Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

# Recognition-memory models and ranking tasks: The importance of auxiliary assumptions for tests of the two-high-threshold model

Simone Malejka [a,b,*], Daniel W. Heck [a,c], Edgar Erdfelder [a]

[a] University of Mannheim, Germany
[b] University College London, United Kingdom
[c] University of Marburg, Germany

ARTICLE INFO

*Keywords:*
Recognition memory
Two-high-threshold model
Multinomial processing tree models
Ranking judgments
Model specification

ABSTRACT

The question of whether recognition memory should be measured assuming continuous memory strength (signal detection theory) or discrete memory states (threshold theory) has become a prominent point of discussion. In light of limitations associated with receiver operating characteristics, comparisons of the rival models based on simple qualitative predictions derived from their core properties were proposed. In particular, *K*-alternative ranking tasks (*K*ARTs) yield a conditional probability of targets being assigned Rank 2, given that they were not assigned Rank 1, which is higher for strong than for weak targets. This finding has been argued to be incompatible with the two-high-threshold (2HT) model (Kellen & Klauer, 2014). However, we show that the incompatibility only holds under the auxiliary assumption that the probability of detecting lures is invariant under target-strength manipulations. We tested this assumption in two different ways: by developing new model versions of 2HT theory tailored to *K*ARTs and by employing novel forced-choice-then-ranking tasks. Our results show that 2HT models can explain increases in the conditional probability of targets being assigned Rank 2 with target strength. This effect is due to larger 2HT lure-detection probabilities in test displays in which lures are ranked jointly with strong (as compared to weak) targets. We conclude that lure-detection probabilities vary with target strength and recommend that 2HT models should allow for this variation. As such models are compatible with *K*ART performance, our work highlights the importance of carefully adapting measurement models to new paradigms.

Without the ability to identify previously encountered information, we would not be able to function properly. Recognition memory is a fundamental ability of our cognitive system. To investigate memory abilities and deficits in basic and applied research, mathematical measurement models disentangle effects of recognition memory from response tendencies and thereby provide more valid memory measures. Over the years, various models have been proposed, tested, revised, and rejected. The models currently best supported by data triggered a debate on the nature of recognition memory. At the heart of this discussion stands the question of whether recognition judgments stem from a direct mapping of graded familiarity signals or whether they originate from discrete memory states. The former view is represented by *signal detection theory* (SDT; Green & Swets, 1966), whereas the latter view is central to threshold theory, most prominently the *two-high-threshold* (2HT) model (Egan, 1958).

The most common approach for testing both rival models is based on a comparison of their goodness-of-fit to empirical old–new recognition data (i.e., *receiver operating characteristics* [ROC] data). Recently, however, this approach has been criticized for requiring strong and debatable auxiliary assumptions (e.g., the choice of familiarity distributions in the SDT model or the mapping of memory states onto confidence-rating responses in the 2HT model; Kellen, Winiger, Dunn, & Singmann, 2021). As an alternative, one can rely on more informative recognition paradigms that offer tests of qualitatively different predictions of the models under a minimal set of assumptions and using a simple paired comparison of observed response frequencies. In line with the idea of strong inference (Platt, 1964), these tests are typically designed as an *experimentum crucis*, in which the empirical result rules out one of the models.

One such critical test was proposed by Kellen and Klauer (2014) for the K-*alternative ranking task* (*K*ART). On each trial of this task, one previously studied item (target) and $K-1$ non-studied items (lures) are presented simultaneously, while participants are asked to rank all $K$

[*] Corresponding author at: Department of Psychology, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany.
*E-mail address:* simone.malejka@uni-koeln.de (S. Malejka).

https://doi.org/10.1016/j.jml.2022.104356
Received 5 May 2021; Received in revised form 10 July 2022; Accepted 18 July 2022
Available online 28 September 2022

items according to their belief that each item is old. Of main interest is the conditional probability of targets being assigned Rank 2 given they were not assigned Rank 1, denoted as $c_2$. More specifically, the crucial question is whether $c_2$ does or does not differ between targets with a weak memory signal versus targets with a strong memory signal. From the rival recognition models, Kellen and Klauer derived the following hypotheses: If the SDT model holds, $c_2$ will increase monotonically with target strength ($\mathcal{H}_{\text{SDT}}$: $c_2^{\text{w}} \leq c_2^{\text{s}}$); whereas if the 2HT model holds, $c_2$ must not differ between weak and strong targets ($\mathcal{H}_{\text{2HT}}$: $c_2^{\text{w}} = c_2^{\text{s}}$).[1] The results of two experiments that manipulated the frequency of target presentation (once vs. three times) were consistent with $\mathcal{H}_{\text{SDT}}$ and in conflict with $\mathcal{H}_{\text{2HT}}$.

Although we agree that extensions and generalizations of old–new recognition tasks provide excellent empirical benchmarks for testing continuous-strength versus discrete-state models, we are cautious about Kellen and Klauer's (2014) conclusions drawn from the *K*ART paradigm. The predictions of the 2HT model they proposed are based on a restrictive auxiliary assumption, namely, that the probability of lure detection is the same in the context of weak versus strong targets when target detection fails (termed lure-detection invariance assumption here). In the following, we first describe the 2HT model for the old–new recognition paradigm and summarize the formal argument regarding ranking judgments put forward by Kellen and Klauer. We then develop an alternative version of the 2HT model without the lure-detection invariance assumption (the 2HT-*K*ART model) and show that this model predicts an increase in $c_2$ with target strength. To test the critical assumption empirically, we investigate the probability of lure detection in *K*ART data. Experiment 1 is a direct replication of Kellen and Klauer's experiments. The data are analyzed with the new 2HT-*K*ART model, providing model-based estimates of lure-detection probabilities given unsuccessful target detection. Experiments 2 and 3 employ a novel forced-choice-then-ranking paradigm, in which the probability of correctly selecting a lure in a 2AFC test conditional on target non-selection in a subsequent 4ART test is analyzed as a proxy for lure detection in the absence of target detection. All three experiments suggest that lures are more easily detected in the presence of a strong than a weak target—even when this target was not detected itself. Thus, the observations of Kellen and Klauer are compatible with a discrete-state account of recognition memory as represented by 2HT theory. This demonstrates the importance of considering auxiliary assumptions when applying measurement models to new experimental paradigms.

## Discrete-state models for old–new recognition tasks

In standard recognition tests, participants have to decide for a randomized list of previously studied targets and non-studied lures whether each item is old or new. Correct decisions include hits (i.e., "old" responses to targets) and correct rejections (i.e., "new" responses to lures), whereas incorrect decisions include misses (i.e., "new" responses to targets) and false alarms (i.e., "old" responses to lures). The relative frequencies of the aforementioned events are fully described by the hit rate (denoted as $H$; the probability of responding "old" to targets) and the false-alarm rate (denoted as $FA$; the probability of responding "old" to lures). However, neither $H$ nor $FA$ represents an appropriate measure of recognition memory because both rates simultaneously increase when old-responding becomes more liberal. In order to disentangle the contributions of memory and response bias, measurement models have been applied to recognition data for more than six decades (since Egan, 1958).

A prominent and frequently used measurement model is the 2HT model, which assumes three latent states (Bröder & Schütz, 2009). With probability $D_{\text{o}}$, the target crosses the old-recognition threshold and an old-detection state is entered (see Fig. 1). Likewise, with probability $D_{\text{n}}$, the lure crosses the new-recognition threshold and a new-detection state is entered. Because the detection parameters are probabilities, item detection can vary between 0 (never detected) and 1 (always detected). However, as soon as a target or a lure is detected as old or new, respectively, the detection state always leads to the correct response. With complementary probabilities to the detection probabilities, an uncertainty state is reached and participants are assumed to guess "old" with probability $g$ dependent on their response bias, but independent of the item type.

To compare the 2HT model to other models, such as SDT, the shape of ROC curves has traditionally been used. Empirical ROCs are obtained by plotting the proportion of correctly recognized old items (hit rate) against the proportion of falsely recognized new items (false-alarm rate) across different levels of confidence or response bias. While the 2HT model is typically assumed to predict a straight line, the SDT models predicts a concave curve. However, in many cases, the shape is not as informative with respect to the underlying model as hoped for (e.g., Erdfelder & Buchner, 1998; Malmberg, 2002; Province & Rouder, 2012). Furthermore, the collection of ROC data is hampered by practical limitations (e.g., Kellen, Klauer, & Bröder, 2013; Malejka & Bröder, 2019; Van Zandt, 2000). As a consequence of the growing skepticism regarding the use of ROC analysis, various novel approaches to discriminate between continuous-strength models and discrete-state models have been proposed recently (e.g., Heck & Erdfelder, 2016; Kellen et al., 2021; Starns, 2021). In particular, tests of the models' core properties in recognition paradigms that extend beyond binary old–new judgments have gained popularity (e.g., Harlow & Donaldson, 2013; Malejka & Bröder, 2016), and Kellen and Klauer (2014) proposed to look at ranking judgments.

## Discrete-state models for *K*-alternative ranking tasks

In two experiments, Kellen and Klauer (2014) presented one target word and three lure words (4ART; Experiment 1), or one target word and two lure words (3ART; Experiment 2), and asked participants to rank the words according to their belief that each was studied previously. Of particular interest in this task is the conditional probability of the target being assigned Rank 2 given that it was not assigned Rank 1, defined as $c_2 = \frac{\pi_2}{1-\pi_1}$, where $\pi_1$ and $\pi_2$ are the unconditional probabilities of the target being assigned Rank 1 and 2, respectively. In both experiments reported by Kellen and Klauer, the observed $c_2$ estimates were higher for targets studied three times than for targets studied only once. While we agree that the SDT model predicts a monotonic increase in $c_2$ as target strength increases, we disagree on the claim that such a result is

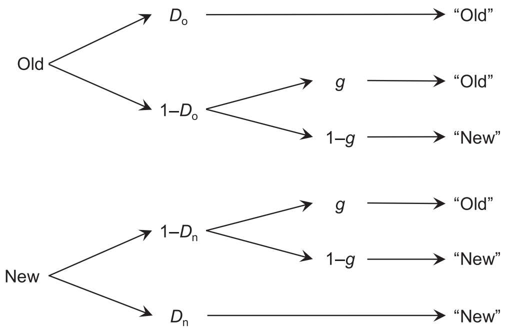

**Fig. 1.** Standard two-high-threshold model for binary old–new recognition tasks with three discrete memory states and a probabilistic decision process in case of response uncertainty. The left-hand side shows the displayed old or new item. The right-hand side shows the response given by the participant. The parameters between items and responses represent transition probabilities between discrete cognitive states.

[1] Please note that we use the term *target strength* to refer to the experimental manipulation of weak and strong targets, whereas the term *memory strength* is reserved for the continuous familiarity signal assumed by SDT.

necessarily incompatible with 2HT theory. Depending on the choice of auxiliary assumptions, the 2HT model can predict either invariance or variability of $c_2$ across different levels of target strength. As detailed in the following sections, these assumptions make up different 2HT model variants and are implemented alongside the model's core properties in order to adapt the standard 2HT model to the ranking task.

In Kellen and Klauer's (2014) adaptation of the 2HT model, the probability $\pi_i$ of a target being assigned Rank $i$ among $K$ alternatives is given by:

$$\pi_i = \begin{cases} D_o + (1 - D_o)\cdot\xi_i & \text{if } i = 1, \\ (1 - D_o)\cdot\xi_i & \text{if } 2 \le i \le K, \end{cases} \tag{1}$$

where $\xi_i$ denotes the probability of assigning Rank $i$ to the target conditional on target non-detection. The target is assigned Rank 1 either when it is detected as old with probability $D_o$, or when it is not detected as old with probability $1 - D_o$ but selected with probability $\xi_i$. Importantly, $\xi_i$ is affected by two distinct psychological processes according to 2HT theory. First, $\xi_i$ depends on the probability of detecting lures as new—a process that results in assigning the lowest ranks to detected lures (similar to assigning Rank 1 to detected targets but opposite in direction). Second, $\xi_i$ depends on guessing among items that did not reach a detection state, which are basically the non-excluded alternatives that are still competing for the highest rank (one non-detected target and up to $K - 1$ non-detected lures). Kellen and Klauer modeled both processes jointly under the assumption that they are both independent of target strength.

Based on Kellen and Klauer's (2014) adaptation to $K$ARTs, what does the 2HT model predict regarding the conditional probability $c_2$? Obviously, Equation (1) states that all *unconditional* probabilities $\pi_i$ of a target being assigned Rank $i$ depend on the target-detection probability $D_o$. In contrast, the conditional probabilities $\xi_i$ of non-detected targets being assigned Rank $i$ are assumed to be independent of $D_o$. That is, when computing the conditional probability $c_2 = \frac{\pi_2}{1-\pi_1}$, all terms that include $D_o$ cancel out, leaving $\xi_i$ as the only relevant factor:

$$c_2 = \frac{(1 - D_o)\cdot\xi_2}{1 - [D_o + (1 - D_o)\cdot\xi_1]} = \frac{(1 - D_o)\cdot\xi_2}{(1 - D_o)\cdot(1 - \xi_1)} = \frac{\xi_2}{\sum_{i=2}^{K}\xi_i}. \tag{2}$$

Because $\xi_i$ is assumed to be independent of $D_o$, it follows that $c_2$ must be independent of target strength. In other words, different values of $D_o$ will not affect the predicted value of $c_2$ in Equation (2). Hence, Kellen and Klauer's 2HT model variant predicts that the $c_2$ values must be equal for weak and strong targets ($\mathcal{H}_{2HT}$: $c_2^w = c_2^s$)—a prediction that is at odds with the data of their Experiments 1 and 2.

Based on the set of assumptions that entered into their derivations, Kellen and Klauer's (2014) conclusions regarding the 2HT model are fully justified. However, a closer look reveals that Kellen and Klauer's adaptation of the 2HT model to $K$ARTs includes two new—seemingly small but important—auxiliary assumptions. These two assumptions, which we termed target-detection dominance assumption and lure-detection invariance assumption, require closer inspection.

Regarding the *target-detection dominance assumption*, Kellen and Klauer (2014) maintained the idea that the optimal ranking decision would be made whenever a target is in the detection state (i.e., detected targets are always assigned Rank 1 and lures are assigned Ranks 2 to $K$). In other words, detected targets dominate the ranking of other items. This assumption may be considered problematic in confidence-rating tasks, in which confidence ratings for detected targets do not necessarily dominate those of lures (Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Malmberg, 2002). However, because ranking judgments do not require explicit or implicit confidence assessments, we agree with Kellen and Klauer that ranking tasks lack requirements that "would lead one to assigning anything else than Rank 1 to detected old items" (p. 1802). We thus follow their claim and base our analyses on the target-detection dominance assumption.

The second auxiliary assumption of Kellen and Klauer (2014), however, is more problematic. According to the *lure-detection invariance assumption*, lures are detected as new with the same probability in the context of test displays with a weak versus a strong non-detected target. In other words, the probability of lure detection, $D_n$, in the 2HT framework is assumed to be invariant under different target strengths. This assumption, which seems very plausible at first glance, let Kellen and Klauer to jointly model guessing and lure detection by parameter $\xi_i$ for all target-strength conditions and without further specifying the exact predictions conditional on target non-detection. As we outline next, this assumption may be overly restrictive. Importantly, the rejection of the 2HT model for the $K$ART hinges on it: The crucial prediction that $c_2$ should be independent of target strength does not hold anymore when $D_n$ increases monotonically with the strength of the target in an item tuple.

## Does the conditional Rank-2 probability increase with the probability of lure detection?

In the following, we show that $\mathcal{H}_{2HT}$: $c_2^w \le c_2^s$ follows from an adaptation of 2HT theory to ranking tasks based on the assumptions that (1) $c_2$ increases with $D_n$ and (2) $D_n$ increases with target strength. The former assumption is derived from a formal model analysis, whereas the latter requires an empirical test. For both purposes, we propose a new multinomial processing-tree (MPT) model based on 2HT theory that is tailored to the $K$ART paradigm: the 2HT-$K$ART model (for general overviews of MPT models, see Batchelder & Riefer, 1999; Erdfelder et al., 2009).

Fig. 2 illustrates the special case of the 2HT-$K$ART model in which participants have to rank a tuple of $K = 4$ items. Similar to the standard 2HT model for binary old–new recognition, the adapted model assumes that targets are detected with probability $D_o$ and lures are detected with probability $D_n$, where the latter is conditional on target non-detection.[2] Furthermore, it is assumed that each lure is detected independently from other lures, and that detected targets and lures are always assigned the first and last ranks, respectively (thus relying on the target-detection dominance assumption). Finally, the model assumes that participants assign ranks to the remaining items in the uncertainty state by pure guessing. For instance, when only two lures are detected (see the three branches with two $D_n$ and one $1 - D_n$ in Fig. 2A), they are assigned Ranks 3 and 4, whereas the non-detected target and the non-detected lure receive Ranks 1 and 2 by guessing (i.e., with a probability of $\frac{1}{2}$ for each of the two possible combinations). The proposed model for the $K$ART paradigm builds on similar psychological assumptions as Province and Rouder's (2012) 2HT model for two-alternative forced-choice tasks, Luce's (1963) low-threshold (LT) model for two-alternative forced-choice tasks, and Kellen, Erdfelder, Malmberg, Dubé, and Criss's (2016) extension of the LT model to ranking tasks (see also Iverson & Bamber, 1997).

Following the notation of Kellen and Klauer (2014), the 2HT-4ART model can be generalized to any number of items $K$ in order to model the probability $\pi_i$ of a target being assigned Rank $i$ among $K$ alternatives as a function of $D_o$ and $\xi_i(D_n)$. However, in contrast to Kellen and Klauer, we explicitly account for lure detection on the $K$ART trials. In particular, we model the probability $\xi_i(D_n)$ of assigning Rank $i$ to a non-detected target as a function of $D_n$:

$$\xi_i(D_n) = \sum_{j=1}^{K} \binom{K-1}{j-1} \cdot D_n^{K-j} \cdot (1 - D_n)^{j-1} \cdot \zeta_{ij}, \tag{3}$$

where the response-mapping function $\zeta_{ij}$ determines the probability of

[2] We restrict the 2HT-$K$ART model to $D_n \in [0, 1)$ because $c_2$ is not defined for $D_n = 1$.

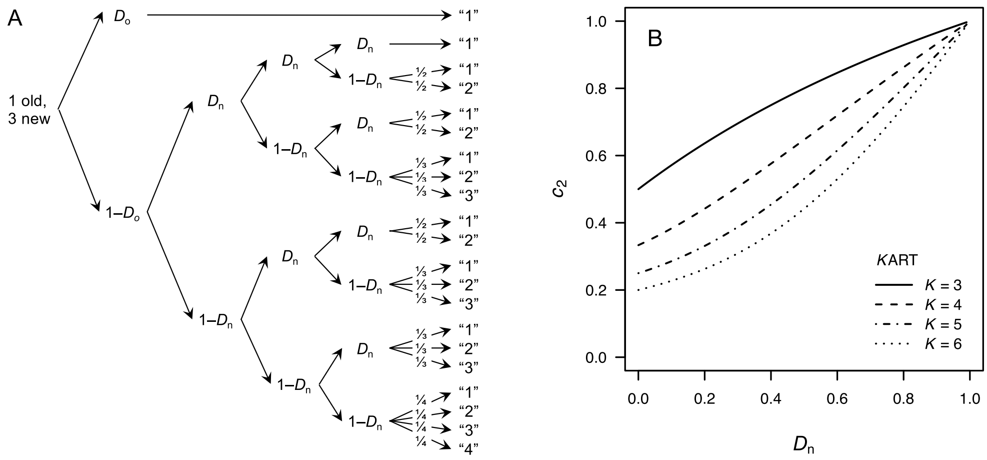

**Fig. 2.** (A) New two-high-threshold model for ranking tasks. The item quadruple on the left-hand side consists of one old item and three new items. The probabilities of target and lure detection (the latter conditional on target non-detection) are denoted as $D_o$ and $D_n$, respectively, with $D_n \in [0,1)$. Guessing probabilities are given by $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$, and depend on the total number of non-detected items. The numbers in quotation marks on the right-hand side represent the four possible ranks that can be assigned to the target depending on the memory-state combination and possible guessing outcome for all four test items. (B) Conditional probabilities $c_2$ of targets in a *K*ART test being assigned Rank 2 given that they were not assigned Rank 1 as a function of the probability of lure detection $D_n$ for different numbers of items *K* per tuple. Note that $c_2$ is only defined for $D_n < 1$.

assigning Rank $i$ to the target conditional on $j$ items being in the uncertainty state (i.e., the target and $j-1$ lures), while $K-j$ lures were already excluded as potential targets. Note that $\zeta_{ij}$ represents memory-unrelated guessing probabilities (i.e., this term neither depends on target detection nor on lure detection) and is fixed to $\frac{1}{j}$ in the 2HT-4ART model shown in Fig. 2A (to reflect equiprobable guessing among non-detected items).

The proposed 2HT-*K*ART model predicts that the conditional probability of assigning Rank 2 to non-detected targets increases when the number of detected lures increases. This effect essentially emerges because the detected lures are excluded from the set of alternatives competing for high ranks. Hence, the conditional probability $c_2$ of assigning Rank 2 to targets increases as the lure-detection probability $D_n$ increases. When assuming equiprobable guessing among non-excluded items, $\zeta_{ij} = \frac{1}{j}$, this prediction follows when inserting Equation (3) into Equation (2):

$$c_2(D_n) = \frac{\xi_2(D_n)}{\sum_{i=2}^{K}\xi_i(D_n)} = \frac{\sum_{j=2}^{K}\binom{K-1}{j-1}\cdot D_n^{K-j}\cdot(1-D_n)^{j-1}\cdot\frac{1}{j}}{\sum_{i=2}^{K}\sum_{j=i}^{K}\binom{K-1}{j-1}\cdot D_n^{K-j}\cdot(1-D_n)^{j-1}\cdot\frac{1}{j}}. \qquad (4)$$

As before, the target-detection probability $D_o$ cancels out. However, the conditional probability $c_2$ still depends on the lure-detection probability $D_n$. As shown in Fig. 2B, $c_2$ increases monotonically as a function of $D_n$ irrespective of the number of alternatives $K$.

In sum, Kellen and Klauer's (2014) 2HT model variant is based on the idea that lure-detection probabilities are independent of target strength, whereas our model variant allows for different lure-detection probabilities that are dependent on target strength—$D_n^s$ for item tuples with a strong target and $D_n^w$ for item tuples with a weak target. Of course, other adaptations of 2HT theory to ranking tasks are possible. For example, it is possible to relax the assumption of equiprobable guessing by assuming that higher levels of target strength would bias participants to guess a target rank of 1 more often. However, this version would violate the 2HT theory's assumption of complete information loss, which is a core assumption of 2HT theory (for a detailed account of complete information loss, see the General Discussion).

Although Kellen and Klauer's (2014) model is more parsimonious than ours in describing behavior in *K*ARTs (e.g., it uses fewer parameters), our model was designed as a means to test the lure-detection invariance assumption that inspired Kellen and Klauer's interpretation of 2HT theory. As such, our model is necessarily more complex. Furthermore, the process theory underlying our model can be read as discretizing a continuous memory signal at the level of the item tuples (see the Appendix). While this may call for a direct comparison of the models controlling for their different complexities, we followed Kellen and Klauer's call for simple tests of competing hypotheses using off-the-shelf statistical techniques and tested the assumption in three qualitative tests as described next.

## Does the probability of lure detection increase with target strength?

Given that $c_2$ increases monotonically with the lure-detection probability $D_n$ according to our 2HT-*K*ART model, it is necessary to test empirically whether $D_n$ actually increases with target strength. First support for this hypothesis comes from recognition-memory studies that rely on fitting the 2HT model to data: $D_o$ and $D_n$ are often constrained to be equal without compromising the model's goodness-of-fit (e.g., Bayen, Murnane, & Erdfelder, 1996; Klauer & Wegener, 1998; Snodgrass & Corwin, 1988). Such findings suggest that manipulations of item detection typically affect $D_o$ and $D_n$ simultaneously, such that any change in $D_o$ is accompanied by a corresponding change in $D_n$, and vice versa.

One might argue that findings on the link between $D_o$ and $D_n$ are not relevant to Kellen and Klauer's (2014) paradigm because they manipulated target strength within the study list and selected lures from a homogeneous pool of new items. In this experimental setup, it may seem unproblematic to assume that $D_n$ must be equal in the context of weak versus strong targets. However, we argue that lure detection in *K*ART tests differs from lure detection in old–new recognition tests: *K*ART tests do not present items individually (i.e., $K = 1$ per trial; single-item recognition); instead, they present multiple items at the same time (i.e., $K \geq 2$ per trial; multiple-item recognition). Hence, the test stimulus is not a *single* item, but a *tuple* of items (i.e., one target and $K-1$ lures). This stands in stark contrast to old–new judgments in which test stimulus and test item are identical (i.e., one target or one lure). Furthermore, we propose that lure detection in *K*ART tests may depend on the context provided by all items that are presented in a given test trial and not only on the currently attended item.

This distinction is supported by a recent finding of Voormann, Spektor, and Klauer (2021), who showed that multiple-item recognition decisions cannot be treated as a sequence of independent single-item recognition decisions. Their participants showed higher hit rates in single-item tests than in multiple-item tests, and data analysis revealed that recognition decisions for words in a word pair were interdependent. Likewise, research on eyewitness identification has shown that simultaneous lineups require comparative judgments, whereas sequential lineups do not (Lindsay & Wells, 1985), and that the eyewitness's knowledge that no more than one person in the lineup could be the perpetrator affects their decision-making (Wixted & Mickes, 2014).

For these reasons, we propose that, in order to arrive at a ranking decision in the *K*ART, all items in the test display are taken into account jointly and evaluated in direct comparison. Successful discrimination then depends on all items in the display, and it becomes implausible to assume that detection of a lure in a weak-target tuple (with probability $D_n^w$) is as likely as detection of a lure in a strong-target tuple (with probability $D_n^s$). Hence, we hypothesize that lure detection is facilitated when the lures are presented in the context of a strong target as compared to a weak target, which results in the prediction that $D_n^w \leq D_n^s$ even when the target has not been detected itself. For a more detailed formal account of familiarity-contrast mechanisms assumed to underlie target and lure detection in *K*ART displays, we refer to the Appendix.

The broader idea that items influence each other's recognition rates is supported by research on ensemble recognition. In this task, participants make old–new recognition judgments for item tuples consisting only of targets or only of lures, thereby including all items of the tuple into their recognition decision (Benjamin, Diaz, & Wee, 2009). In a recent investigation of the paradigm, Dubé, Tong, Westfall, and Bauer (2019) found support for aggregation of statistical representations in recognition memory. However, the precise processes involved in producing the average estimates remain unclear. Outside of the recognition-memory literature, vision research has long been investigating ensemble recognition (e.g., Dubé, 2019). Although readers may argue that visual perception and short-term memory cannot illustrate the workings of long-term recognition memory, the structural similarity between recognition-ranking tasks and ensemble recognition tasks stimulates the question whether there is also a conceptual similarity of the involved processes. Of course, models and theories need to be validated anew for a structurally similar task taken from a different domain. Therefore, we believe that it is important to test whether and how the probability of detecting an item—as an essential parameter of any 2HT model—depends on the other items in a tuple.

The question whether $D_n$ increases monotonically with the strength of the target in the *K*ART display is ultimately an empirical one. To answer it, we present two approaches to test the lure-detection invariance assumption: a model-based replication approach using Kellen and Klauer's (2014) 4ART paradigm analyzed with the 2HT-4ART model (Experiment 1) and an approach based on a new forced-choice-then-ranking paradigm (Experiments 2 & 3). In the former approach, the measure of interest is the probability of lure detection conditional on target non-detection obtained from fitting the proposed 2HT-4ART model in Fig. 2A to 4ART data. In the latter approach, the measure of interest is the probability of correctly selecting a lure in the presence of a non-detected target. The measure is obtained in two novel forced-choice-then-ranking tasks, which allow analyzing lure-selection rates in a 2AFC test conditional on target non-selection in a 4ART test. The results of both approaches suggest that the probability of lure detection is higher in the context of strong than weak targets, thereby casting doubt on the lure-detection invariance assumption.

## Experiment 1

Experiment 1 was a direct replication of Kellen and Klauer's (2014) experiments to test the crucial effect of $c_2^w < c_2^s$ in a 4ART paradigm. The sample size, the repetition scheme of words in the study phase, and the number of lures in the tuples presented at test followed their Experiment 1. The reward scheme during the test was adopted from their Experiment 2, such that participants received points contingent on the rank assigned to the target.

To test whether the strength of targets in item tuples influences the probability of lure detection, we first looked at the individual goodness-of-fit scores of the proposed 2HT-4ART model assuming pure, equiprobable guessing as described above. In this model, the lure-detection probability $D_n$ is conditional on target non-detection because $D_n$ only affects the predicted rank frequencies for the target when target detection fails (cf. Fig. 2A). This allows to test the null hypothesis that $D_n$ is equal for weak and strong targets ($\mathcal{H}_0$: $D_n^w = D_n^s$) against the alternative hypothesis that $D_n$ is higher for strong than for weak targets ($\mathcal{H}_1$: $D_n^w < D_n^s$). If the person-specific estimates of $D_n$ are larger for test displays with a strong target than for test displays with a weak target, it follows that the probability of lure detection must be context-dependent.

### Data availability

All raw data, analysis code, and multinomial processing-tree models used in this article are publicly available through the Open Science Framework at https://osf.io/yx39t/. All computations were performed with R (R Core Team, 2019) in combination with JAGS (Plummer, 2017).

### Method

#### Participants

Twenty-two undergraduate students (19 females, 3 males) from the University of Mannheim participated in exchange for €3.50 or partial course credit with a maximum performance-based reward of €3.00, respectively. The sample size was selected a priori to match Kellen and Klauer's (2014) Experiment 1. A sensitivity analysis using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) revealed that this sample size suffices to detect an above-medium effect of size $d_z = 0.56$ in the test of interest (i.e., a directional Wilcoxon signed-rank test assuming an underlying normal distribution, significance level $\alpha = .05$, and power $1 - \beta = .80$).[3] All participants were native or fluent speakers of German, and their mean age was 22.14 years ($SD = 3.83$, range = 18–33). The study took place in a computer laboratory with individual cubicles, and participants were tested in small groups of up to five people.

#### Materials

Following the selection criteria provided by Kellen and Klauer (2014), German nouns were taken from Lahl, Göritz, Pietrowsky, and Rosenberg (2009). Word lengths ranged from four to eight letters and all words were of medium valence (3.5 to 6.5 on an 11-point scale) and low arousal (0.5 to 4.5 on an 11-point scale). For each participant, 150 and 450 words were randomly selected from the word pool to serve as old and new items, respectively.

#### Design and procedure

The experiment comprised a study phase and a test phase. During the study phase, target strength was manipulated as a within-subjects factor, such that 75 of the targets were presented once (weak targets) and 75 were presented three times (strong targets). As in Kellen and Klauer's (2014) Experiment 1, each of three blocks in the study list contained all strong words interspersed with one third of the weak words. The presentation order within each block was randomized, but it was ensured that at least three other words were presented between the presentation of a strong word and its next presentation. Each word was presented individually for 600 ms at the center of the computer screen in black

[3] Although the Wilcoxon signed-rank test is distribution-free, calculating numerical values in a power analysis requires the user to specify a response distribution because the effect of a deviation from symmetry depends on the specific form of this distribution (for details, see G*Power, 2020).

**Table 1**
Mean Probability Estimates (and Standard Deviations) in the Ranking Task of Kellen and Klauer (2014, Experiment 1) and of Experiments 1, 2, and 3 Reported Here.

| Experiment | Target | $\pi_1$ | $\pi_2$ | $\pi_3$ | $\pi_4$ | $c_2$ |
|---|---|---|---|---|---|---|
| Kellen & Klauer | Weak | .38 (.08) | .23 (.05) | .21 (.05) | .18 (.05) | .37 (.09) |
| | Strong | .55 (.15) | .19 (.07) | .14 (.06) | .13 (.05) | .43 (.09) |
| Experiment 1 | Weak | .40 (.11) | .21 (.04) | .21 (.08) | .18 (.05) | .37 (.09) |
| | Strong | .57 (.15) | .19 (.05) | .13 (.06) | .11 (.06) | .46 (.09) |
| Experiment 2 | Weak | .54 (.13) | .20 (.06) | .15 (.06) | .12 (.06) | .44 (.11) |
| | Strong | .64 (.16) | .16 (.06) | .12 (.07) | .08 (.06) | .47 (.13) |
| Experiment 3 | Weak | .45 (.10) | .20 (.04) | .16 (.05) | .18 (.07) | .38 (.09) |
| | Strong | .63 (.14) | .15 (.05) | .11 (.05) | .11 (.06) | .43 (.10) |
| Experiment 3* | Weak | .50 (.14) | .18 (.08) | .13 (.07) | .19 (.11) | .38 (.20) |
| | Strong | .68 (.15) | .13 (.07) | .08 (.06) | .11 (.09) | .44 (.22) |

*Note.* $\pi_i$ = estimated probability of the target being assigned Rank *i*; $c_2$ = estimated conditional probability of the target being assigned Rank 2 given that it was not assigned Rank 1; Experiment 3 = results based on all 4ART displays; Experiment 3* = results based on 4ART displays that followed a 2AFC display with a target–lure pair.

Arial font on gray background with all letters capitalized. Before the next word appeared, a blank screen was shown for 100 ms.

A practice phase for the 4ART test immediately followed the study phase. The first practice display explained the ranking of the four words and the second practice display exemplified the reward scheme. The 4ART test consisted of 150 displays (with a total of 150 targets and 450 lures). All test displays showed one target and three randomly selected lures in one rectangular box each. The four boxes were arranged in a 2 × 2 matrix around the center of the screen, and the positions of the target and the three lures were randomly determined. In line with Kellen and Klauer's (2014), ranks could be assigned to each word by clicking in the boxes with the computer mouse. The first word selected received Rank 1 and the rank number appeared above the word's box, the second item selected received Rank 2, and so forth. At any stage of the ranking, participants could erase all assigned ranks by clicking "DELETE." When all items were ranked, participants could confirm the given response and proceed to the next item quadruple by clicking "OK." The reward scheme allowed participants to win one point for each assignment of the target to Rank 1 and lose two, three, or four points for assignments of the target to Rank 2, 3, or 4, respectively. Participants were told that no feedback would be provided and that they could not lose money from the fixed show-up fee of €3.50, but that they could win up to €3.00 when cashing in the maximum number of 150 points. After the test, participants were thanked, debriefed, and compensated for participation.

### *Results*

#### *Analysis of probability estimates*

Table 1 shows the mean estimated probabilities $\pi_i$ of the target being assigned Rank *i* among *K* alternatives and the mean conditional probabilities $c_2$ of the target being assigned Rank 2 given that it was not assigned Rank 1. These mean probability estimates closely replicate the ones reported by Kellen and Klauer (2014) for their Experiment 1. To avoid strong, and perhaps misspecified, distributional assumptions when testing statistical hypotheses, we followed Kellen and Klauer by conducting non-parametric tests across participants. A directional Wilcoxon signed-rank test showed that the difference in accuracy between weak and strong targets as measured by individual $\pi_1^{\mathrm{w}}$ and $\pi_1^{\mathrm{s}}$ parameters was significant, $V = 249$, $p < .001$, $d_z = 1.78$. This shows that the study-repetition manipulation successfully led to better memory for strong than for weak targets. Fig. 3A plots the individual $c_2$ estimates for strong targets against the individual $c_2$ estimates for weak targets. Most observations lie above the main diagonal, which suggests that strong targets are more often assigned Rank 2 than weak targets when both were not assigned Rank 1. This impression was supported by a significant directional Wilcoxon signed-rank test examining whether the mean $c_2^{\mathrm{s}}$ value was higher than the mean $c_2^{\mathrm{w}}$ value, $V = 222$, $p = .001$, $d_z = 0.90$.

Sharing Kellen and Klauer's (2014) concern that the Wilcoxon test is based on a different number of observations for each individual and each target-strength condition, which results in different standard errors of the estimates for $c_2$, we used the hierarchical Bayesian model reported in Kellen and Klauer's Supplemental Materials to compute a Bayes factor (BF). The BF quantifies the evidence for $\mathcal{H}_1$: $\delta > 0$ relative to $\mathcal{H}_0$: $\delta = 0$, where $\delta$ is the effect size capturing the mean difference between $c_2^{\mathrm{w}}$ and $c_2^{\mathrm{s}}$ on the group level. Because we were interested in an order-restricted test of the $c_2$ estimates (i.e., a directional test), we used a normal prior for $\delta$ that was truncated from below at zero. The estimated BF indicated that the data were more than 33 times more likely to have occurred under $\mathcal{H}_1$ than under $\mathcal{H}_0$. This can be considered as strong evidence in favor of $\mathcal{H}_1$ according to Jeffreys (1961).

#### *Model-based analysis*

The 2HT model for the 4ART with equiprobable guessing among non-excluded alternatives competing for Rank 1 in Fig. 2A was fitted to the data of each individual separately. In total, the responses of 20 of the 22 participants were described well by the model, all $G^2(2) \leq 5.47$, $p \geq .065$, whereas the responses of two participants were not described well, both $G^2(2) \geq 8.75$, $p \leq .013$. The mean maximum-likelihood parameter estimates (and the corresponding standard errors of the mean) for item detection were .19 (.03) for $D_{\mathrm{o}}^{\mathrm{w}}$, .39 (.04) for $D_{\mathrm{o}}^{\mathrm{s}}$, .06 (.02) for $D_{\mathrm{n}}^{\mathrm{w}}$, and .15 (.03) for $D_{\mathrm{n}}^{\mathrm{s}}$.[4]

Fig. 4A plots the individual $D_{\mathrm{o}}$ estimates for strong targets against the individual $D_{\mathrm{o}}$ estimates for weak targets. This allows checking whether more repetitions of targets during the study phase resulted in a higher probability of target detection in the 2HT model during testing. As expected, all estimates—except two—lie above the main diagonal. A directional Wilcoxon signed-rank test confirmed that the $D_{\mathrm{o}}^{\mathrm{s}}$ estimates were on average significantly larger than the $D_{\mathrm{o}}^{\mathrm{w}}$ estimates, $V = 248$, $p < .001$, $d_z = 1.62$, supporting the impression that strong targets had indeed a higher probability of being detected as old.

Regarding the test of the crucial lure-detection invariance assumption, Fig. 4B plots the individual $D_{\mathrm{n}}$ estimates of the lure-detection parameter for contexts with strong targets against contexts with weak targets. A majority of 14 participants had higher $D_{\mathrm{n}}^{\mathrm{s}}$ estimates than $D_{\mathrm{n}}^{\mathrm{w}}$ estimates. As predicted, a directional Wilcoxon signed-rank test

[4] As a robustness analysis, we also fitted a hierarchical version of the 2HT model for the 4ART (Klauer, 2010) using the R package TreeBUGS (Heck, Arnold, & Arnold, 2018). The hierarchical group-level estimates and credibility intervals did not differ substantially from the posterior means and standard deviations aggregated across the individual parameter estimates reported in the main text. The model fit to the observed response frequencies and correlations was slightly below the conventional threshold of .05 with posterior predictive *p*-values of .045 and .039 for test quantities $T_1$ (assessing whether the model can recover the observed mean category frequencies) and $T_2$ (assessing whether the model can recover the observed covariance structure), respectively. Note, however, that the conventional threshold cannot be applied with the same meaning as in a $G^2$-test because posterior predictive *p*-values are not uniformly distributed under $\mathcal{H}_0$.

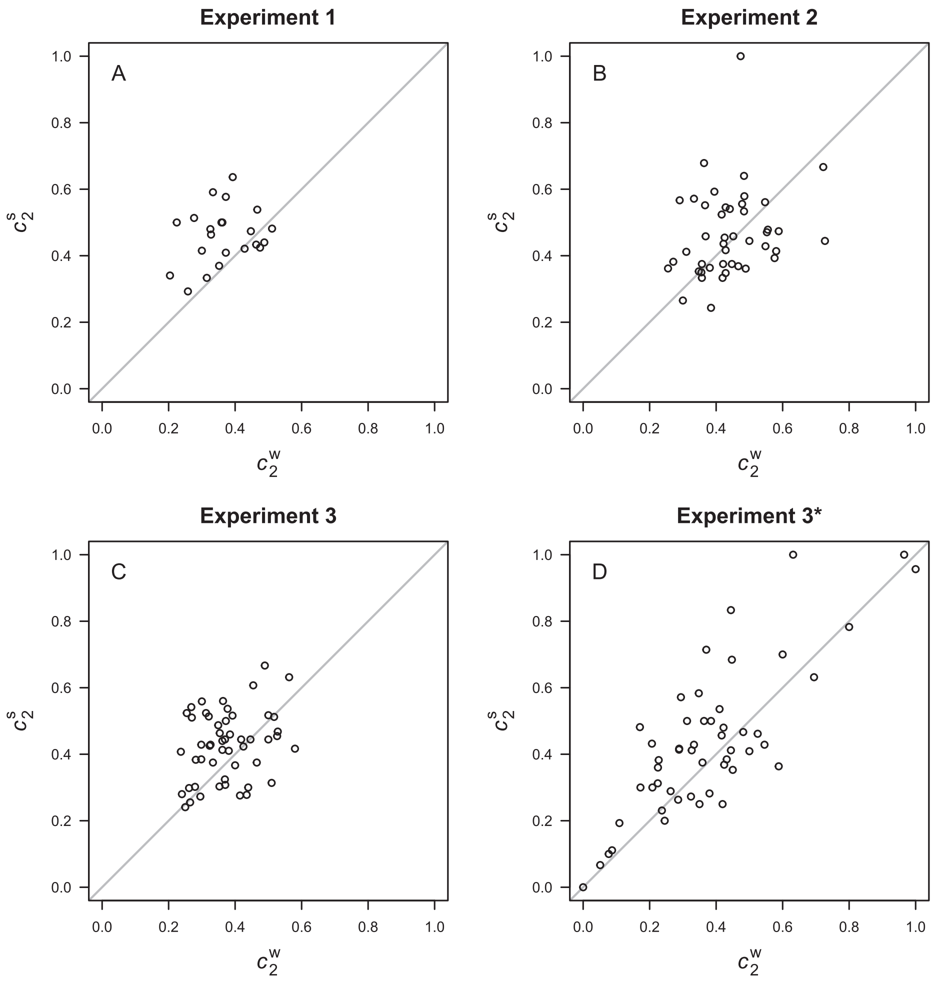

**Fig. 3.** Individual probability estimates of assigning Rank 2 to weak versus strong targets given they were not assigned rank 1 in the 4ART tests of Experiments 1, 2, and 3. $c_2$ = estimated conditional probability of the target being assigned Rank 2 given that it was not assigned Rank 1. (A) Experiment 1, (B) Experiment 2, (C) Experiment 3 with results based on all 4ART displays, (D) Experiment 3* with results based on 4ART displays that followed a 2AFC display with a target–lure pair.

confirmed that the probabilities of lure detection were on average significantly higher in the context of a strong target than in the context of a weak target, $V = 174$, $p = .005$, $d_z = 0.67$.

### Discussion

Experiment 1 replicated Kellen and Klauer's (2014) finding that the conditional probability $c_2$ is higher for strong than for weak targets. This finding is inconsistent with Kellen and Klauer's assumption that the probability of lure detection in 2HT theory is unaffected by the strength of the non-detected target in the test display, and as such inconsistent with their 2HT model. In contrast, the finding is consistent with our 2HT-4ART model that dispenses with the lure-detection invariance assumption and allows $D_n$ to differ between displays with a weak and a strong non-detected target. This supports our argument that the lure-detection invariance assumption in Kellen and Klauer's model is overly restrictive and at odds with empirical data.

Although Experiment 1 clearly shows that the 2HT-4ART model is consistent with ranking data, one could criticize that this 2HT model version (a) is more flexible than the one considered by Kellen and Klauer (2014) and (b) has not been directly validated so far (especially with respect to the crucial model parameter $D_n$). Even though the model-fitting results are plausible and in line with previous results on the 2HT model, we deemed it necessary to provide additional evidence for our argument. Therefore, we developed a novel paradigm to test the crucial prediction that the probability of lure detection in the 2HT model increases with the strength of the target in the same test display—even when this target is not assigned Rank 1 in a ranking task, implying that it was not detected.

## A novel forced-choice-then-ranking paradigm

Experiments 2 and 3 implemented a novel paradigm including two successive recognition tasks based on the same targets, which we called the forced-choice-then-ranking (2AFC-*K*ART) paradigm. We developed this paradigm to test the lure-detection invariance assumption in a more direct way and thereby provide converging evidence on the $D_n$ results obtained with the 2HT-*K*ART model in Experiment 1. After the study phase, in which weak and strong targets were presented once and three times, respectively, participants completed (1) a two-alternative forced-choice (2AFC) test in which they had to choose the word in a test pair that was more likely to be new and (2) a 4ART test identical to the one already used in Experiment 1. The 4ART test either followed as a separate block after the end of the 2AFC block (Experiment 2) or every 4ART test display immediately extended the corresponding 2AFC display from two to four items (Experiment 3). Note that the 2AFC test asked for the new rather than the old item for two reasons. First, this task mirrors our interest in lure detection. Second, it has the practical advantage that an arbitrary number of 2AFC displays with a lure–lure pair can be added as fillers to the 2AFC displays with a target–lure pair of main interest. Such fillers are necessary to ensure that participants cannot restudy the targets in the 2AFC test prior to the 4ART test

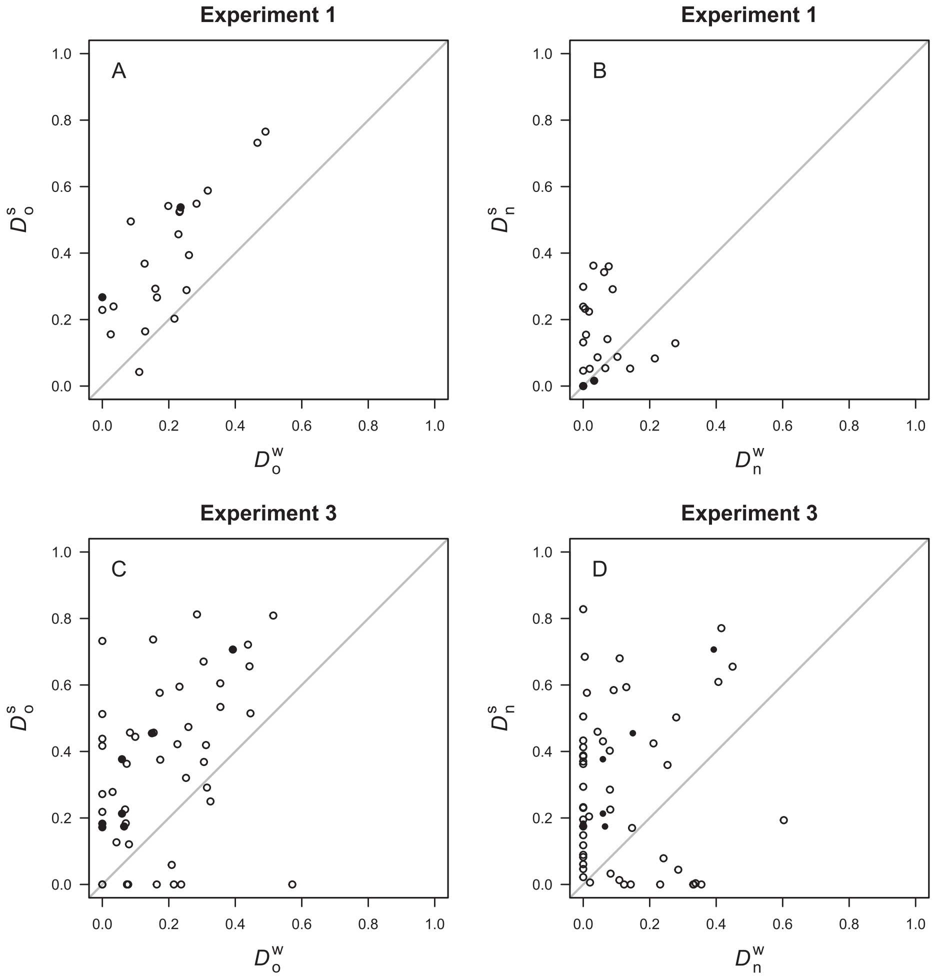

**Fig. 4.** Individual probability estimates of target and lure detection for test displays with a weak versus strong target obtained from fitting the 2HT models in Experiments 1 and 3. $D_o$ = estimated probability of target detection, $D_n$ = estimated probability of lure detection conditional on target non-detection. (A–B) Experiment 1 based on the 2HT-4ART model, (C–D) Experiment 3 based on the 2HT-2AFC-4ART model. Full points highlight participants for which the models showed a misfit as indicated by $p < .05$ in the $G^2$-test.

(Experiment 2) nor will they always need to assign Rank 1 to one word of the 2AFC pair (Experiment 3). Because there are only two possible states of the world (old vs. new word), selecting the lure should not pose a greater challenge than selecting the target to participants. However, the switch from selecting the lure in the 2AFC test and selecting the target in the subsequent 4ART test may require more attention in Experiment 3, in which both questions followed on a trial-by-trial basis. Therefore, when analyzing the data, one must pay attention to the response strategies used to solve the consecutive tasks.

As will become clear in the following, by considering only those 2AFC pairs with a target that cannot have been detected on the subsequent 4ART test according to 2HT theory (i.e., received a rank larger than 1), we can directly test whether the strength of a non-detected target affects the probability of lure detection in the 2AFC test. For this purpose, we used the lure-selection rate (i.e., the probability of correctly selecting the lure in a target–lure pair; denoted as *LS*) in the 2AFC test as our core dependent measure. According to 2HT theory, this measure depends on the probabilities of detecting the target ($D_o$), detecting the lure given target non-detection ($D_n$), and lucky guessing ($g$) in the 2AFC context. However, there is only one possible combination of events that results in lure non-selection, namely when neither the target nor the lure is detected and the participant guesses incorrectly with probability of .50. Hence, the probability of lure non-selection (1 − *LS*) is given by the product $(1 - D_o) \cdot (1 - D_n) \cdot .50$, which immediately implies:

$$LS = 1 - (1 - D_o) \cdot (1 - D_n) \cdot .50. \tag{5}$$

Now let us consider the conditional *LS* probabilities given critical test pairs in which the target is known to be detected ($LS_+$) versus known to be non-detected ($LS_-$). By inserting $D_o = 1$ and $D_o = 0$, respectively, into Equation (5), we obtain:

$$LS_+ = 1, \tag{6}$$

$$LS_- = 1 - (1 - D_n) \cdot .50. \tag{7}$$

Thus, the conditional $LS_-$ measure depends only on the core parameter of interest, that is, the probability of lure detection given target non-detection $D_n$.[5] In fact, by solving Equation (7) for $D_n$, we can write $D_n$ as a function of $LS_-$:

$$D_n = 2 \cdot LS_- - 1. \tag{8}$$

How can *LS*, $LS_+$, and $LS_-$ be estimated empirically? By taking advantage of the forced-choice-then-ranking paradigm, we can estimate the *LS* parameters not only based on all target–lure pairs in the 2AFC

[5] Note that this also holds when target and lure detection are not stochastically independent as assumed in our derivation. Since $D_n$ is the conditional probability of lure detection given failure of target detection (1 −$D_o$), Equation (7) becomes generally valid, irrespective of the independence assumption.

test, but also separately for two different subsets of pairs depending on the outcome of the subsequent 4ART test: First, for the subset of pairs with the target assigned Rank 1 in the 4ART test ($LS_1$), and second, for the subset of pairs with the target being assigned Ranks 2 to 4 ($LS_{>1}$).

Of course, Rank-1 judgments for targets do not necessarily imply that the target was detected. Rank 1 could also have been selected due to lucky guessing among the items in the uncertainty state. Thus, the empirical $LS_1$ estimate cannot be expected to approach the $LS_+$ parameter of 1, which was derived for the subset of detected targets (cf. Equation (6)). However, more important for our purposes is the fact that targets with ranks larger than 1 in the 4ART test must necessarily be non-detected according to Kellen and Klauer's (2014) target-detection dominance assumption. Thus, it follows that the $LS_{>1}$ estimate is based only on those target–lure pairs in the 2AFC test in which the target was not detected in the subsequent 4ART test. If we assume that non-detected targets in the 4ART test were also not detected in the preceding 2AFC test, $LS_{>1}$ may serve as an empirical estimate of parameter $LS_-$, that is, the lure-selection rate conditional on target non-detection (cf. Equation (7)).

Based on this assumption, separate $LS_{>1}$ estimates for 2AFC displays with weak targets ($LS^{\mathrm{w}}_{>1}$) and strong targets ($LS^{\mathrm{s}}_{>1}$) allow for a test of the lure-detection invariance assumption. According to this assumption, the lure-selection rates conditional on non-detection of weak and strong targets must be equal ($\mathcal{H}_0$: $LS^{\mathrm{w}}_{>1} = LS^{\mathrm{s}}_{>1}$). In contrast, if lure detection depends on the strength of the target in the test display, conditional lure-selection rates must be larger for strong targets compared to weak targets ($\mathcal{H}_1$: $LS^{\mathrm{w}}_{>1} < LS^{\mathrm{s}}_{>1}$). Moreover, by inserting $LS^{\mathrm{w}}_{>1}$ and $LS^{\mathrm{s}}_{>1}$ for $LS_-$ in Equation (8), we can derive estimates of the lure-detection probabilities $D^{\mathrm{w}}_{\mathrm{n}}$ and $D^{\mathrm{s}}_{\mathrm{n}}$ of 2HT theory, respectively.

## Experiment 2

In Experiment 2, the 2AFC test and the 4ART test were completed in two separate blocks. Besides the core prediction that $LS^{\mathrm{w}}_{>1} < LS^{\mathrm{s}}_{>1}$, we expected a general increase in target strength in the final 4ART test compared to Experiment 1 because participants were exposed to all targets for an additional time during the 2AFC test. This should lead to an increase in the number of targets being assigned Rank 1 in the 4ART test (i.e., larger $\pi^{\mathrm{w}}_1$ and $\pi^{\mathrm{s}}_1$ parameters compared to those in Experiment 1). As a result, the effect of $c^{\mathrm{w}}_2 < c^{\mathrm{s}}_2$ may diminish, or perhaps even disappear, for the 4ART test of Experiment 2 because participants have seen the weak target words twice by this time (once during study and once during the preceding 2AFC test) and the strong target words four times (three times during study and once during the 2AFC test). As such, weak targets could possibly catch up in relative target strength and become more similar to strong targets in the 4ART test because the increment in target strength from two to four presentations may be smaller compared to that from one to three presentations. However, similar to non-detected targets that were assigned Rank 1, an assimilation of weak and strong targets in the 4ART test would not threaten our crucial test of the conditional $LS$ estimates. Therefore, we did not make a strong prediction of whether this would occur. If the probability $\pi_1$ of assigning Rank 1 to a target increases with repeated exposure, the $LS_{>1}$ values will include fewer trials, thus rendering our test even more strict.

### *Method*

#### *Participants and materials*

None of the participants from Experiment 1 were recruited for Experiment 2. Forty-four individuals (28 females, 16 males) from the University of Mannheim's participant pool for psychological studies received €5.00 and a performance-based reward between €3.00 and €9.00 for participation. Because we expected a smaller difference in the $c_2$ values for strong versus weak target displays than in Experiment 1, we extended the period of data collection and tested as many participants as possible. This led to a doubling in sample size. A sensitivity analysis with G*Power 3.1 (Faul et al., 2009) showed that the achieved sample size of $N = 44$ suffices to detect an effect of size $d_z = 0.39$ in a directional Wilcoxon signed-rank test assuming an underlying normal distribution, significance level $\alpha = .05$, and power $1 - \beta = .80$. All participants were native or fluent speakers of German, and all except one person were undergraduate students. The sample's mean age was 23.05 years ($SD =$ 5.32, range = 18–41). To closely resemble Experiment 1, the study was conducted in the same computer laboratory, the word pool was selected by slightly extended criteria (valence of 3.0 to 7.0 and arousal of 0.5 to 4.5 on 11-point scales), the repetition scheme was maintained, and the same number of targets was studied. For each participant, 150 and 900 words were randomly drawn from the word pool to serve as old and new items in the recognition tests, respectively.

#### *Design and procedure*

The experiential session comprised a study phase, a 2AFC test, and a 4ART test. The study phase and the 4ART test were identical to Experiment 1. Between them, the 2AFC test with 300 displays in random order was administered. All test displays showed two items in rectangular boxes next to each other centered on the screen. The box for each item was randomly determined. For each of these item pairs, participants were asked to select the word that they thought was more likely to be new by clicking in its box. Participants were informed beforehand that 150 of the 300 displays would include a target and a randomly selected lure, while the other 150 displays would include two lures. Participants were also informed beforehand that they could win one reward point for each correct response, which implies a certain win of one point in case of two lures. To maintain a constant level of motivation, participants could take a short break after half of the 2AFC trials.

The lure–lure pairs were included to keep participants from actively restudying the words that they thought were the targets, and the number of gained points was only displayed at the end of the experiment to avoid trial-by-trial feedback. Selecting a lure in a list with only target–lure pairs and receiving negative feedback would have allowed participants to infer that the non-selected item must be the target. Given that another test was expected, such an inference could have prompted participants to focus their attention selectively on these items. As we were interested in lure detection, paying too much attention to the presumed targets could decrease the effort to detect lures. To ensure understanding of the unusual task structure, the 2AFC test only started after participants had correctly stated that they should find the word that was more likely new.

Following the 2AFC test, the 4ART test with 150 displays including all 150 targets and 450 different lures started. Each 4ART display consisted of one target and three lures in boxes randomly arranged in a 2 × 2 matrix on the computer screen, exactly as in Experiment 1. We used lures different from those of the 2AFC test to ensure that lures do not become familiar to the extent that they act as very weak targets, thus interfering with the true targets on the 4ART test. Again, as in Experiment 1, two practice displays explained the task and the reward scheme. In addition, the 4ART test only started after participants successfully indicated that the old word was now to be identified. For the 4ART test, the order of target presentation was randomized anew. At the end of the session, participants were thanked, debriefed, and compensated for participation.

### *Results*

#### *4ART analysis*

Table 1 shows the mean estimates of the unconditional probabilities $\pi_i$ and the conditional probabilities $c_2$. Ensuring that the study-repetition manipulation was successful, a directional Wilcoxon signed-rank test indicated that the estimates for $\pi^{\mathrm{s}}_1$ across participants were significantly larger than those for $\pi^{\mathrm{w}}_1$, $V = 896$, $p < .001$, $d_z = 1.17$. However, the difference between the $c^{\mathrm{s}}_2$ and $c^{\mathrm{w}}_2$ estimates was no longer significant, $V = 554.5$, $p = .246$, $d_z = 0.16$. As outlined above, we did not have a strong a-priori prediction, but expected such a finding, which is attributable to the assimilating effect of the preceding 2AFC test on weak and strong targets. Fig. 3B plots the individual $c_2$ estimates for strong against weak targets. The pairs of $c_2^{\mathrm{w}}$ and $c_2^{\mathrm{s}}$ are generally closer to the main diagonal than in Experiment 1. In line with this impression, the BF of the hierarchical model analysis was 0.98, which indicates evidence neither for $\mathcal{H}_0$: $\delta = 0$ nor for $\mathcal{H}_1$: $\delta > 0$, where δ is again the effect size capturing the mean difference between $c_2^{\mathrm{w}}$ and $c_2^{\mathrm{s}}$ on the group level. While keeping in mind that comparisons across different experiments without random assignment of participants need to be treated with caution, our data indicate that the unconditional probabilities $\pi_1^{\mathrm{w}}$ and $\pi_1^{\mathrm{s}}$ of assigning Rank 1 to targets were descriptively higher in Experiment 2 than in Experiment 1. This provides evidence for the expected boost in strength of both target types as a result of the additional presentation in the preceding 2AFC test.

### *2AFC analysis*

Table 2 shows the mean estimates of the lure-selection rates (*LS*), the lure-selection rates conditional on Rank-1 judgments in the 4ART test ($LS_1$), and the lure-selection rates conditional on rank judgments larger than 1 ($LS_{>1}$) separately for lures paired with a weak versus a strong target in the 2AFC test. The lure–lure pairs were not considered in the reported analyses. Directional Wilcoxon signed-rank tests showed that all three *LS* estimates were significantly higher for lures next to a strong target than for a weak target, all $V \geq 820, p < .001, d_z \geq 0.63$. Most importantly, the result for the $LS_{>1}$ estimates conditional on target non-detection provided evidence against the lure-detection invariance assumption, showing that lure detection in the 2HT model was more likely in the context of a strong than a weak target, even when the target was not detected in the 4ART test ($\mathcal{H}_1$: $LS_{>1}^{\mathrm{w}} < LS_{>1}^{\mathrm{s}}$). To assess interindividual heterogeneity, Fig. 5A and B plot the individual *LS* estimates for all targets and for the subset of targets with a rank larger than 1, respectively. In both panels, the majority of points lie above the main diagonal, showing that the expected effect was observed for the majority of participants.

Similar to the crucial test of the conditional probabilities $c_2$, the $LS_{>1}$ estimates are based on different numbers of observations across individuals and target-strength conditions. Therefore, we adapted the hierarchical Bayesian model using the number of correct lure selections in the 2AFC test conditional on target non-selection (instead of the frequencies of Rank-2 assignments) and the number of incorrect target selections (instead of the frequencies of Rank-3 and Rank-4 assignments) as observed response frequencies. The estimated BF indicated that the data were over 260 times more likely to have occurred under $\mathcal{H}_1$ than under $\mathcal{H}_0$, which provides decisive evidence for $\mathcal{H}_1$: $\delta > 0$, where δ is the effect size capturing the mean difference between $LS_{>1}^{\mathrm{w}}$ and $LS_{>1}^{\mathrm{s}}$ on the group level.

In addition to providing a test of our core hypothesis, the observed mean *LS* estimates in Table 2 can be used to derive estimates for the probabilities of lure and target detection of 2HT theory in the 2AFC test. Based on Equation (8), the observed conditional probabilities $LS_{>1}$ for weak and strong targets on the group level from Table 2 imply estimates of $D_{\mathrm{n}}^{\mathrm{w}} = .08$ and $D_{\mathrm{n}}^{\mathrm{s}} = .28$, respectively. Moreover, inserting these estimates in Equation (5) along with the observed unconditional *LS* estimates for weak and strong targets from Table 2 and solving for the corresponding target-detection probabilities results in estimates of $D_{\mathrm{o}}^{\mathrm{w}} = .20$ and $D_{\mathrm{o}}^{\mathrm{s}} = .33$. These estimates are remarkably similar to the model-based parameter estimates obtained for Experiment 1, despite different memory-test paradigms (4ART vs. 2AFC) and non-random assignment of participants to Experiments 1 and 2. The only exception is the higher $D_{\mathrm{n}}^{\mathrm{s}}$ parameter, which may hint at the possibility that $D_{\mathrm{n}}$ in the 2AFC test is larger than $D_{\mathrm{n}}$ in the 4ART test. This idea is supported by the process model outlined in the Appendix: Comparing a lure against a target in the 2AFC test will often result in a stronger familiarity contrast than comparing a lure against the mean of a target and two lures in the 4ART test.

**Table 2**
Mean Estimates of Lure-Selection Rates (and Standard Deviations) for Item Pairs With a Weak versus a Strong Target in the 2AFC Tests of Experiments 2 and 3.

| Experiment | Target | *LS* | $LS_1$ | $LS_{>1}$ |
|---|---|---|---|---|
| Experiment 2 | Weak | .63 (.09) | .70 (.11) | .54 (.09) |
| | Strong | .76 (.11) | .82 (.11) | .64 (.13) |
| Experiment 3 | Weak | .64 (.09) | .95 (.06) | .29 (.15) |
| | Strong | .78 (.10) | .96 (.06) | .35 (.18) |

*Note.* $LS$ = estimated probability of correctly selecting the lure in a target–lure pair; $LS_1$ = estimated probability of correctly selecting the lure conditional on the target being assigned Rank 1 in the subsequent 4ART display; $LS_{>1}$ = estimated probability of correctly selecting the lure conditional on the target not being assigned Rank 1 in the corresponding 4ART display.

## *Discussion*

As outlined above, the forced-choice-then-ranking task allows for an additional, independent test of the lure-detection invariance hypothesis embedded in Kellen and Klauer's (2014) 2HT model. Of primary interest here are the conditional probabilities $LS_{>1}$ in the 2AFC test, which refer only to target–lure pairs with a target that was not detected in the subsequent 4ART test. Based on the assumption that these targets were also not detected in the preceding 2AFC test, participants can only rely on lure detection or guessing in order to decide which item of the pair is more likely new. Empirically, we found unequivocal evidence in Experiment 2 for $\mathcal{H}_1$: $LS_{>1}^{\mathrm{w}} < LS_{>1}^{\mathrm{s}}$. This finding supports our prediction that the probability of detecting a lure in 2HT theory increased with the strength of the target next to the lure—even when the target itself was not detected.

To reiterate, this argument rests on the assumption in 2HT theory that targets detected in the 2AFC test will also be detected in the subsequent 4ART test, or equivalently that targets not detected in the 4ART test were also not detected previously in the 2AFC test. For Experiment 2, one might object that detecting a target at time point $t_1$ and not detecting it at a later time point $t_2$ reflects the nature of forgetting. Accordingly, the subset of targets not detected in the 4ART test could still include words that were detected in the 2AFC test, thereby threatening our conclusion. However, the general argument that the state of an item changes between both tests ignores two important aspects. First, the time difference $t_2 - t_1$ between the 4ART and the 2AFC tests was very small, which leaves not much time for interim forgetting. Second and more importantly, forgetting studies typically compare the performance in a *single* memory test administered at two points in time, usually including a different set of learned targets. In contrast, we employed two *separate* memory tests in succession using exactly the same targets. Such a repeated memory-testing design resembles testing-effect studies more closely than forgetting studies (e.g., Roediger & Karpicke, 2006; Rowland, 2014). According to the well-known testing effect, repeated testing leads to an improvement in memory across time for previously retrieved items as compared to non-retrieved items (e.g., Halamish & Bjork, 2011; Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012; Rowland & DeLosh, 2015). Thus, if anything, the detection probability of targets is more likely to increase rather than to deteriorate during this short time interval.

Our data support this idea. The higher $\pi_1$ estimates in Experiment 2 compared to Experiment 1 suggest that the preceding 2AFC test increased participants' performance. Although the majority of findings in the testing-effect literature refers to long retention intervals and recall tests, the 2AFC test still provides participants with the opportunity to restudy recognized and presumed targets. Whether participants actually use this opportunity, we cannot control. However, by interspersing the target–lure pairs with lure–lure pairs and by informing participants that only half of the displays included targets, we prevented intentional restudying of those words that were thought to be the targets. Similarly, the non-significant difference between the $c_2$ estimates in the 4ART test may also be the result of the learning effect from the 2AFC test, causing an assimilation of strong and weak targets. As in monotonic, negatively accelerated hyperbolic learning curves, the greatest increments in learning may occur during the first few occurrences of the material (see

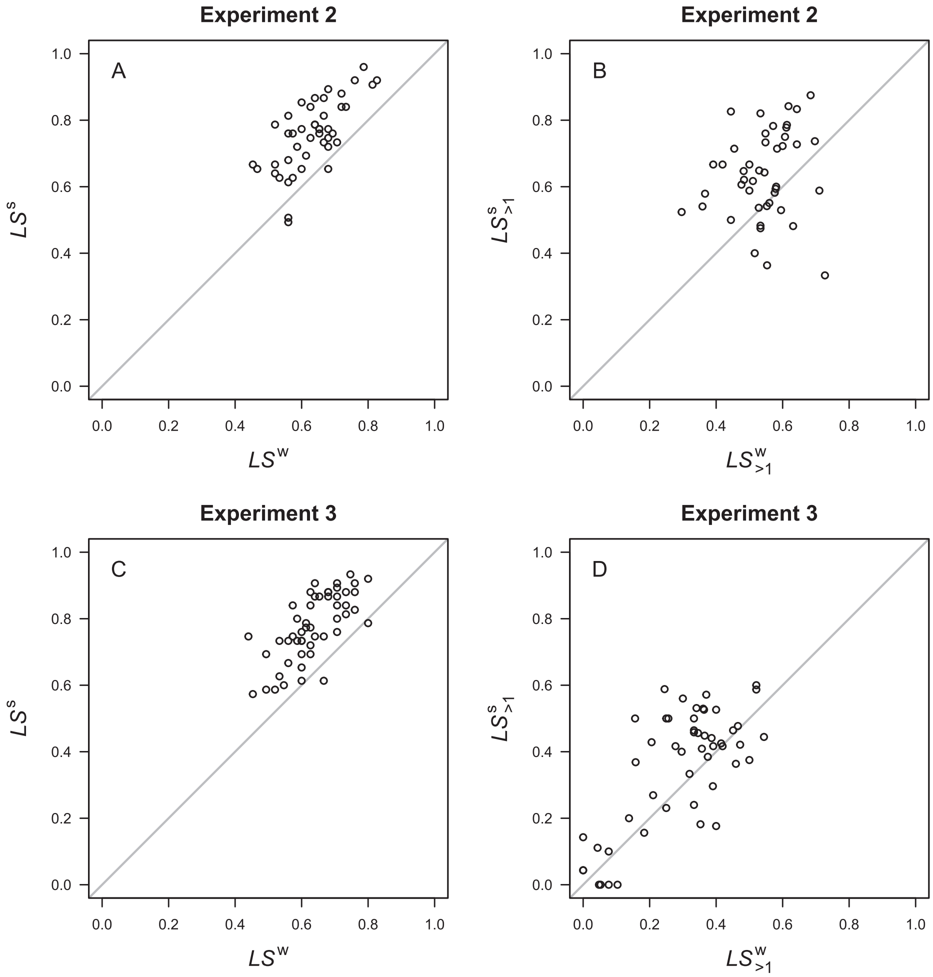

**Fig. 5.** Individual lure-selection rates and lure-selection rates conditional on target non-detection in the 2AFC tests of Experiments 2 and 3. $LS$ = estimated probability of correctly selecting the lure in a target–lure pair, $LS_{>1}$ = estimated probability of correctly selecting the lure in a target–lure pair conditional on the target not being assigned Rank 1 in the corresponding 4ART display. (A–B) Experiment 2, (C–D) Experiment 3.

Greene, 1992). Hence, the difference between one and two presentations may be more pronounced than between two and three.

Alternatively, the observed increase in Rank-1 judgments and the failure to replicate the effect of $c_2^w < c_2^s$ in Experiment 2 could also be explained by participants' awareness of which words were repeated (namely the targets). It is not unlikely that participants recognize a certain word on the 4ART test and remember that it was already presented during the 2AFC test. If they then make the reasonable conclusion that this word must be a target, it would boost performance over all targets. This would certainly lead to more targets being detected. However, this per se is unproblematic for our critical test because all of these targets would be excluded in the calculation of the $LS_{>1}$ probabilities, leading to a stricter test of our hypothesis.

Although it seems unlikely from a threshold perspective that participants detect a target in the 2AFC test and fail to detect the same target in the subsequent 4ART test, we cannot definitely rule out that such events occur. The main reason for this concern is that familiarity contrasts underlying detection probabilities in the 2AFC and the 4ART tests are assumed to be conditionally independent, unless targets or lures occur repeatedly in the same test trial (see the Appendix). Because the 2AFC display and the corresponding later 4ART display were not part of the same test trial in Experiment 2, target and lure detection in 2HT models may probabilistically vary between both tasks. Hence, our test based on the $LS_{>1}$ estimates might be contaminated by some test pairs in which the target was actually detected in the 2AFC test. As this may compromise the logic of Experiment 2, we decided to replicate the crucial test of $LS^w_{>1} < LS^s_{>1}$ with an improved experimental procedure.

## Experiment 3

In Experiment 3, the 2AFC test and the 4ART test were not completed in two separate blocks, but each 2AFC question was immediately followed by the corresponding 4ART question in the same test trial, with the 4ART question including the 2AFC item pair and two additional items. From a threshold perspective, this procedure ensures that detection states for the 2AFC items remain stable across tasks, and thus target detection in the 2AFC display goes along with target detection in the subsequent 4ART display. Thus, at least for this refined procedure, the $LS_{>1}$ estimates derived from forced-choice-then-ranking paradigm can be interpreted as correct lure selections conditional on target non-detection. We expected the following results for the new test procedure. First, a replication of the $LS^w_{>1} < LS^s_{>1}$ effect found in Experiment 2 (assuming that lure detection is in fact facilitated by increasing target strength even when the target is not detected). Second, a replication of the $c_2^w < c_2^s$ effect found in Experiment 1 (because in contrast to Experiment 2 participants can no longer rely on second-guessing which words are repeated in the 4ART part of the test).

### *Method*

#### *Participants and materials*

Fifty-two students (37 females, 13 males, 2 non-binary people) of the

University of Mannheim were recruited for Experiment 3—none of which had previously participated in Experiment 1 or 2. Participants could choose between €5.00 or partial course credit for their participation, and received a performance-based reward between €3.00 and €9.00 in both cases. All participants were native or fluent speakers of German, and all except two were undergraduate students. The sample's mean age was 22.31 years (*SD* = 4.02, range = 18–38). Participant 4 was excluded from all further analyses because of a correct response rate of 100% for strong targets, ruling out calculation of relevant statistics. As in Experiment 2, we used a longer data-collection period than in Experiment 1 and tested as many participants as possible aiming at a larger sample size than in Kellen and Klauer's (2014) experiments. A sensitivity analysis with G*Power 3.1 (Faul et al., 2009) revealed that the sample size of $N = 51$ allows to detect a relatively small effect of size $d_z = 0.36$ in a directional Wilcoxon signed-rank test assuming an underlying normal distribution, significance level $\alpha = .05$, and power $1 - \beta = .80$. The experiment was conducted in the same computer laboratory as the previous experiments. The word pool from Experiment 2 was used and the repetition scheme was maintained. For each participant, 200 and 600 words were randomly drawn from the word pool to serve as old and new items in the recognition test, respectively.

### *Design and procedure*

The experiment comprised a study phase and a single test phase. The study phase was identical to the study phases in Experiments 1 and 2, except that more targets were studied to compensate for the lure–lure pairs in the 2AFC test, as each 2AFC display had to be followed by a 4ART display including a target. A total of 100 words were studied once and 100 words were studied three times, followed immediately by the test phase with 200 trials. On each trial of the test phase, two words were shown for the 2AFC question ("Which word is more likely a new word?"), and when the participant had made a response, two more words appeared for the 4ART question ("Please rank the words from most likely old to most likely new."). As before the words were displayed in rectangular boxes. The first two boxes were shown next to each other in the center of the screen. When the additional two words appeared, all four boxes were arranged according to a 2 × 2 matrix and the position of each word was randomly determined anew to ensure that participants had to read all words presented on the screen when answering the 4ART question.

All 200 item quadruples in the test phase consisted of one (weak or strong) target and three lures. On 75 of the 200 test trials, the word pair shown in the 2AFC display consisted of a randomly selected lure and a strong target, and on another 75 trials, the word pair consisted of a randomly selected lure and a weak target. On the remaining 50 trials, participants saw two lures (half of which were followed by 4ART displays with a weak target or a strong target, respectively). The order of trials was randomized. To avoid a drop in motivation, participants could take a break after half of the trials. The lure–lure pairs were included to make sure that participants could not infer that one word of the pair had to be a target. Such a conclusion would have prompted them to always assigning Rank 1 to one word of the 2AFC pair. Unknown to participants, only 25% lure–lure pairs were used to avoid having unnecessarily long study and test lists.[6]

Prior to the test phase, as in Experiment 2, participants completed two practice displays, which explained the tasks and the reward scheme. Participants were then quizzed on the questions they had to answer for the item pairs and the item quadruples, and they were only allowed to proceed without re-reading the instructions if their answers were correct. Finally, participants were informed that some displays would include two lures, but they were not told about the exact proportion of lure–lure pairs. After completing the test phase, participants were thanked, debriefed, and compensated for participation.

## *Results*

### *4ART analysis*

The mean estimates of the unconditional probabilities $\pi_i$ and the conditional probabilities $c_2$ are shown in Table 1. They were calculated based on all 4ART displays (denoted as Experiment 3) and based on 4ART displays that were preceded by a target–lure pair on the 2AFC display (denoted as Experiment 3*). Directional Wilcoxon signed-rank tests indicated that $\pi_1^s$ was significantly higher than $\pi_1^w$ across participants for all displays, $V = 1225, p < .001, d_z = 2.11$, as well as for the subset of displays following a target–lure pair, $V = 1320, p < .001, d_z = 2.00$. The difference between the $c_2^s$ and $c_2^w$ estimates was also significant for all displays, $V = 957, p = .003, d_z = 0.43$, as well as for the subset of displays following a target–lure pair, $V = 915, p = .004, d_z = 0.43$. Fig. 3C and D depict the individual $c_2$ estimates for strong targets against weak targets for all displays and the relevant subset of displays, respectively. In line with the results of the statistical tests, both figures are more similar to Experiment 1 than Experiment 2 in the sense that the majority of observations lie above the main diagonal. This visual impression was also supported by the BF of the hierarchical model analysis comparing $\mathcal{H}_1$: $\delta > 0$ to $\mathcal{H}_0$: $\delta = 0$, where $\delta$ is the effect size capturing the mean difference between $c_2^w$ and $c_2^s$ on the group level. The BF was 38 for all displays and 22 for the subset of displays following a target–lure pair.

### *2AFC analysis*

The mean estimates of the unconditional and conditional lure-selection rates for target–lure pairs are shown in Table 2. The lure–lure pairs were not considered in these analyses. The unconditional lure-selection rates (*LS*) were very similar to Experiment 2, whereas the lure-selection rates conditional on Rank-1 judgments in the 4ART test ($LS_1$) were higher and the lure-selection rates conditional on rank judgments larger than 1 ($LS_{>1}$) were consequently lower. Fig. 5C and D plot the individual *LS* estimates for all target–lure pairs (*LS*) and for the subset of pairs including a target that was later assigned a rank larger than 1 ($LS_{>1}$), respectively. As in Experiment 2, the majority of participants showed the expected effects as they lie above the main diagonals.

Directional Wilcoxon signed-rank tests indicated that the *LS* estimates were significantly higher for strong than for weak targets, $V = 1319.5, p \leq .001, d_z = 1.92$, whereas the difference in the $LS_1$ estimates failed to reach the level of statistical significance, $V = 532, p = .051, d_z = 0.22$, most likely because of a ceiling effect. More importantly, the directional Wilcoxon signed-rank test for the $LS_{>1}$ estimate was significant, $V = 988.5, p = .001, d_z = 0.48$. Likewise, the BF showed that the alternative hypothesis ($\mathcal{H}_1$: $LS_{>1}^w < LS_{>1}^s$) was 38 times more likely than the null hypothesis ($\mathcal{H}_0$: $LS_{>1}^w = LS_{>1}^s$), which can be interpreted as providing very strong evidence against the lure-detection invariance assumption.

Although the results of Experiment 3 were in line with our prediction of higher $LS_{>1}$ estimates in the context of strong than weak targets, the observed mean $LS_{>1}$ estimates for weak and strong targets in Table 2 are smaller than .50, leading to negative estimates of 2HT theory's $D_n$ if we were to apply Equation (7) to the data of Experiment 3. In words of 2HT theory, the majority of participants performed below chance level in the 2AFC test when focusing on item pairs with a target that was later not detected in the 4ART test (see Fig. 5D). How can this be explained? A plausible explanation is suggested by the nature of the forced-choice-

[6] A consequence of the unequal ratio of target–lure and lure–lure pairs is the higher likelihood that the target was already presented in the 2AFC display and did not only appear in the 4ART display. Hence, from a threshold perspective, participants may (a) no longer show equiprobable guessing (i.e., guessing which word is the non-detected target with probabilities of $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$ depending on the number of detected lures) and (b) rather show more consistent responding between both displays (i.e., a tendency to assign the two words from the 2AFC display to Ranks 1 and 4 in the 4ART display, while basically ignoring the two additional words that appear in the 4ART display and assigning them to Ranks 2 and 3). We account for both possibilities in our data analysis.

then-ranking task employed here. Because each 2AFC display is immediately followed by a 4ART display including the same item pair (making it a 2AFC-4ART trial), participants may have been inclined to answer consistently in both parts of each trial. Consistent judgments are generally considered desired behavior. This could have encouraged participants to match their 4ART response with their preceding 2AFC response whenever they believed that one of the 2AFC items was the target (a belief fostered by the fact that there were three times as many target–lure pairs than lure–lure pairs).

Let us assume that consistent responding occurs for both correct and incorrect 2AFC judgments and irrespective of target strength. If participants aiming at consistent responding incorrectly select the target in the 2AFC display, they will consequently assign Rank 1 to the lure and a rank of larger than 1 to the target in the 4ART display. Note that these consistent errors would diminish our core measure, the $LS_{>1}$ estimate, as pairs with the target being assigned a rank larger than 1 in the 4ART test are included in this measure. Conversely, if participants aiming at consistent responding correctly select the lure in the 2AFC display, they will consequently assign Rank 1 in the 4ART display to the other item correctly assumed to be the target. These correct lure selections in the 2AFC test, however, are excluded from our $LS_{>1}$ estimates. Hence, if consistent responding happens on many trials, a relatively large proportion of incorrect 2AFC judgments are included in our $LS_{>1}$ measure and a large proportion of correct judgments are excluded from it—resulting in systematic below-chance $LS_{>1}$ estimates. More importantly, if we assume that 2AFC errors occur more frequently when the target is weak than when it is strong, the consistent-response bias in $LS_{>1}$ estimates would be sufficient to explain the observed $LS^{w}_{>1} < LS^{s}_{>1}$ pattern, without the necessity to assume any effect of target strength on lure-detection probabilities.

*Model-based analysis*

To address the concern mentioned above and to test our post-hoc hypothesis that consistent responding can explain the below-chance performance in the 2AFC displays of Experiment 3, we added a model-based analysis by adapting our 2HT model for the ranking task of Experiment 1 (Fig. 2A) to the forced-choice-than-ranking task of Experiment 3 (Fig. 6). Importantly, rather than modeling the rank of the target in the 4ART display, we modeled the combined results of the 2AFC and the 4ART display as one 2AFC-4ART trial. As before with the $LS_{>1}$ estimates, we only looked at 2AFC-4ART trials with target–lures pairs in the 2AFC part. The 2AFC part of each trial had two possible response outcomes (L for correctly selecting the lure and T for

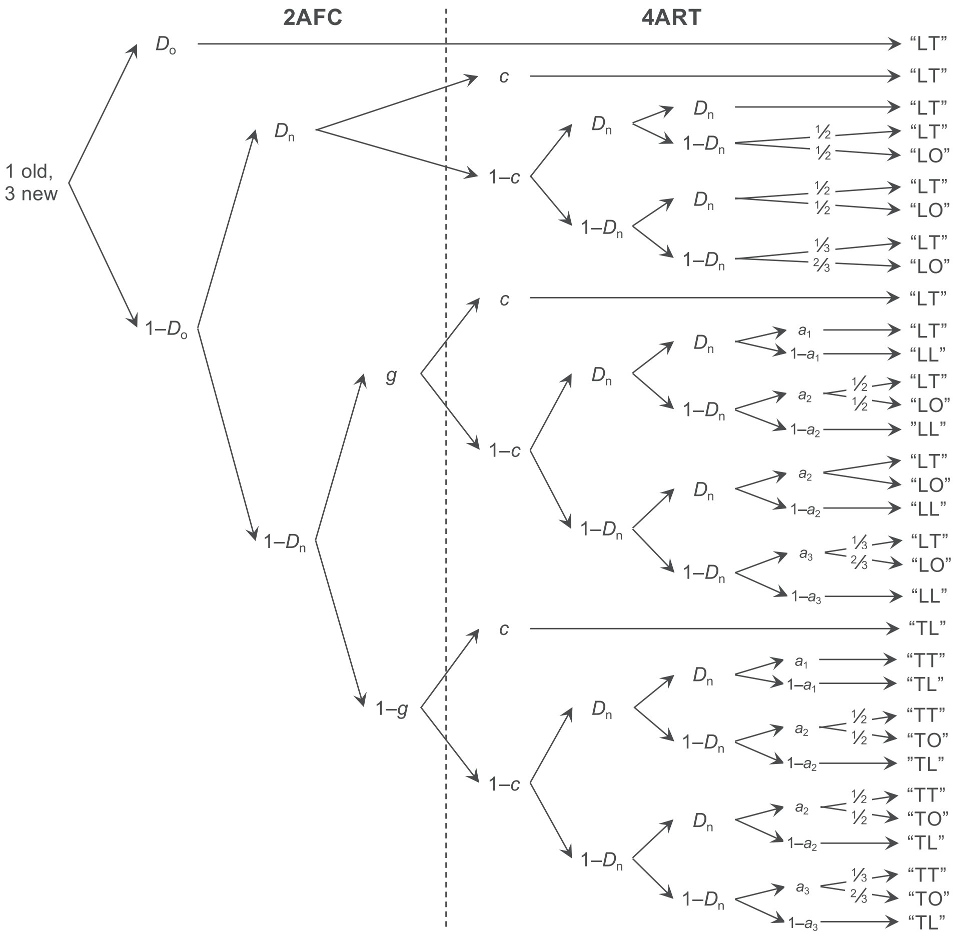

**Fig. 6.** New two-high-threshold model for the forced-choice-then-ranking task in Experiment 3. The item quadruple on the left-hand side consists of one old item and three new items. The probabilities of target and lure detection (the latter conditional on target non-detection) are denoted as $D_o$ and $D_n$, respectively. The lure-guessing probability for the 2AFC display of each trial is given by $g = .50$. The probability of consistent responding within the 2AFC-4ART trial is called $c$. The guessing probabilities for the 4ART display depend on participants tendency to avoid response categories that are incompatible with the task structure (given by $a_1$, $a_2$, and $a_3$ depending on the total number of task-compatible response options) and pure guessing (given by $\frac{1}{2}$ and $\frac{1}{3}$ depending on the total number of non-detected items). The letters in quotation marks on the right-hand side represent the six response categories, which result from crossing the response options in the 2AFC part and the 4ART part of a trial. The dashed vertical line separates the cognitive states involved in the 2AFC part of each trial from the cognitive states in the 4ART part.

incorrectly selecting the target), whereas the 4ART part had three possible response outcomes (T for correctly assigning Rank 1 to the target, L for incorrectly assigning Rank 1 to the lure from the 2AFC pair, and O for incorrectly assigning Rank 1 to one of the two other lures not present in the 2AFC pair). This leads to six possible 2AFC-4ART outcome combinations in total: LT, LL, LO, TT, TL, and TO. These six categories are repeated in two separate trees—one for trials with a weak target and the other for trials with a strong target.

Fig. 6 illustrates the processing-tree structure of the 2HT-2AFC-4ART model. In the 2AFC display, participants either detect the target with probability $D_\mathrm{o}$ or fail to detected it. When target detection fails, participants can detect the lure with conditional probability $D_\mathrm{n}$ or fail to detect it. When target and lure detection both fail, participants are assumed to guess a response with fixed probability $g = .50$. In the 4ART display, participants aim at consistent responding with probability $c$ and will—after a correct 2AFC response—assign Rank 1 to the target and do not process the two additional lures. When participants decide to process the two additional lures, they can detect each with independent probability $D_\mathrm{n}$. Importantly, the target and the lure of the 2AFC display will not change their memory state from the 2AFC to the 4ART display because they are repeated within the same test trial (i.e., there are only two more items added). When the target and at least one lure are undetected in the 4ART display, participants will arrive at a ranking response by guessing among these items. This guessing process can be pure, such that the guessing probabilities can be fixed to equiprobable guessing of $\frac{1}{2}$ and $\frac{1}{3}$, or it can be motivated by participants' tendency to avoid a response that would be incompatible with the task structure (i.e., LL for selecting the lure in the 2AFC part and assigning Rank 1 to that lure in the 4ART part, and TT for selecting the target in the 2AFC part and assigning Rank 1 to the target in the 4ART part of each trial). It is assumed that participants will avoid such a response category with unknown probabilities $a_1$, $a_2$, and $a_3$, depending on whether they have to decide among an incompatible response and one, two, or three other responses, respectively. For instance, when the lure was selected through guessing in the 2AFC part of a trial and the two additional lures in the 4ART part were detected, only the non-detected target and the non-detected lure from the 2AFC part compete for Rank 1. As participants then want to avoid the LL response, they will more likely select the target.

The $c$ parameter and all guessing probabilities are assumed to be independent of item type, participants' memory state, and target strength. Hence, applied simultaneously to the weak-target and strong-target conditions, the model accounts for $2\cdot(6-1) = 10$ free category probabilities with 8 free parameters. As the model was developed post hoc to account for below-chance performance on the 2AFC test, the design of the 2AFC-4ART test was not ideally suited for a model-based analysis. Many participants showed empty cells with observed zero frequency for response categories LL and TT (for weak-target and strong-target displays). Although this lends support to the idea that participants avoid inconsistent responses, it is problematic for model fitting as empty cells can lead to some parameters being non-identifiable. Therefore, we added a constant of 0.1 to every cell frequency, which is common practice in MPT modeling (see Hu, 1991; Klauer, Stahl, & Erdfelder, 2007; Rothkegel, 1999).

The 2HT model for the 2AFC-4ART paradigm with equiprobable guessing shown in Fig. 6 was fitted to the data of each individual separately. In total, responses of 44 of the 51 participants were described well by the model, all $G^2(2) \leq 5.90$, $p \geq .052$, whereas responses of seven participants were not described well, $G^2(2) \geq 6.07$, $p \leq .048$. The mean maximum-likelihood estimates for consistent responding ($c = .46$) and target-guessing ($a_1 = .61$, $a_2 = .68$, and $a_3 = .79$) were reasonable and within the expected range. More importantly, the mean maximum-likelihood estimates (and the corresponding standard errors of the mean) for item detection were .17 (.02) for $D^\mathrm{w}_\mathrm{o}$, .35 (.03) for $D^\mathrm{s}_\mathrm{o}$, .12 (.02) for $D^\mathrm{w}_\mathrm{n}$, and .28 (.03) for $D^\mathrm{s}_\mathrm{n}$, which closely resemble the estimates calculated from the lure-selection rates in Experiment 2.[7]

Fig. 4C plots the individual $D_\mathrm{o}$ estimates for strong targets against the individual $D_\mathrm{o}$ estimates for weak targets. The majority of estimates lie above the main diagonal. A directional Wilcoxon signed-rank test showed that the $D^\mathrm{s}_\mathrm{o}$ estimates were on average significantly larger than the $D^\mathrm{w}_\mathrm{o}$ estimates, $V = 1064$, $p < .001$, $d_z = 0.74$, supporting the idea that more target repetitions during studying resulted in a higher probability of target detection during testing. Fig. 4D plots the individual $D_\mathrm{n}$ estimates of the lure-detection parameter for strong targets against weak targets. The majority of participants showed higher $D^\mathrm{s}_\mathrm{n}$ than $D^\mathrm{w}_\mathrm{n}$ estimates, and a directional Wilcoxon signed-rank test confirmed that the probabilities of lure detection in the 2HT model were on average significantly higher in the context of a strong target than in the context of a weak target, $V = 1044$, $p < .001$, $d_z = 0.57$. This again supports our core hypothesis that target strength does not only affect the probability of target detection in 2HT theory, but also the probability of lure detection conditional on target non-detection.

### Discussion

Experiment 3 replicates the result of higher lure-selection rates conditional on strong targets with ranks larger than 1 as compared to weak targets with ranks larger than 1 ($LS^\mathrm{w}_{>1} < LS^\mathrm{s}_{>1}$) that was found in Experiment 2 in an improved experimental design. Because each 2AFC display was immediately followed by the corresponding 4ART display (as opposed to the block-wise design in Experiment 2), it is reasonable for any threshold model to assume that the detection state of a target in 2AFC displays remains stable when two more lures are added for the 4ART display. A possible problem was that a bias toward consistent responding across 2AFC and 4ART judgments alone could account for the observed $LS^\mathrm{w}_{>1} < LS^\mathrm{s}_{>1}$ pattern in Experiment 3. However, by explicitly modeling consistent responding in a 2HT model for the 2AFC-4ART paradigm, we showed that $D^\mathrm{w}_\mathrm{n} < D^\mathrm{s}_\mathrm{n}$ still holds for conditional lure-detection probabilities.

Taken together, the forced-choice-then-ranking procedure used in Experiment 3 overcomes the main concern we had with Experiment 2. Because every target in Experiment 2 was tested in the 2AFC test and the 4ART test, participants might have recognized which word was present on both tests and inferred that this word must be the target. Nevertheless, another concern could not be ruled out completely in Experiment 3. The context of target presentation still changes between the 2AFC display and the 4ART display as two additional lures appear for the 4ART test. In principle, such a change in context could activate different components of memory, which in turn could affect the target-detection probability in the 2HT model. However, first, it is extremely implausible for any threshold model to assume that a target just detected in the 2AFC display will transition to a non-detection state in the 4ART display. Second, even if the context change affects the target-detection probability, it will make detection more likely, and the trials on which the target was assigned Rank 1 are excluded in the crucial $LS_{>1}$ analysis. Hence, if anything, the new procedure provided an even stricter test than Experiment 2.

## General discussion

The debate between continuous-strength and discrete-state models of recognition memory has been an active field of research for decades

[7] As a robustness analysis, we also fitted a hierarchical version of the 2HT model for the 2AFC-4ART paradigm (Klauer, 2010). The hierarchical group-level estimates and credibility intervals did not differ substantially from the posterior means and standard deviations aggregated across the individual parameter estimates reported in the main text. The model fitted the observed response frequencies and covariance structure very well as indicated by large posterior predictive $p$-values of .212 and .654 for test quantities $T_1$ and $T_2$, respectively.

(e.g., Bröder & Schütz, 2009; Dubé & Rotello, 2012; Dubé, Starns, Rotello, & Ratcliff, 2012; Egan, 1958; Green & Swets, 1966; Kellen, Klauer, & Bröder, 2013; Province & Rouder, 2012; Wixted, 2007; Yonelinas & Parks, 2007). The corresponding models—the SDT model and the 2HT model—make different assumptions about memory representations and decision processes. The SDT model assumes that familiarity signals elicited by test items vary on a latent strength-of-familiarity continuum according to probability distribution for targets and a probability distribution for lures, and that a criterion placed along the familiarity continuum determines the recognition response. The 2HT model assumes that test items enter one of three discrete memory states (old-detection, new-detection, and non-detection) and that a probabilistic response decision is made in case of item non-detection.

The classical approach of evaluating the goodness-of-fit of the rival models to ROC data did not reveal clear-cut conclusions and requires strong distributional assumptions. As a remedy, the focus of the debate has recently shifted towards new methods for testing the competitors. Here we investigated the underlying assumptions of an *experimentum crucis* reported by Kellen and Klauer (2014) that compares simple observed response frequencies in a ranking task for which the models make competing predictions. Specifically, using a multiple-paradigm approach and taking a 2HT perspective, we tested whether lures in a $K$-alternative ranking task are more easily detected in the context of a strong versus a weak target—even when the target itself was not detected. The answer to this question is central for predictions of the 2HT model put forward by Kellen and Klauer, but also bears relevance to the more general question of how to adapt recognition models to new experimental paradigms.

### *Summary*

Kellen and Klauer's (2014) adaptation of the 2HT model to ranking tasks predicts that the conditional probability of targets being assigned Rank 2 given that they were not assigned Rank 1 ($c_2$) must be equal for weak and strong targets. Consequently, when Kellen and Klauer found that the $c_2$ estimates were higher for strong than for weak targets, they dismissed the 2HT model for $K$ARTs. However, the prediction only holds for 2HT models that assume invariance of the lure-detection probability ($D_\mathrm{n}$) to changes in target strength—a strong auxiliary assumption that we termed lure-detection invariance assumption. We questioned this assumption as it entered into Kellen and Klauer's formal argument.

After formally showing that an alternative adaptation of 2HT theory to the $K$ART paradigm without the lure-detection invariance assumption—the 2HT-$K$ART model—predicts higher $c_2$ values for strong targets, we empirically tested this crucial assumption in different ways. In Experiment 1, we replicated the effect of target strength on the conditional probabilities of assigning Rank 2 to targets ($c_2^\mathrm{w} < c_2^\mathrm{s}$). The 2HT-$K$ART model provided a higher estimate of the lure-detection probability in the context of a strong than a weak target even when the target was not detected ($D_\mathrm{n}^\mathrm{w} < D_\mathrm{n}^\mathrm{s}$). In Experiments 2 and 3, we implemented a new forced-choice-then-ranking paradigm and analyzed the lure-selection rates for lures in a 2AFC pair conditional on targets with ranks other than Rank 1 in the subsequent 4ART test, which can be interpreted as target non-detection according to the target-detection dominance assumption. The results supported the findings of Experiment 1 in two empirical tests ($LS_{>1}^\mathrm{w} < LS_{>1}^\mathrm{s}$). Furthermore, the mean estimates of $D_\mathrm{n}^\mathrm{w}$ and $D_\mathrm{n}^\mathrm{s}$ observed in Experiment 1 were matched by corresponding estimates in Experiment 2 and by mean estimates from a 2HT-2AFC-4ART model accounting for below-chance performance on the 2AFC displays of Experiment 3 through the process of consistent responding.

### *Possible remaining questions*

#### *Variants of 2HT models and the principle of complete information loss*

MPT models assume that observable responses depend on the latent cognitive state from which they emerge and not on the conditions or processes that led into this state (Krantz, 1969; Province & Rouder, 2012). For example, when an item is not detected and thus in the uncertainty state according to 2HT, all information about the item's history will be lost (assumption of complete information loss; Heck & Erdfelder, 2016; Kellen & Klauer, 2015; Swagman, Province, & Rouder, 2015). This entails that non-detected targets and lures are indistinguishable in the uncertainty state. Hence, the question arises how it is possible that the strength of a non-detected target may nevertheless affect the conditional lure-detection probability $D_\mathrm{n}$. Although this may appear like a violation of the information loss principle upon first sight, it is not.

To see this, it is important to acknowledge that the proposed 2HT adaptations to different ranking tasks do not consider items in isolation. In contrast to the standard 2HT model for old–new recognition, the 2HT-$K$ART model and the 2HT-2AFC-$K$ART model both refer to test displays consisting of one target and $K-1$ lures. Each of the $K$ test items in a test display can then be either in a detection state or in a non-detection state, which results in $2^K$ possible memory-state combinations for an item tuple as the relevant level of analysis. The probability of entering a detection state for each item depends on properties of the entire item tuple—one property being the strength of the target. As detailed in the Appendix, the detection probabilities can be conceived as results of familiarity contrasts between each item in the display and the $K-1$ items that form its context. It is thus to be expected that the probability of detecting a lure in the context of a strong target ($D_\mathrm{n}^\mathrm{s}$) exceeds the probability of detecting the same lure in the context of a weak target ($D_\mathrm{n}^\mathrm{w}$) simply because the familiarity contrast is likely to be more extreme in the former case. Notably, this does not depend on whether the conditionally independent target-detection process itself results in a positive or a negative outcome. In sum, our proposed 2HT models for ranking tasks assumes that ranking judgments depend on the specific $K$ART-tuple of memory states from which they emerge. Judgments are not influenced by the conditions and processes that led into this specific memory-state combination. Hence, there is no violation of the information-loss principle in the 2HT adaptations for ranking tasks proposed here.

#### *The necessity of 2HT adaptations to different judgment tasks*

It is important to remember that the standard 2HT model in Fig. 1 was originally developed for binary responses (e.g., yes–no, same–different, old–new). In order to model ranking judgments in a $K$ART test, the $K$ items presented as the test tuple need to be compared—even when assuming that they were initially processed in isolation. This makes the task a multiple-item discrimination task and not a single-item recognition task. Hence, the standard 2HT model requires appropriate modifications to account for this specific task structure. Importantly, these are not ad-hoc modifications, but adaptations of models to different research paradigms (i.e., how the models can handle multiple-item discrimination in 2AFC tasks, ranking tasks, and combinations thereof). This is in line with the philosophy underlying MPT modeling, according to which measurement models need to be adapted and tested for each experimental paradigm anew even when they were derived from the same underlying psychological theory (Erdfelder et al., 2009) and therefore adhere to the same set of core assumptions (Kellen et al., 2021). Put differently, our 2HT model is not a post-hoc modification of Kellen and Klauer's (2014) 2HT model to account for $c_2^\mathrm{w} < c_2^\mathrm{s}$. It is simply a different adaptation of 2HT theory to ranking tasks, which performed better in Experiments 1 to 3 of our current research than Kellen and Klauer's adaptation.

### *Implications for ranking tasks*

What are the implications of our results with respect to the modeling of ranking data? Importantly, most of Kellen and Klauer's (2014) conclusions still hold. Our results do not affect the conclusion that variants of continuous-strength models, such as the SDT model, are in line with their critical test. This is not only true for SDT assuming normal distributions, but for many alternative continuous familiarity distributions. However, whereas Kellen and Klauer concluded that the 2HT theory did not pass their critical test, we showed that 2HT theory is well compatible with the observation of $c_2^{\mathrm{w}} < c_2^{\mathrm{s}}$. As supported by the three experiments reported here, the probability of lure detection monotonically increases with target strength. Discrete-state models in the 2HT framework that take this finding into account are therefore capable of predicting $c_2^{\mathrm{w}} < c_2^{\mathrm{s}}$. Thus, it can be argued that ranking data are in line with versions of discrete-state models in the 2HT framework that assume that discrete states mediate between a latent familiarity continuum and observable responses.

Further support for discrete-state models was reported by Kellen et al. (2016). The authors pointed out that Luce's (1963) LT model is also in line with Kellen and Klauer's (2014) critical test. The LT model assumes just two memory states—detection and non-detection—rather than three states as the 2HT model. The threshold between the two states of the LT model is supposed to be "low" in the sense that it can be exceeded by lures and targets with probabilities $q_{\mathrm{n}}$ and $q_{\mathrm{o}}$, respectively (with $q_{\mathrm{n}} \leq q_{\mathrm{o}}$). Thus, in contrast to the 2HT model, both item types can reach the same detection state, which allows for the possibility that a lure is assigned Rank 1 and the detected target is assigned Rank 2. Hence, the finding $c_2^{\mathrm{w}} < c_2^{\mathrm{s}}$ is fully compatible with the LT model.

### *Implications for Kellen and Klauer's critical test*

Given that variants of the SDT model, the 2HT model, and the LT model pass the test proposed by Kellen and Klauer (2014), can we conclude that their critical test is much less diagnostic than previously thought? We would not fully subscribe to such a conclusion because the test criterion is strong enough to rule out a number of alternative discrete-state models discussed in the literature. In particular, any threshold model that (1) allows only targets but not lures to enter a detection state, (2) assumes that items in the detection state are not always ranked higher (i.e., smaller rank values) than items in non-detection states, or (3) assumes a fixed, invariant probability $D_{\mathrm{n}} = a$ for lure detection is in conflict with the observation that conditional Rank-2 probabilities increase with target strength. One prominent model that falls in this class of to-be-rejected models is Blackwell's (1953) one-high-threshold (1HT) model. The 1HT model can be seen as a special case of the 2HT model based on the assumption that $D_{\mathrm{n}} = 0$, making the model effectively a two-state model without the possibility to detect lures. Notably, the 1HT model is known to be at odds not only with ranking data but also with old–new recognition data (see Kinchla, 1994).

This leads us to an important point. Although our work contributes to the debate between continuous-strength and discrete-state models by demonstrating that ranking tasks do not enjoy the *experimentum crucis* character ascribed by Kellen and Klauer (2014), our work does not provide a direct comparison of whether SDT or 2HT models describe *K*ART behavior better. Future work should therefore aim at developing different SDT and 2HT variants for the *K*ART paradigm (potentially coupled with alternative tasks like the 2AFC task), fit these to appropriate data, and directly compare their performances.

### *Implications for model adaptations to new experimental paradigms*

What are the implications of our results with respect to using well-known models and applying them to new experimental paradigms? Our work suggests that binary old–new decisions and ranking decisions in recognition memory are not the same (see also Voormann, Spektor, & Klauer, 2021), and we should therefore treat them differently (e.g., our 2HT-*K*ART model makes no prediction for the case of $K = 1$). In particular, evidence obtained by the new 2ART-*K*ART model suggests that test items need to be modeled as components of a test display when they are presented and processed simultaneously with other test items. Put differently, the complete vector of (target and lure) detection and non-detection states needs to be considered for each item tuple, instead of modeling each item in isolation. This account is compatible with Luce's (1963) LT model and Province and Rouder's (2012) 2HT model for the standard 2AFC task. In the latter case, an item pair is either in a state of left-item detection, right-item detection, or non-detection. Because the distribution of responses from a cognitive state does not depend on target strength, conditional independence holds. However, the probability of entering each state does depend on target strength: The higher the target strength is, the more likely it is that the item pair will enter a state of target detection (i.e., only the target is detected, or the target and the lure are detected) or a state of only-lure detection (i.e., only the lure is detected).

As our work highlights the importance of the theoretical treatment of auxiliary assumptions, future work should carefully account for different auxiliary assumptions for threshold theories. At the same time, different 2HT model variants will have to share the same core assumptions (e.g., complete information loss, high-thresholds). This may not necessarily allow for a straightforward model hierarchy (e.g., high-threshold placement on the item level versus the tuple level). However, it shows that recent calls for combined modeling approaches in which one model is applicable to different paradigms (e.g., Jang, Wixted, & Huber, 2009; Kellen, Klauer, & Singmann, 2012) need to bear in mind that task-specific mechanisms may require additional model components in order to handle experimental data (e.g., Kellen et al., 2021; McAdoo, Key, & Gronlund, 2019).

### *Implications for the continuous–discrete modeling debate*

What are the implications of our results with respect to the continuous–discrete modeling debate of recognition-memory data? Our work shows that using the new variants of continuous-strength and discrete-state models, and carefully applying them to new paradigms, will allow discriminating between them more sharply. For example, Kellen and Klauer (2015) proposed an additional qualitative criterion for recognition-confidence ratings based on weak versus strong targets. They showed that the 2HT model meets this test criterion, whereas the SDT model does not. In contrast, by modeling confidence ratings and response times jointly, Starns (2021) provided evidence against 2HT and 1LT theorythat can be accounted for by SDT and 2LT theory. The latter is instantiated by a three-state recognition model similar in structure to the 2HT model, but with two low-thresholds rather than two high-thresholds (see also Starns, Dubé, & Frelinger, 2018; Voormann, Rothe-Wulf, Starns, & Klauer, 2020).

Even more recently, Voormann et al. (2021) showed that the 2HT model is more parsimonious in describing 2AFC data than the SDT model. In contrast, using a 2AFC test with pairs of targets and lures previously classified as old, Ma, Starns, and Kellen (2021) showed that the dual-process model (Yonelinas, 1994) outperforms both the SDT and the 2HT model. Most recently, using another critical distribution-free test, Meyer-Grant and Klauer (2021) showed that an SDT model with monotonic rank order probabilities outperforms an 2HT model in a simultaneous detection and identification task. Thus, when combining the conclusions of recent research as well as our present work, we may conclude that both model classes have gained support and suffered losses in different paradigms. If recognition memory is considered to be context-dependent to the extent that different tasks require either continuous-strength or discrete-state recognition (e.g., Malejka & Bröder, 2019; McAdoo & Gronlund, 2019; McAdoo, Key, & Gronlund, 2018), which recognition strategy is applied may well depend on the task at hand and how it can be solved in the most efficient way (Malmberg, 2008; Stevens, 1961).

### *Closing remark*

Although we questioned one crucial auxiliary assumption required for Kellen and Klauer's (2014) prediction and interpretation of their results, we definitely subscribe to their closing remark. The theoretical and empirical concerns regarding ROC analysis call for alternative approaches to study recognition memory and to compare rival measurement models. Testing core properties of the rival models in modified recognition paradigms or in recognition paradigms extended by other judgments from memory (such as confidence ratings, response times, or source-memory judgments) are elegant and more informative alternatives. While old–new recognition judgments provide only a limited, often noisy database and require a fairly large number of auxiliary assumptions, novel paradigms for testing recognition-memory models may provide critical tests with fewer and weaker assumptions. Furthermore, the simple qualitative, but yet very specific predictions derived from the models for these paradigms allow direct model comparisons based on standard hypothesis tests rather than complex model-selection techniques. However, when testing established measurement models in novel recognition-memory paradigms, the competing models need to be adapted carefully to the specific design and structure of the task. Of course, continuous-strength and discrete-state theories of recognition memory are supposed to account for data across as many experimental paradigms as possible. Yet it is important to consider the paradigm-specific auxiliary assumptions that enter into the derivation of precise, testable measurement models from these general theories.

## Funding agency

This research was supported by the University of Mannheim's Graduate School of Economic and Social Sciences (GESS) and the research training group Statistical Modeling in Psychology (SMiP; grant GRK 2277) funded by the German Research Foundation (DFG).

## CRediT authorship contribution statement

**Simone Malejka:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Daniel W. Heck:** Conceptualization, Formal analysis, Methodology, Software, Writing – review & editing. **Edgar Erdfelder:** Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors would like to thank Almut Baronius, Caroline Böhm, Nico Lindheimer, Sophia Kremer, and Silja Wübbelmann for their assistance with data collection, as well as Christoph Klauer, André Aßfalg, Chad Dubé, and Jeff Starns for their valuable feedback on an earlier version of this article.

## Appendix A. Familiarity-contrast model for target and lure detection in *K*-alternative ranking tasks

In this appendix, we outline a stochastic process model of familiarity contrasts that may underlie detection of targets and lures in *K*-alternative ranking tasks (*K*ARTs). The model provides an explanation of why lure-detection probabilities will generally increase with the strength of the target in the same display (i.e., $D_\text{n}^\text{w} \leq D_\text{n}^\text{s}$), irrespective of whether the target item itself is detected or not. In other words, even when the target remains undetected, its strength may affect detection of lures in the same test display. To show this, we assume that attention switches from one item to the next in the *K*ART display while performing the ranking task. In each of the *K* steps, an *independent* automatic process (described below) takes place, which determines whether the attended item (target vs. lure) will enter the detection or the non-detection state. Based on the vector of detection or non-detection states determined in these *K* steps, ranking judgments are derived as defined in the 2HT-*K*ART model illustrated in Fig. 2A of the main text.

More specifically, we assume that the detection probability for each attended item $k$ ($k = 1, \cdots, K$) in a *K*ART display—$D_\text{o}$ when $k$ is the target and $D_\text{n}$ when $k$ is a lure—is determined by a familiarity contrast $\psi(x_k)$ between item $k$ and its context (i.e., the other $K - 1$ items in the display):

$$\psi(x_k) = x_k - \sum_{j \neq k} w_j \cdot x_j, \tag{A1}$$

where $x_k$ denotes the latent familiarity value of item $k$ sampled from its underlying familiarity distribution, which has density $f_\text{o}(\cdot)$ when $k$ is the target and density $f_\text{n}(\cdot)$ otherwise. Correspondingly, $x_j$ denotes the latent familiarity value of item $j$ ($j \neq k$). Finally, $w_j \in [0, 1]$ is a weight with the constraint that $\sum_{j \neq k} w_j = 1$. For simplicity, we may assume that all weights are equal, $w_j = \frac{1}{K-1}$, in which case $\sum_{j \neq k} w_j \cdot x_j$ reduces to the arithmetic mean of the context items' familiarity values (for a similar idea of averaging familiarity values in the context of SDT, see Wixted, Vul, Mickes, & Wilson, 2018).

Once the familiarity contrast has been determined for item $k$, attention switches to a new item $k'$ in the same display, automatically resulting in a new familiarity contrast $\psi(x_{k'})$. This is repeated until all $K$ items were attended. In other words, joint processing of the $K$ items in a *K*ART display results in $K$ item-specific contrasts—each measuring the mean familiarity difference between an attended item and its context. Importantly, when switching attention to a different item $k'$ ($k' \neq k$) in the same display for which a detection state has not yet been determined, we assume that its familiarity contrast $\psi(x_{k'})$ is based on a vector of familiarities sampled *independently* from all $K$ familiarity distributions.

Hence, to obtain the familiarity contrast of the target in a 4ART display, four values are sampled from each of the familiarity distributions underlying the four test items. To obtain the familiarity contrast of any lure in the same test display, four *independent* values are sampled, ensuring conditional independence of familiarity values sampled for different attended items. In addition, when an attended item occurs repeatedly within the same test trial (e.g., in the forced-choice-then-ranking paradigm in Experiment 3), it is not necessary to resample any familiarity values because the item's memory state has already been determined. In such a case, the memory state of the repeated item is maintained, that is, detected items remain in the detection state and non-detected items in the uncertainty state.

When test items are selected at random from the same item pool (as was the case in both Kellen & Klauer, 2014, and all of our own experiments), all lures are characterized by the same familiarity distribution. Thus, when target strength is manipulated between two levels (weak vs strong), only three types of familiarity distributions need to be distinguished—one for strong targets, $f_\text{o}^\text{s}(\cdot)$, one for weak targets, $f_\text{o}^\text{w}(\cdot)$, and one for new items, $f_\text{n}(\cdot)$, with means in descending order. This implies four possible contrast types in *K*ART displays: (1) a strong attended target is compared with $K - 1$ lures, (2) a weak attended target is compared with $K - 1$ lures, (3) an attended lure is compared with a strong target and $K - 2$ lures, and (4) an attended lure is compared with a weak target and $K - 2$ lures. These four contrast types determine the possible distributions of the familiarity contrasts as well as the corresponding detection probabilities—$D_o^s$, $D_o^w$, $D_n^s$, and $D_n^w$, respectively. The detection probabilities, $D_n$ and $D_o$, are defined as follows:

$$D_n = P(\psi(x_n) \leq h_l), \tag{A2}$$

$$D_o = P(\psi(x_o) > h_u). \tag{A3}$$

In these equations, $h_l$ and $h_u$ denote lower and upper thresholds on the contrast dimension, respectively, such that item $k$ in a $K$ART display enters the detection state when either $\psi(x_k) \leq h_l$ or $\psi(x_k) > h_u$ (with $h_l \leq h_u$). In line with 2HT theory, both thresholds are assumed to be "high" in the sense that only familiarity contrasts of targets may exceed $h_u$ and only familiarity contrasts of lures may fall below $h_l$. Assuming contrasts based on equal weights $w_j = \frac{1}{K-1}$ as specified above, this implies $D_o^s \geq D_o^w$ (because contrast values are likely to be larger for strong than for weak targets) and $D_n^s \geq D_n^w$ (because contrast values are likely to be smaller, or more negative, for lures when their context includes a strong rather than a weak target). Attended items with $h_l \leq \psi(x_k) \leq h_u$ are in the uncertainty state, in which pure guessing applies. Based on this assignment of attended items to memory states as implied by their familiarity contrasts, ranking responses are then probabilistically derived from the 2HT-$K$ART model in Fig. 2A of the main text.

One might argue that the familiarity-contrast model proposed here resembles SDT as it assumes continuous familiarity distributions on the latent level. We do not object against such a view, which is admissible for basically any threshold model (e.g., Macmillan & Creelman, 2005; Malejka & Bröder, 2019). Note, however, one important difference between the proposed 2HT-$K$ART process model and the SDT model: According to the former, discrete memory states fully mediate between the latent continuous level and observable responses in the $K$ART. That is, only a finite number of memory states matter in determining response probabilities (as illustrated in Fig. 2A of the main text), whereas familiarity contrasts only serve the purpose to determine detection probabilities. Once items are assigned to detection and non-detection states, the contrast values that led to this assignment become completely irrelevant.

However, in contrast to the 2HT model for binary old–new recognition, our 2HT model for the ranking task discretizes memory strength at the level of the item tuples and not at the level of the individual items within the tuple. 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 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.

In a nutshell, we suggest that ranking judgments in $K$ARTs depends on multiple familiarity contrasts among all items in the test display. Moreover, lure detection is facilitated when the lures are presented in the context of a strong target as compared to a weak target. This is because the contrast is more likely to fall below the relevant threshold $h_l$ in the former case, irrespective of whether the target is actually detected in the same display or not. This results in the prediction that $D_n^w < D_n^s$ even when the target has not been detected itself.

## References

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review, 6*(1), 57–86. https://doi.org/10.3758/bf03210812

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 197–215. https://doi.org/10.1037//0278-7393.22.1.197

Blackwell, H. R. (1953). *Psychological thresholds: Experimental studies of methods of measurement*. (Bulletin No. 36). Ann Arbor, MI: University of Michigan, Engineering Research Institute.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*(1), 84–115. https://doi.org/10.1037/a0014351

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

Dubé, C. (2019). Central tendency representation and exemplar matching in visual short-term memory. *Memory & Cognition, 47*(4), 589–602. https://doi.org/10.3758/s13421-019-00900-0

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., 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 & Language, 67*(3), 389–406. https://doi.org/10.1016/j.jml.2012.06.002

Dubé, C., Tong, K., Westfall, H., & Bauer, E. (2019). Ensemble coding of memory strength in recognition test. *Memory & Cognition, 47*(5), 936–953. https://doi.org/10.3758/s13421-019-00912-w

Egan, J.P., (1958). *Recognition memory and the operating characteristic* (Technical Note AFCRC-TN-58-51). Bloomington, IN: Indiana University Hearing and Communication Laboratory.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology, 217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127*(1), 83–97. https://doi.org/10.1037/0096-3445.127.1.83

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

G*Power (2020). *G*Power 3.1 manual* [Software manual]. https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf.

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

Greene, R. L. (1992). *Human memory: Paradigms and paradoxes*. Hillsdale, NJ: Erlbaum.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 801–812. https://doi.org/10.1037/a0023219

Harlow, I. M., & Donaldson, D. I. (2013). Source accuracy data reveal the thresholded nature of human episodic memory. *Psychonomic Bulletin & Review, 20*(2), 318–325. https://doi.org/10.3758/s13423-012-0340-9

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review, 23*, 1440–1465. https://doi.org/10.3758/s13423-016-1025-6

Hu, X. (1991). *Statistical inference program for multinomial binary tree models* [Computer software]. University of California at Irvine.

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). Mahwah, NJ: 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

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quarterly Journal of Experimental Psychology, 65*(5), 962–975. https://doi.org/10.1080/17470218.2011.638079

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Kellen, D., Erdfelder, E., Malmberg, K. J., Dubé, C., & Criss, A. (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. (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., 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

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101*(1), 166–171. https://doi.org/10.1037/0033-295x.101.1.166

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 680–703. https://doi.org/10.1037/0278-7393.33.4.680

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the "Who said what" paradigm. *Journal of Personality and Social Psychology, 75*(5), 1155–1178. https://doi.org/10.1037/0022-3514.75.5.1155

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76*(3), 308–324. https://doi.org/10.1037/h0027238

Lahl, O., Göritz, A. S., Pietrowsky, R., & Rosenberg, J. (2009). Using the world-wide web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods, 41*(1), 13–19. https://doi.org/10.3758/brm.41.1.13

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556–564. https://doi.org/10.1037/0021-9010.70.3.556

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. (2021). 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*. Advance online publication. https://doi.org/10.1037/xlm0001107.

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

Malejka, S., & Bröder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & Cognition, 44*(1), 63–72. https://doi.org/10.3758/s13421-015-0549-8

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

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

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*(4), 335–384. https://doi.org/10.1016/j.cogpsych.2008.02.004

McAdoo, R. M., & Gronlund, S. D. (2019). 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. (2018). Stimulus effects and the mediation of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(11), 1814–1823. https://doi.org/10.1037/xlm0000550

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

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

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Plummer, M. (2017). *JAGS version 4.3.0* [Computer software manual]. http://mcmc-jags.sourceforge.net/.

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

R Core Team (2019). *R: A language and environment for statistical computing* [Computer software manual]. https://www.R-project.org/.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 31*(4), 696–700. https://doi.org/10.3758/bf03200748

Rowland, C. A. (2014). The effect of testing versus restudy in retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of interval practice at short retention intervals. *Memory, 23*(3), 403–419. https://doi.org/10.1080/09658211.2014.889710

Snodgrass, J. G., & Corwin, J. L. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 30–50. https://doi.org/10.1037//0096-3445.117.1.34

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., 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

Stevens, S. S. (1961). Toward a resolution of the Fechner-Thurstone legacy. *Psychometrika, 26*(1), 35–47. https://doi.org/10.1007/bf02289683

Swagman, A. R., Province, J. M., & Rouder, J. N. (2015). Performance on perceptual word identification is mediated by discrete states. *Psychonomic Bulletin & Review, 22*, 265–273. https://doi.org/10.3758/s13423-014-0670-x

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 582–600. https://doi.org/10.1037/0278-7393.26.3.582

Voormann, A., Rothe-Wulf, A., Starns, J. J., & Klauer, K. C. (2020). Does speed of recognition predict two-alternative forced-choice performance? Replicating and extending Starns, Dubé, and Frelinger (2018). *Quarterly Journal of Experimental Psychology, 74*(1), 122–134. https://doi.org/10.1177/1747021820963033

Voormann, A., Spektor, M. S., & Klauer, K. C. (2021). The simultaneous recognition of multiple words: A process analysis. *Memory & Cognition, 49*(4), 787–802. https://doi.org/10.3758/s13421-020-01082-w

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176. https://doi.org/10.1037/0033-295x.114.1.152

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*(2), 262–276. https://doi.org/10.1037/a0035940

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

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2*(6), 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800