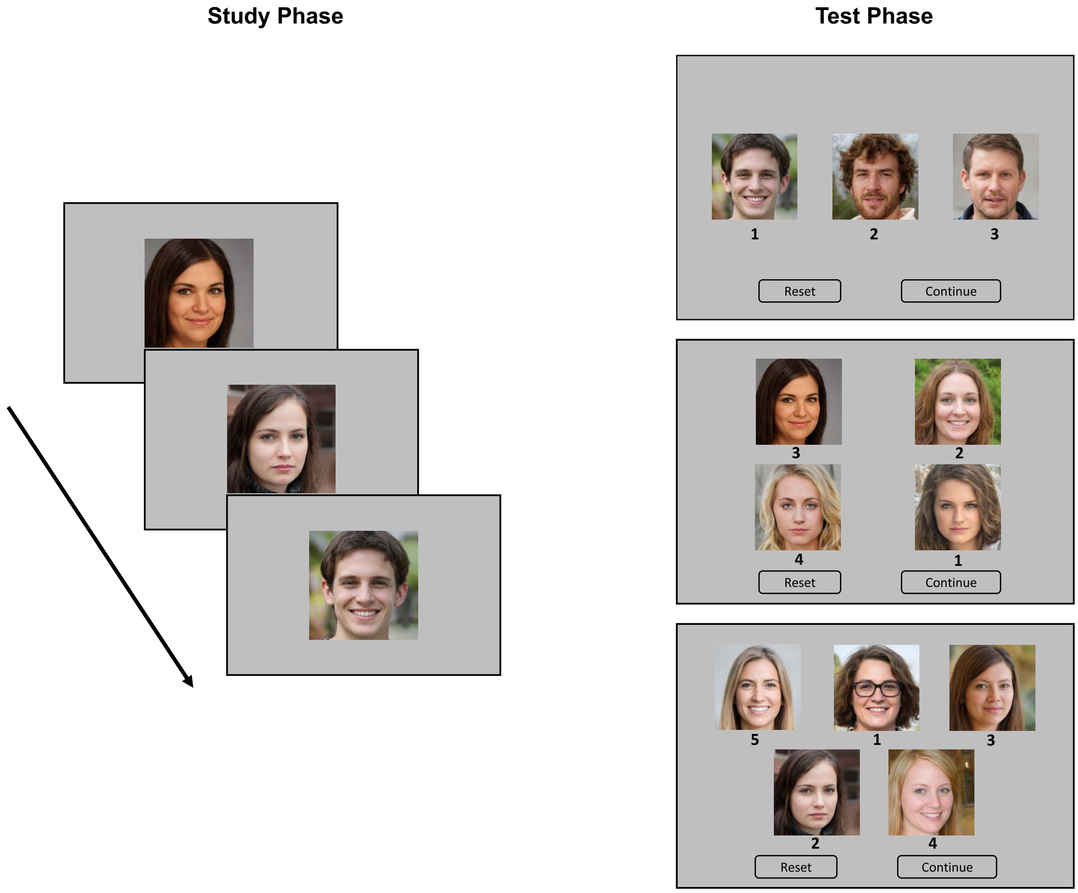
For the 2HTC model, more parameters must be estimated, which renders the generalization approach unfeasible. Nonetheless, we can assess whether the model is able to account for the observed data patterns at all. We decided against explicitly modeling the actual contrast mechanism (which would require auxiliary assumptions and thereby impose additional constraints on the parameters of the 2HT model) and merely enforced that $D_\mathrm{n}^{K=3} \geq D_\mathrm{n}^{K=4} \geq D_\mathrm{n}^{K=5}$. As already established above, the observed data lead to values for $t^{K=3}(\hat{c}_2^{K=3})$, $t^{K=4}(\hat{c}_2^{K=4})$, and $t^{K=5}(\hat{c}_2^{K=5})$ that are clearly at odds with this restriction and, as a consequence, also with the estimated parameters ($\hat{D}_\mathrm{n}^{K=3} = .067, \hat{D}_\mathrm{n}^{K=4} = .067, \hat{D}_\mathrm{n}^{K=5} = .061$; see small light gray bars in the subplot of Figure 5).

Limitations

Despite the persuasiveness of these results, it is important to emphasize that our line of reasoning rests on specific theoretical premises, and it might be argued that some of them should be put up for discussion. For instance, one could contest the implicitly made assumption that the lower threshold h_l does not depend on K. For theoretical reasons, however, we find that objection unconvincing: Unlike response criteria in SDT models, thresholds in the 2HT/2HTC model do not demarcate what a decision maker considers sufficient evidence to make a decision. Instead, they correspond to the levels of latent memory information at which the decision maker becomes (un)aware of the item's true status (i.e., it being either old or new). Given the nonstrategic nature of these thresholds, we have difficulties conceiving a good explanation for why they should depend on K (see also Kellen et al., in press, who pointed out other problems associated with condition-dependent thresholds).

Another problem could be seen in the assumption that the individual memory signals are given equal weighting when calculating the weighted average $\sum_{j \neq k} w_j \psi'_j$ (i.e., $w_j = 1/(K-1)$ for all $j \neq k$). Indeed, w_j could, in principle, depend on the item's latent memory signal ψ'_j. But this seems to conflict with the assumption that continuous memory information cannot be directly accessed by the decision maker. Put differently, such a move would require a convincing theoretical justification for why latent memory signals

Figure 3
Illustration of the Experimental Design



Note. In each trial of the test phase, three, four, or five images were presented. The target was ranked first in the test trial depicted in the top display, third in the test trial depicted in the middle display, and second in the test trial depicted in the bottom display. See the online article for the color version of this figure.

directly influence the weighting of items but not the actual decision process.

Aside from these conceptual problems, both ideas also face a critical empirical obstacle: When fitting the 2HT model without order constraint on the D_n parameters, estimates are almost unchanged. In fact, when simply enforcing D_o and D_n to remain constant between conditions, goodness-of-fit is not significantly impaired ($\hat{D}_o = .227$, $\hat{D}_n = .065$; $\chi^2_{LR}(4) = 1.79$, $p = .77$). Considering the inconsistency with the estimates of D_n derived via $\hat{c}_2^{K=3}$, $\hat{c}_2^{K=4}$, and $\hat{c}_2^{K=5}$, this implies that internal conflicts arise when the model tries to account for different facets of the entire data set. Thus, it seems that the 2HT model is generally irreconcilable with the present data—with or without a contrast mechanism.

Conclusion

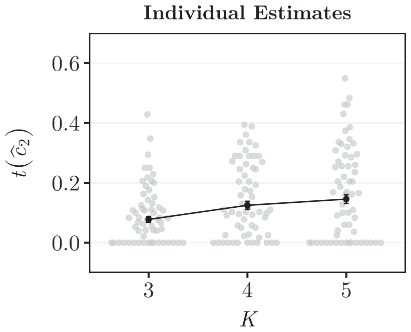
In summary, the results of the present study grant several novel insights: First, the 2HT model seems to be ill-equipped to appropriately handle data obtained from a ranking task in which K is manipulated. Second, our empirical findings speak against the notion that a straightforward contrast mechanism underlies detection whenever multiple items are tested simultaneously. Taken together, this clearly dispels the concern that overlooking such a mechanism has been responsible for the rejection of the 2HT model by Kellen and Klauer (2014; cf. Malejka et al., 2022). The present study thereby renews confidence in the conclusion drawn by these authors (see also Kellen et al., in press). Moreover, the UVG model appears to provide a tenable, consistent, and parsimonious characterization of the present data. This further bolsters the empirical support for SDT-based models in the context of recognition memory (Kellen et al., 2021).

Constraints on Generality

Our sample consisted of predominantly young, female undergraduate students of the University of Freiburg. Previous research has demonstrated quantitative memory differences depending on age and gender (Fraundorf et al., 2019; Herlitz & Lovén, 2013). However, we are not aware of any theoretical or empirical arguments pointing toward corresponding *qualitative* differences with respect to the nature of the involved cognitive processes, which is why we expect our results to generalize to samples of different

Figure 4

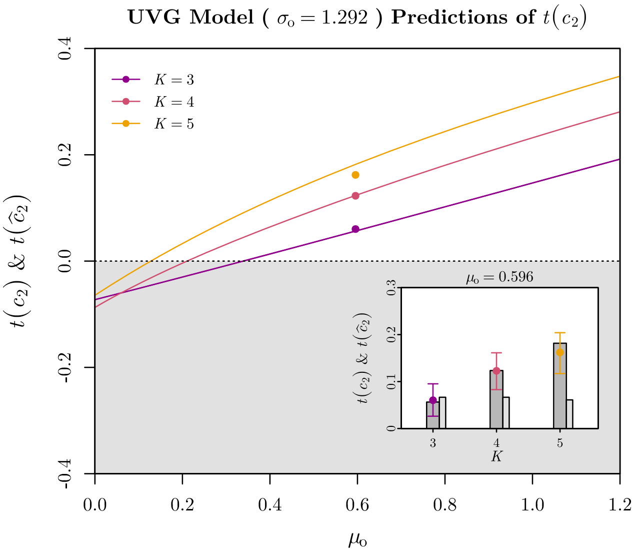
Estimates of $t(\hat{c}_2)$ as a Function of K



Note. Gray dots represent individual estimates and black dots population estimates. Error bars show ±1 *SE* (model-based). The means of the individual estimates of $\hat{c}_2^{K=3}$, $\hat{c}_2^{K=3}$, and $\hat{c}_2^{K=3}$ are 0.548, 0.397, and 0.317, respectively; the means of the individual estimates of $t^{K=3}(\hat{c}_2^{K=3})$, $t^{K=4}(\hat{c}_2^{K=3})$, and $t^{K=5}(\hat{c}_2^{K=3})$ are 0.095, 0.143, and 0.170, respectively, and the model-based population estimates for $t^{K=3}(\hat{c}_2^{K=3})$, $t^{K=4}(\hat{c}_2^{K=3})$, and $t^{K=5}(\hat{c}_2^{K=3})$ are 0.078, 0.125, and 0.145, respectively. SE = standard error.

Figure 5

Predictions of the UVG Model for $t(c_2)$ as a Function of the Mean of the Old Item Distribution μ_o



Note. Standard deviation has been fixed to the maximum-likelihood estimate ($\sigma_o = 1.292$). Points indicate $t^{K=3}(\hat{c}_2^{K=3} = 0.544) = 0.060$, $t^{K=4}(\hat{c}_2^{K=4} = 0.396) = 0.123$, and $t^{K=5}(\hat{c}_2^{K=5} = 0.313) = 0.162$ based on data aggregated across participants. Subplot: Observed $t(\hat{c}_2)$ values (colored points) compared to predicted $t(c_2)$ values (dark gray bars: UVG; small light gray bars: 2HTC). Parameter estimates of the UVG model ($\mu_o = 0.596$ and $\sigma_o = 1.292$) are only based on the relative frequencies of old items being ranked first, whereas parameter estimates of the 2HTC model ($\hat{D}_n^{K=3} = .067$, $\hat{D}_n^{K=4} = .067$, $\hat{D}_n^{K=5} = .061$; $\hat{D}_o^{K=3} = .240$, $\hat{D}_o^{K=4} = .223$, $\hat{D}_o^{K=5} = .222$) are based on all available data. Error bars show 95% confidence intervals. UVG = unequal-variance Gaussian; 2HTC = two-high-threshold contrast. See the online article for the color version of this figure.

demographics. In addition, we presented participants with images, whereas Malejka et al. (2022) and Kellen and Klauer (2014) used words as stimuli. While we have no tangible reason to expect our results to differ for word rankings, this generalization is ultimately an empirical question. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

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

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

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189. https://doi.org/10.1006/jmps.1999.1282

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

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

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. M. (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

Egan, J. P. (1958). *Recognition memory and the operating characteristic*. Indiana University, Hearing and Communication Laboratory. https://psycnet.apa.org/record/1960-00788-001

Fraundorf, S. H., Hourihan, K. L., Peters, R. A., & Benjamin, A. S. (2019). Aging and recognition memory: A meta-analysis. *Psychological Bulletin*, *145*(4), 339–371. https://doi.org/10.1037/bul0000185

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

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9–10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401–4410).

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, *2*(3–4), 160–165. https://doi.org/10.1007/s42113-019-00037-y

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., Meyer-Grant, C. G., Singmann, H., & Klauer, K. C. (in press). Critical testing in recognition memory: Selective influence, single-item generalization, and the high-threshold hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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. (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

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

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

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

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

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425. https://doi.org/10.1016/S1364-6613(02)01964-2

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Sarafoglou, A., Kuhlmann, B. G., Aust, F., & Haaf, J. M. (2024). Refining Bayesian hierarchical MPT modeling: Integrating prior knowledge and ordinal expectations. *Behavior Research Methods*, *56*(7), 6557–6581. https://doi.org/10.3758/s13428-024-02370-y

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

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

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics & Data Analysis*, *54*(9), 2094–2102. https://doi.org/10.1016/j.csda.2010.03.016

(*Appendices follow*)

Appendix A

Formal Model Specifications for the Ranking Paradigm

Let $R_{r,K}$ denote the probability of assigning rank r out of K to the old item. For both models it holds that the conditional probability that the old item is assigned rank two, given that it was not assigned rank one, is given by

$$c_2^K = \frac{R_{2,K}}{1 - R_{1,K}} = \frac{R_{2,K}}{\sum_{i=2}^{K} R_{i,K}}. \tag{A1}$$

According to SDT models, the items are ranked according to the strength of their associated memory signals (i.e., the item associated with the strongest signal is assigned rank one, the item with the second strongest signal is assigned rank two, and so forth). Formally, it therefore holds that

$$R_{r,K} = \binom{K-1}{r-1} \int_{-\infty}^{\infty} f_{\psi'_{\mathrm{o}}}(x) F_{\psi'_{\mathrm{n}}}(x)^{K-r} (1 - F_{\psi'_{\mathrm{n}}}(x))^{r-1} \mathrm{d}x, \tag{A2}$$

where $f_{\psi'_{\mathrm{o}}}$ is the density function of old-item memory signals and $F_{\psi'_{\mathrm{n}}}$ is the distribution function of new-item memory signals.

The 2HT model, on the other hand, assumes that if the old item is correctly detected, it is always assigned rank one. Detected new items are randomly assigned to the lowest ranks and the remaining ranks are randomly distributed among all nondetected items. Formally, it holds that

$$R_{r,K} = \begin{cases} D_{\mathrm{o}} + (1 - D_{\mathrm{o}})\xi(r, D_{\mathrm{n}}), & \text{if } r = 1 \\ (1 - D_{\mathrm{o}})\xi(r, D_{\mathrm{n}}), & \text{if } 2 \leq r \leq K \end{cases}, \tag{A3}$$

where

$$\xi(r, D_{\mathrm{n}}) = \sum_{j=r}^{K} \binom{K-1}{r-1} D_{\mathrm{n}}^{K-j} (1 - D_{\mathrm{n}})^{j-1} \frac{1}{j}, \tag{A4}$$

is the probability of assigning rank r to an old item that was not detected. It now follows from Equations A1 and A3 that

$$c_2^K = \frac{\xi(2, D_{\mathrm{n}})}{\sum_{i=2}^{K} \xi(i, D_{\mathrm{n}})}, \tag{A5}$$

under the 2HT model (see, e.g., Equation 2 in Malejka et al., 2022). Hence, c_2^K does not depend on D_{o}.

For $K = 3$, Equation A5 simplifies to

$$c_2^{K=3} = \frac{2D_{\mathrm{n}} + 1}{D_{\mathrm{n}} + 2}, \tag{A6}$$

and the corresponding inverse function is consequently given by

$$t^{K=3}(c_2^{K=3}) := \frac{1 - 2c_2^{K=3}}{c_2^{K=3} - 2}. \tag{A7}$$

Analogously, one can derive the functions

$$t^{K=4}(c_2^{K=4}) := \frac{1 - c_2^{K=4} - \sqrt{8c_2^{K=4} - 2(c_2^{K=4})^2 - 2}}{c_2^{K=4} - 3}, \tag{A8}$$

and

$$t^{K=5}(c_2^{K=5}) := \frac{A^2 - A(2c_2^{K=5} - 3) - B}{A(3c_2^{K=5} - 12)}, \tag{A9}$$

where

$$\begin{aligned} A = \sqrt[3]{5}\Big(&63(c_2^{K=5})^2 - 7(c_2^{K=5})^3 \\ &+ 3\sqrt{6}((c_2^{K=5})^6 - 18(c_2^{K=5})^5 \\ &+ 124(c_2^{K=5})^4 - 394(c_2^{K=5})^3 \\ &+ 529(c_2^{K=5})^2 - 168c_2^{K=5} + 16)^{\frac{1}{2}} - 153c_2^{K=5} + 27\Big)^{\frac{1}{3}}, \end{aligned} \tag{A10}$$

and

$$B = 5(c_2^{K=5})^2 - 30c_2^{K=5} + 15. \tag{A11}$$

(*Appendices continue*)

Appendix B

Bayesian-Hierarchical Analysis of the Critical Order Constraint

For the Bayesian-hierarchical model analysis, we adopted a latent trait approach; that is, we assumed that the probit-transformed individual D_n parameters (i.e., $d_{i,K} := \Phi^{-1}((D_\mathrm{n})_{i,K})$ for $i \in \{1, \ldots, 60\}$ and $K \in \{3{,}4{,}5\}$) are random variables drawn from a trivariate normal distribution with mean vector $\boldsymbol{\mu}_d$, representing three population means of the $d_{i,K}$ parameters for $K \in \{3{,}4{,}5\}$, and covariance matrix $\boldsymbol{\Sigma}_d = \mathrm{diag}\,(\boldsymbol{\sigma}_d)\,\boldsymbol{\Omega}_d\,\mathrm{diag}\,(\boldsymbol{\sigma}_d)$, representing parameter variability and correlations between participants. Since Bayes factors depend on the choice of prior distributions, we placed informative hyper-priors on the population-level means and variances such that the prior predictive distributions accurately reflected our assumptions on plausible parameter values (i.e., we treated the priors explicitly as part of our model; see also Sarafoglou et al., 2024). We placed Gaussian hyper-priors on the probit-transformed population means, such that $(\boldsymbol{\mu}_d)_j \sim \mathcal{N}(-1, 0.5)\forall K$, indicating a higher plausibility for smaller (i.e., <0.5) values of D_n. For the hyper-prior of the population variances, we used gamma hyper-prior distributions, such that $(\boldsymbol{\sigma}_d)_K \sim \Gamma(2, 3)\forall K$, resulting in unimodal prior distributions of the individual D_n parameters (Sarafoglou et al., 2024). The correlation matrix $\boldsymbol{\Omega}_d$ was assigned a Lewandowski–Kurowicka–Joe hyper-prior distribution with the shape parameter set to one.

We used the software Stan (Carpenter et al., 2017) to draw 400,000 samples from the posterior distribution of the unconstrained model. To evaluate the order constraint, we calculated a Bayes factor comparing the unconstrained model with an order-constrained model in line with the 2HTC model using the unconditional encompassing prior method, a generalization of the Savage–Dickey density ratio (Wetzels et al., 2010). With this method, the Bayes factor can be estimated through the proportion of posterior samples satisfying the order constraint relative to the corresponding proportion of prior samples.

Appendix C

Bayes-Factor Analyses of the Critical Order Constraint

To calculate the pertinent Bayes factor for each participant, we first note that all possible values of $c_2^{K=3}$, $c_2^{K=3}$, and $c_2^{K=3}$ form a unit cube. Furthermore, all plausible values under the 2HT model must satisfy the constraint that $c_2^K > 1/(K-1)$—the lower bound corresponding to guessing-level performance. This amounts to 1/2 × 2/3 × 3/4 = 1/4 of the unit cube's volume, over which the unconstrained model ($\mathcal{M}_0$) defines a uniform prior.

Under the constrained model ($\mathcal{M}_1$), we additionally impose the constraint that $t^{K=3}(c_2^{K=3}) \geq t^{K=4}(c_2^{K=4}) \geq t^{K=5}(c_2^{K=5})$ and—analogously to $\mathcal{M}_0$—$\mathcal{M}_1$ defines a uniform prior over all its permissible values of $c_2^{K=3}$, $c_2^{K=3}$, and $c_2^{K=3}$ within the unit cube. The corresponding volume amounts to

$$\nu := \int_{1/2}^{1}\int_{1/3}^{g_4(c_2^{K=3})}\int_{1/4}^{g_5(c_2^{K=4})} 1\mathrm{d}c_2^{K=5}\mathrm{d}c_2^{K=4}\mathrm{d}c_2^{K=3} < \frac{1}{4}, \tag{C1}$$

where

$$g_k(\cdot) := (t^{-1})^{K=k}(t^{K=k-1}(\cdot)), \tag{C2}$$

and $(t^{-1})^{K=k}$ denotes the inverse function of $t^{K=k}$ (i.e., the 2HT model predictions for $c_2^{K=k}$ as a function of D_n).

The marginal likelihood of $\mathcal{M}_0$ is therefore given by

$$\mathcal{L}_{\mathcal{M}_0} = 4\int_{1/2}^{1}\int_{1/3}^{1}\int_{1/4}^{1}\prod_{k=3}^{5} f_{\mathcal{B}}(\hat{c}_2^{K=k}n_k, n_k, c_2^{K=k})\mathrm{d}c_2^{K=5}\mathrm{d}c_2^{K=4}\mathrm{d}c_2^{K=3}, \tag{C3}$$

where

$$f_{\mathcal{B}}(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{C4}$$

(*Appendices continue*)

$$\frac{\mathcal{L}_{\mathcal{M}_1}}{\mathcal{L}_{\mathcal{M}_0}} = \frac{1}{4\nu} \frac{\int_{1/2}^{1} \int_{1/3}^{g_4\left(c_2^{K=3}\right)} \int_{1/4}^{g_5\left(c_2^{K=4}\right)} \prod_{k=3}^{5} f_{\mathcal{B}}(\hat{c}_2^{K=k} n_k, n_k, c_2^{K=k}) \mathrm{d}c_2^{K=5} \mathrm{d}c_2^{K=4} \mathrm{d}c_2^{K=3}}{\int_{1/2}^{1} \int_{1/3}^{1} \int_{1/4}^{1} \prod_{k=3}^{5} f_{\mathcal{B}}(\hat{c}_2^{K=k} n_k, n_k, c_2^{K=k}) \mathrm{d}c_2^{K=5} \mathrm{d}c_2^{K=4} \mathrm{d}c_2^{K=3}}, \tag{C6}$$

is the probability mass function of a binomial distribution and n_k is the number of observations in which the old item was not assigned rank one. The marginal likelihood of $\mathcal{M}_1$, on the other hand, is given by

$$\mathcal{L}_{\mathcal{M}_1} = \frac{1}{\nu} \int_{1/2}^{1} \int_{1/3}^{g_4\left(c_2^{K=3}\right)} \int_{1/4}^{g_5\left(c_2^{K=4}\right)} \prod_{k=3}^{5} f_{\mathcal{B}}(\hat{c}_2^{K=k} n_k, n_k, c_2^{K=k}) \mathrm{d}c_2^{K=5} \mathrm{d}c_2^{K=4} \mathrm{d}c_2^{K=3}. \tag{C5}$$

Hence,

(see Equation C6 above)

yields the targeted Bayes factor.

Received July 23, 2024
Revision received September 18, 2024
Accepted October 14, 2024 ■