Journal of Mathematical Psychology 55 (2011) 251–266

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Evaluating models of recognition memory using first- and second-choice responses

David Kellen *, Karl Christoph Klauer

Albert-Ludwigs-Universität Freiburg, Germany

ARTICLE INFO

Article history:
Received 6 August 2010
Received in revised form
8 November 2010
Available online 5 January 2011

Keywords:
Recognition memory
Second-choice responses
Signal detection theory
Discrete-state models
Hybrid models

ABSTRACT

Swets, Tanner Jr., and Birdsall (1961) proposed a 4-alternative forced-choice task with two choices (4AFC-2R) for distinguishing between the Equal-Variance Signal Detection model and the One-High Threshold model. This task was recently implemented in the field of recognition memory (Parks & Yonelinas, 2009), a field in which several candidate models have been proposed. One advantage of the 4AFC-2R task is that it permits parameter estimation and goodness of fit testing, something which so far was only possible through the use of Receiver Operating Characteristic (ROC) functions for the more complex candidate models. The present article provides a thorough characterization and comparison of the main recognition memory models in the context of this task. Results are illustrated by a reanalysis of Parks and Yonelinas' original data, revealing a preference for hybrid approaches to recognition memory, more specifically for the dual-process model (Yonelinas, 1997), whereas pure signal detection models performed poorly. The present analysis provides an assessment of the merits and limitations of this task, highlighting future research applications.

The ability to recognize previously acquired information is one of the aspects of human memory that has been extensively studied by means of mathematical models (for a review, see Malmberg, 2008). In the past decades several models have been proposed and discarded, each with distinct assumptions and predictions: Some approaches assume that mnesic information is adequately represented as an all-or-none process (e.g., Batchelder & Riefer, 1990; Klauer & Kellen, 2010), while others postulate a continuous representation of mnesic evidence (e.g., Wixted, 2007) or a combination of both possibilities (e.g., DeCarlo, 2002; Yonelinas, 1997). The major models, represented in Fig. 1, can thus be divided into three classes: discrete-state, continuous, and hybrid models.

The Signal Detection model (SDT; Macmillan & Creelman, 2005) assumes that a single, continuous mnesic process, termed familiarity, describes the individuals' decisions based on mnesic information. Given that all items – studied and non-studied – have some degree of familiarity, it is possible to represent them by their respective familiarity distributions, and the ability to discriminate studied items from distractors is defined by the distance between the two distributions, in this case represented by parameter μ. In this framework, an item is declared old or new in a recognition task by comparing its familiarity with an established response criterion, denoted by c. If the item's familiarity value is higher than the response criterion, the item is declared as old, otherwise it is rejected. Within the SDT framework one can distinguish two versions of the model; the Equal-Variance Signal Detection model (EVSDT), which assumes that signal and noise distributions have the same standard deviation, and the Unequal-Variance Signal Detection model (UVSDT), which is a more general version that allows for the signal distribution standard deviation, indexed by σ, to assume a different value from the distractor distribution standard deviation.

The One-High Threshold model (1HTM Blackwell, 1963) assumes that decisions based on mnesic information can be described by two discrete states, a "remember" state and a "guessing" state. Remembering an item during test is a probabilistic event given that only a portion of the studied items will surpass a specific memory threshold, the portion being defined by parameter D_o. The items that are not remembered (which occurs for all distractors) trigger a guessing process defined by parameter g. An extension of this discrete-state approach is the Two-High Threshold model (2HTM Snodgrass & Corwin, 1988) that includes an additional state of "distractor detection" indexed by parameter D_n. This parameter attempts to describe the active rejection of distractors, a phenomenon usually associated with a host of metacognitive strategies (e.g., Strack & Bless, 1994).

The Dual-Process Model (DPSDT Mandler, 1980; Yonelinas, 1997) is a hybrid approach that combines aspects of the previous two model classes, adopting a continuous familiarity process (equivalent to EVSDT) to describe more vague mnesic states, and an additional threshold component that describes episodic remembrance, termed recollection and defined by parameter R. A second hybrid approach is the Finite Mixture Signal Detection Model (MSDT DeCarlo, 2002, 2010), according to which the signal

* Corresponding address: Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany.
E-mail address: david.kellen@psychologie.uni-freiburg.de (D. Kellen).

0022-2496/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2010.11.004

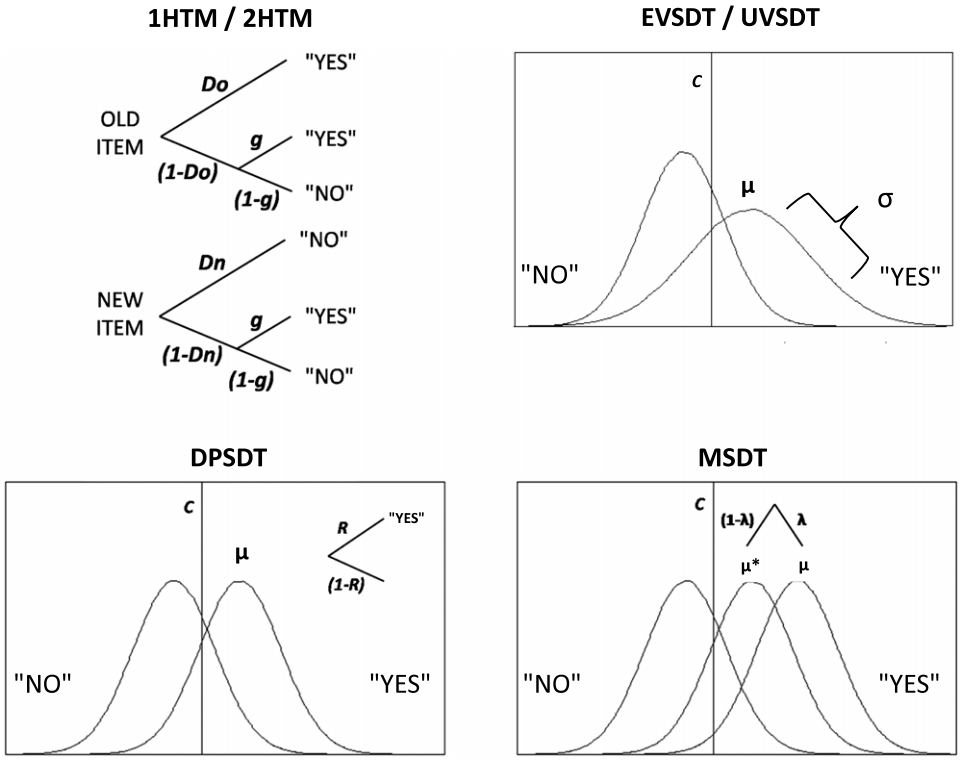

Fig. 1. Representation of the main recognition memory models.

distribution is comprised of a mixture of two latent equal-variance normal distributions—one corresponding to items that were attended during study, with mean μ, and a second distribution for unattended items with mean μ^* and with $\mu^* \leq \mu$. The proportion of attended items among studied items is defined by parameter λ.

Traditionally, recognition memory models are tested by means of their predictions regarding Receiver Operating Characteristic (ROC) functions. ROCs map hits and false alarm probabilities across several levels of response bias. ROCs are informative given that the shape of the expected function varies greatly depending on the assumed processes and underlying probability distributions (see Swets, 1986). For instance, both discrete-state models predict linear ROCs, while SDT predicts curvilinear ones. Both hybrid models can produce intermediate shapes as well as more complex ones (e.g., DeCarlo, 2002).

Although ROC functions are defined as functions obtained through the use of direct bias manipulations, either in terms of changes in item base rates or in the payoff matrix (see Van Zandt, 2000), they are almost always obtained through the use of confidence-rating scales, which are then compiled in order to emulate a bias manipulation (for exceptions, see the review by Bröder & Schütz, 2009). Despite their widespread use in the literature, ROC functions based on confidence ratings are in fact a relatively non-informative method for testing models and theories, given that the candidate models are often indistinguishable by means of the function's shape, that is, the models make similar predictions when extended to the confidence-rating response format appropriately (Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Klauer & Kellen, 2010; Krantz, 1969; Lockhart & Murdock, 1970; Malmberg, 2002). In order to overcome such limitations, Bröder and Schütz (2009) collected ROCs by manipulating test item base rates, and obtained linear-shaped functions, in contrast to the almost ubiquitous nonlinear shape that has been observed in the past decades and that led to the (possibly premature) dismissal of pure discrete-state approaches (Wixted, 2007). These results make it desirable to explore alternative methods to distinguish between these models. Another advantage of having alternative methods is to avoid what has been called mono-operation bias (Shadish, Cook, & Campbell, 2002, Chap. 3), a bias that is incurred when conclusions rely excessively on a single method, in this case confidence-rating ROCs (see Ratcliff & Starns, 2009).

1. Beyond the ROC: model selection by means of first- and second-choice accuracies

Alternative methods were already considered in the early developments of SDT (Green & Swets, 1966); one such method is based on the predictions that models make for first- and second-choice responses in multiple alternative forced-choice paradigms (Swets et al., 1961). As previously described, according to SDT, responses are based on the test item familiarity or evidence value and the comparison of this value with a previously established response criterion. In the case of multiple alternatives, responses are not based on the comparison of a stimulus with an established response criterion, but on the direct comparison of the alternatives presented at test and their familiarity values. The assumption is that the alternative with the highest evidence value is chosen, and this difference in the nature of the comparison eliminates the need for parameter c. Given that responses are based on these comparisons, individuals order their preferences according to the associated evidence values. If performance is above chance level (that is, on average, correct alternatives have a higher evidence value than incorrect alternatives), two simple predictions result for EVSDT: first, the accuracy of second-choice responses, after an incorrect first choice, must be above chance level. Second, the accuracy of first and second choices should be correlated (see section "Models for First- and Second-Choice Accuracy in 4AFC-2R" for formal derivations of those predictions).

These predictions contrast with the ones that were made by 1HTM; according to this model, on each trial, participants either detect the correct alternative, or else they simply guess. Thus, incorrect first choices occur because the correct alternative did not

Table 1
Means and standard deviations (in parenthesis) for the proportions of correct scores from Parks and Yonelinas' (2009) experiments.

Condition	Single choice	First choice	Second choice	Correlation between single and second choice
Item	0.53 (0.15)	0.51 (0.17)	0.42 (0.10)	0.56*
Pair	0.51 (0.15)	0.53 (0.17)	0.38 (0.10)	−0.07
Sentence	0.57 (0.19)	0.55 (0.17)	0.36 (0.19)	0.13
Compound	0.53 (0.17)	0.55 (0.16)	0.39 (0.10)	0.39*

* $p < 0.05$.

surpass the detection threshold. Given this sub-threshold status of the correct alternative, there is no evidence available to distinguish between alternatives, which should result in chance level accuracy for second-choice responses. In consequence, there should be no correlation between first- and second-choice accuracies; the latter should always be at chance level.

The predictions made by both models were originally tested with a visual perception task by Swets et al. (1961), using a 4-alternative forced-choice task with two choices (4AFC-2R). The pattern of results obtained showed that second-choice accuracy (conditional on incorrect first choice) is above the 33% correct response proportion that a chance level performance would produce. Also, a positive correlation was found between the two choices, which favors even further the Signal Detection model in comparison to the One-High Threshold model. The results presented by Swets et al. (1961) were later on replicated in many other studies using several perception modalities (see Green & Swets, 1966, Chap. 4), generalizing the findings, and establishing this paradigm as an alternative route for comparing SDT with other model approaches.

Nevertheless, the fact that the 1HTM model cannot account for the above chance performance in second choices does not represent a rejection of the discrete-state approach in general. The 2HTM model can naturally account for such data patterns: When the correct alternative is not detected, the distractor detection parameter D_n can still reduce the total number of alternatives considered (see García-Pérez, 1990). By not considering incorrect alternatives that are detected as wrong, the "guessing" performance of the 2HTM is expected to be above chance. As shown later, positive correlations between first-choice correct and second-choice correct (conditional on first-choice incorrect) are also consistent with the 2HTM model.

2. Testing recognition memory models with the 4AFC-2R task

Recently, Parks and Yonelinas (2009) used a 4AFC-2R recognition memory task in order to distinguish between different candidate models, namely the UVSDT and the DPSDT. In addition to the 4AFC-2R trials, participants responded to a single-choice 4AFC condition, in order to control for possible differences in accuracy caused by the inclusion of a second choice. In two experiments, Parks and Yonelinas analyze the differences between single-item recognition and pair recognition, both in terms of the accuracy of second-choice responses, and its relationship with first-choice accuracy. The differences between item recognition and pair recognition are extremely important for the discrimination between these two models, given that both SDT in general and the DPSDT make very specific predictions. SDT models assume that a familiarity process underlies performance in the item and the pair recognition tasks, and therefore predicts that in both tasks, performance should conform to a signal detection process, implying the presence of a positive correlation between the accuracies of first- and second-choice responses (at least for EVSDT). For the DPSDT model, performance in item recognition is based on both recollection and familiarity processes, and therefore DPSDT was argued to predict a correlation between both responses that falls between the predictions of EVSDT and the 1HTM. In contrast, on the pair recognition task, an accurate discrimination of correct word pairs should be based only on the recollection process, given that rearranged word pairs (the distractors in this task) are assumed to be equally familiar as intact pairs, and therefore the characteristics of first- and second-choice responses should conform to those of 1HTM.

The results obtained by Parks and Yonelinas (2009) are summarized in Table 1 and reveal interesting differences between single-item and pair recognition: In Experiment 1, there is a positive correlation between first- and second-choice responses in the Item Memory Condition, but there is no correlation in the Pair Memory Condition. The results suggest the existence of a familiarity-based process underlying performance in single-item recognition, while a threshold-like recollection process seems to underlie the pair recognition data.

In a second experiment, the focus was on pair recognition, more specifically on the effects of encoding on subsequent first- and second-choice performances. The encoding manipulation either encouraged an individual encoding of paired items (Sentence Condition; evaluating each word separately) or a compound encoding of the word pair (Compound Condition; evaluating the pair as a whole). The use of this kind of encoding manipulation has previously produced results (e.g., Diana, Yonelinas, & Ranganath, 2008) suggesting that a more compound processing of items tends to "unitize" the encoding item and source information, and subsequently allows for a central role of familiarity in source discrimination. In the case of pair recognition, "unitization" should lead to familiarity having a greater role in pair recognition, thus leading to a positive correlation between first-choice and second-choice responses, as expected on the basis of an EVSDT-like familiarity process. The results obtained were consonant with the authors' predictions, revealing a positive correlation between the two choices when the pairs were studied in a "unitized" fashion, while no correlation was present when participants studied the items in a more individualized way.

The introduction of 4AFC-2R as a method to distinguish models of recognition memory represents an important contribution to the research in this field, with the potential of changing the way the candidate models are usually tested. Given all the vulnerabilities and limitations of ROCs that have been discussed in the recent literature (Benjamin, Diaz, & Wee, 2009; Bröder & Schütz, 2009; Klauer & Kellen, 2010; Malmberg, 2002; Mueller & Weidemann, 2008; Ratcliff & Starns, 2009; Rosner & Kochanski, 2009), this is a welcome addition to the existing test strategies. But, despite its merits, the study of Parks and Yonelinas (2009) does not provide a full assessment of this method for model testing, and reveals some limitations in different aspects. One of them concerns the fact that the 4AFC-2R task has only been validated as a means to qualitatively distinguish between EVSDT and 1HTM in the previous work; its ability to distinguish between more complex models such as the ones reviewed above has not been comprehensively evaluated so far.[1]

[1] Parks and Yonelinas (2009) implemented experimental controls designed to eliminate the so-called "recall-to-reject" strategy (e.g., Rotello & Heit, 2000) for distractor detection, which led them to remove the 2HTM model from consideration. But recall-to-reject is only one of several strategies underlying distractor detection parameter D_n, other possibilities being rejection on the basis of incongruency (Higham & Brooks, 1997), idiosyncratic associations and metacognitive inferences based on them (Strack & Bless, 1994), among others.

A second point is related to the way models are approached: Whenever possible, models should be tested and compared in terms of their ability to account for the observed frequency data in addition to their ability to account for some summary statistic such as the correlation between first and second choices. Even if one model predicts a strong correlation between the responses, and the other does not, the presence of a correlation in the data does not represent unequivocal evidence in favor of the first model. Despite the correlation, the likelihood that the data came from the second model can still be large, and even greater than the likelihood for the first. Also, the individual parameter values obtained in actual fittings of the models might turn out to be implausible according to the parameters' theoretical interpretations in terms of underlying psychological processes, providing useful information with the potential to discriminate between models. Overall, we argue that in order to assess the adequacy of different candidate models, goodness of fit tests should also be implemented in conjunction with some of the wide range of methods for model assessment and selection that are available (Myung & Pitt, 2004).

3. Models for first- and second-choice accuracies in 4AFC-2R

The fact that mathematical models are formally and unambiguously represented is one of their many advantages in comparison to verbal theories, and the present case is no exception, allowing for quantitative and qualitative assessments of the candidate models. The 4AFC-2R paradigm provides data for three response categories: (1) first-choice correct and second-choice incorrect; (2) first-choice incorrect and second-choice correct; (3) first-choice incorrect and second-choice incorrect.

Note that Parks and Yonelinas (2009) also had a condition with only one choice, adding the two response categories (4) single-choice correct, and (5) single-choice incorrect. Because of the (empirically supported) assumption that the probabilities of single-choice correct (Category 4) and first-choice correct (Category 1) are equal, these response categories do not add structurally new information.

In this section, we introduce the candidate models considered in the literature for the 4AFC-2R task. We present the model equations and then characterize each model in terms of the probabilities of first-choice correct (Category 1) and second-choice correct (Category 2) that they can account for. Because the probabilities for Categories 1–3 have to sum to 1, characterizing models in terms of probabilities of Categories 1 and 2 is sufficient. The characterization will lead to insights into the relationships between these models. The candidate models are UVSDT, EVSDT, 2HTM, 1HTM, DPSDT, and MSDT.

Let π_i be a generic term for the unconditional probability of ith choice being correct, $i = 1, \ldots, 4$, for any of the models considered, noting that these probabilities must sum to one. Furthermore, let $F_{(\mu,\sigma)}$ and $f_{(\mu,\sigma)}$ be the distribution function and probability density, respectively, of the normal distribution with mean μ and standard deviation σ, with F and f being these functions for the standard normal distribution (i.e., $F = F_{(0,1)}$ and $f = f_{(0,1)}$).

3.1. UVSDT and EVSDT

The UVSDT model assumes that the familiarity values of the studied items and distractors are both normally distributed, with the respective probability density functions $f_{(\mu,\sigma)}$ and $f_{(0,1)}$. While the familiarity distribution for the distractors has, without loss of generality, a fixed mean of 0 and a standard deviation of 1, the mean (μ) and standard deviation (σ) of the signal distribution are free parameters to be estimated.

In the UVSDT model, π_1 is defined as the probability that the studied item has a larger familiarity value than the three incorrect alternatives as defined by Eq. (1) (for more details, see Wickens, 2002, Chap. 6), while π_2 corresponds to the probability that the studied item will have a larger familiarity than *just two* of the incorrect alternatives:

$$\pi_1(\mu, \sigma) = \int_{-\infty}^{\infty} F^3(x) f_{(\mu,\sigma)}(x)\, dx \tag{1}$$

$$\pi_2(\mu, \sigma) = 3\int_{-\infty}^{\infty} F^2(x)(1 - F(x)) f_{(\mu,\sigma)}(x)\, dx \tag{2}$$

with $\mu \geq 0$ and $\sigma > 0$.

We characterize each model in terms of the space that they cover in the set of probabilities (π_1, π_2) of first-choice correct and second-choice correct. For UVSDT, the model space is given in Theorem 1.

Theorem 1. *The set of probabilities (π_1, π_2) described by the UVSDT model with $\mu \geq 0$ and $\sigma > 0$ is $\{(\pi_1, \pi_2) : \frac{1}{8} < \pi_1 < 1$ and $\frac{1}{2} - \pi_1 \leq \pi_2 < 3(\pi_1^{\frac{2}{3}} - \pi_1)$ for $\pi_1 < \frac{1}{2}$ and $0 < \pi_2 < 3(\pi_1^{\frac{2}{3}} - \pi_1)$ for $\pi_1 \geq \frac{1}{2}\}$.*

The proofs of all theorems, as well of the identifiability of the candidate models, can be found in the Appendix. The model space is depicted in Fig. 2. In Fig. 2(a), the values on the horizontal and vertical axes are the probabilities π_1 and π_2, respectively; in Fig. 2(b), they are π_1 and π_2^c, respectively, with π_2^c being the conditional probability of second-choice correct, given first-choice incorrect (i.e. $\pi_2^c = \frac{\pi_2}{1-\pi_1}$).

As can be seen, the model space is enclosed by a lower bound and an upper bound. The UVSDT is unique among the models considered here in that it is the only model that can account for below chance performance in terms of both π_1 and π_2. For example, π_1 has a lower bound of $\frac{1}{8}$, and for all $\pi_1 \geq \frac{1}{2}$, π_2 (and therefore π_2^c) has a lower bound of 0 (the guessing baseline for π_2^c is $\frac{1}{3}$). Considering Fig. 2(b), it is also seen that the model is consistent with curves through the space that increase as a function of π_1 as π_1 ranges from $\frac{1}{8}$ to 1, implying positive correlations between first-choice correct and second-choice incorrect, as well as with curves that decrease all the way, implying negative correlations, another aspect that is unique for UVSDT.

As previously mentioned, EVSDT is a special case of UVSDT with $\sigma = 1$. As can be seen in Fig. 2, the predictions of EVSDT are restricted to a single curve. Let $e_i(\mu)$, $i = 1, \ldots, 4$, be the unconditional probability of the ith choice correct specifically for the EVSDT model with parameter μ. Also, let the conditional probability of second-choice correct, given an incorrect first choice, under the EVSDT model be given by $c_2(\mu)$, $c_2(\mu) = \frac{e_2(\mu)}{1-e_1(\mu)}$. An important property shared by c_2 and e_1 is that they are strictly increasing in μ (see the Appendix). This is the mathematical foundation for the prediction of a positive correlation between π_1 and $\pi_2^c = c_2$ under EVSDT (Parks & Yonelinas, 2009). Its reflection in Fig. 2(b) is the fact that the curve describing EVSDT is increasing as a function of π_1.

3.2. 2HTM and 1HTM

The 2HTM model assumes that individuals can detect a specific proportion of studied items as previously studied with probability D_o. When such detection does not occur, individuals engage in a guessing process, excluding beforehand any alternative that might have been detected as incorrect (with item-wise probability D_n). Correct second-choice responses can only occur when the correct alternative was not detected, and at least one of the incorrect

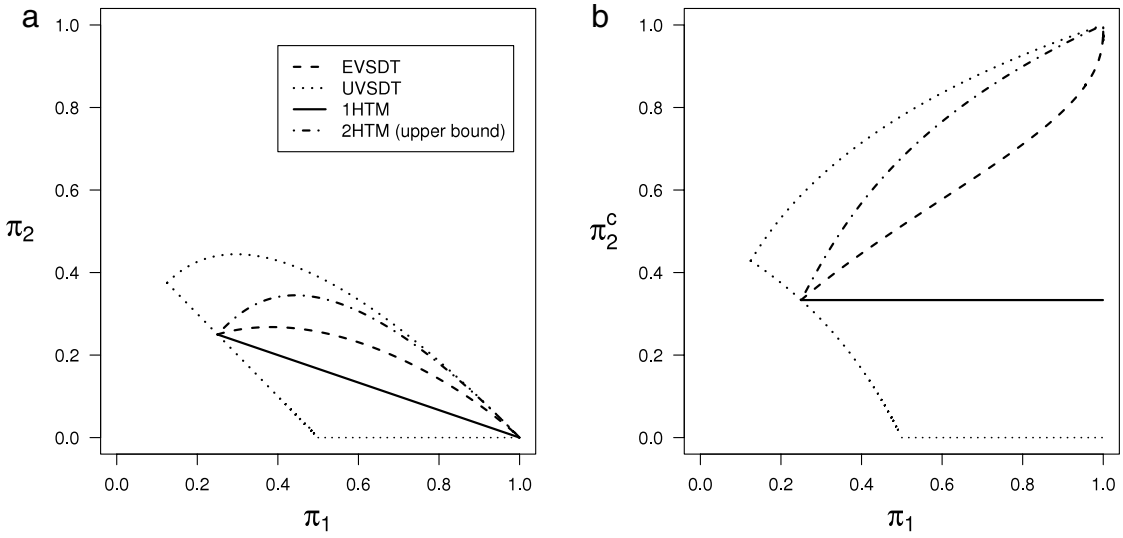

Fig. 2. Predictions of (π_1, π_2) and (π_1, π_2^c) (panels A and B, respectively) for each model across their respective parameter ranges (for details, see the theorems in the Appendix). For EVSDT and 1HTM, the predictions are limited to their respective curves (no areas are defined). The predictions of 2HTM correspond to the area enclosed by the curves labeled 2HTM (upper bound) and 1HTM (lower bound; see Theorem 2). The predictions of both DPSDT and MSDT correspond to the area enclosed by the curves for EVSDT and 1HTM (Theorems 3 and 4). The predictions of the UVSDT model correspond to the area enclosed by the curve labeled UVSDT (Theorem 1).

alternatives was not detected as incorrect as well. This leads to the following model equations:

$$\pi_1(D_o, D_n) = D_o + (1 - D_o)G(D_n) \tag{3}$$

$$\pi_2(D_o, D_n) = (1 - D_o)H(D_n) \tag{4}$$

with $0 \leq D_o, D_n \leq 1$.

$G(D_n)$ and $H(D_n)$ represent the probabilities of arriving at a correct response through distractor detection and/or guessing, conditional on the absence of detection of the correct alternative for first and second choices, respectively.

$$\begin{aligned} G(D_n) &= D_n^3 + \frac{3}{2}D_n^2(1 - D_n) + D_n(1 - D_n)^2 + \frac{1}{4}(1 - D_n)^3 \\ &= \frac{1}{4}\left(1 + D_n + D_n^2 + D_n^3\right) \end{aligned}$$

$$H(D_n) = G(D_n) - D_n^3.$$

It follows that G ranges from $\frac{1}{4}$, the guessing baseline, to 1 and that it is strictly increasing in D_n. In contrast, H is non-monotonous and ranges between $\approx\frac{1}{2.9}$ and 0. The model space for 2HTM is characterized in Theorem 2:

Theorem 2. *The set of probabilities (π_1, π_2) described by the 2HTM model is $\{(\pi_1, \pi_2) : \frac{1}{4} \leq \pi_1 \leq 1$ and $\frac{1}{3}(1 - \pi_1) \leq \pi_2 \leq H(G^{-1}(\pi_1))\}$, where G^{-1} is the inverse of function G. Furthermore, setting $a = \sqrt{3 - 30\pi_1 + 81\pi_1^2}$ and $b = 10 - 54\pi_1 + 6a$, $H(G^{-1}(\pi_1))$ can be written explicitly as $\frac{1}{6}(2 - 18\pi_1 + (-4 + 18\pi_1 + 2a)b^{\frac{1}{3}} + (-1 + 9\pi_1 + a)b^{\frac{2}{3}})$.*

According to Theorem 2, $\pi_1(D_o, D_n)$ has a minimum value of 0.25 (when both D_o and D_n equal 0), whereas π_2 ranges between a guessing baseline of $\frac{1}{3}$ for the conditional probability of second-choice correct, given incorrect first choice (see Fig. 2(b)) and an upper bound that consistently falls below the upper bound for UVSDT. In terms of conditional probability π_2^c, both lower and upper bounds are non-decreasing functions of π_1, suggesting positive correlations between π_1 and π_2^c for the 2HTM, but depending on the trajectory traced by given data in the model space, negative correlations are of course not excluded.

The 1HTM model is a special case of 2HTM with distractor detection parameter D_n equal to 0. This restriction makes both G and H assume the fixed value of $\frac{1}{4}$. The 1HTM line in Fig. 2 thus defines the choice probabilities predicted by the restricted model. As can be seen in Fig. 2(b), a zero correlation is predicted between π_1 and π_2^c, because π_2^c equals $\frac{1}{3}$, irrespective of the value of π_1.

3.3. DPSDT and MSDT

Considering the hybrid models, DPSDT postulates two ways of reaching a correct first response; either by recollection of the studied item (in a process equal to that captured by D_o for 1HTM), quantified by parameter R, or by a familiarity-based process, equivalent to EVSDT. Therefore,

$$\pi_1(\mu, R) = R + (1 - R)e_1(\mu) \tag{5}$$

$$\pi_2(\mu, R) = (1 - R)e_2(\mu) \tag{6}$$

with $\mu \geq 0$ and $0 \leq R \leq 1$.

In contrast, MSDT assumes that the familiarity of the studied items is better described by a mixture of two normal distributions with equal standard deviations, corresponding to items that were studied attentively or not. The probability density functions for these two distributions are $f_{(\mu,1)}$ and $f_{(\mu^*,1)}$, respectively. The proportion of studied items that were studied attentively and are therefore described by the first distribution is denoted by the mixture parameter λ.

Given that MSDT has three parameters to be estimated, the means of both studied item distributions, μ and μ^*, and the mixture parameter λ, the model is overparameterized in the current experimental design, and the parameters are not identified. For identifiability, at least one parameter restriction must be imposed. A common choice is to set the mean of the unattended items' distribution equal to the mean of the distractor distribution (see DeCarlo, 2002, i.e. $\mu^* = 0$). The model equations are as follows:

$$\pi_1(\mu, \mu^*, \lambda) = \lambda e_1(\mu) + (1 - \lambda)e_1(\mu^*) \tag{7}$$

$$\pi_2(\mu, \mu^*, \lambda) = \lambda e_2(\mu) + (1 - \lambda)e_2(\mu^*) \tag{8}$$

with $\mu \geq \mu^* \geq 0$ and $0 \leq \lambda \leq 1$.

The model spaces for DPSDT and MSDT are characterized in Theorems 3 and 4, respectively:

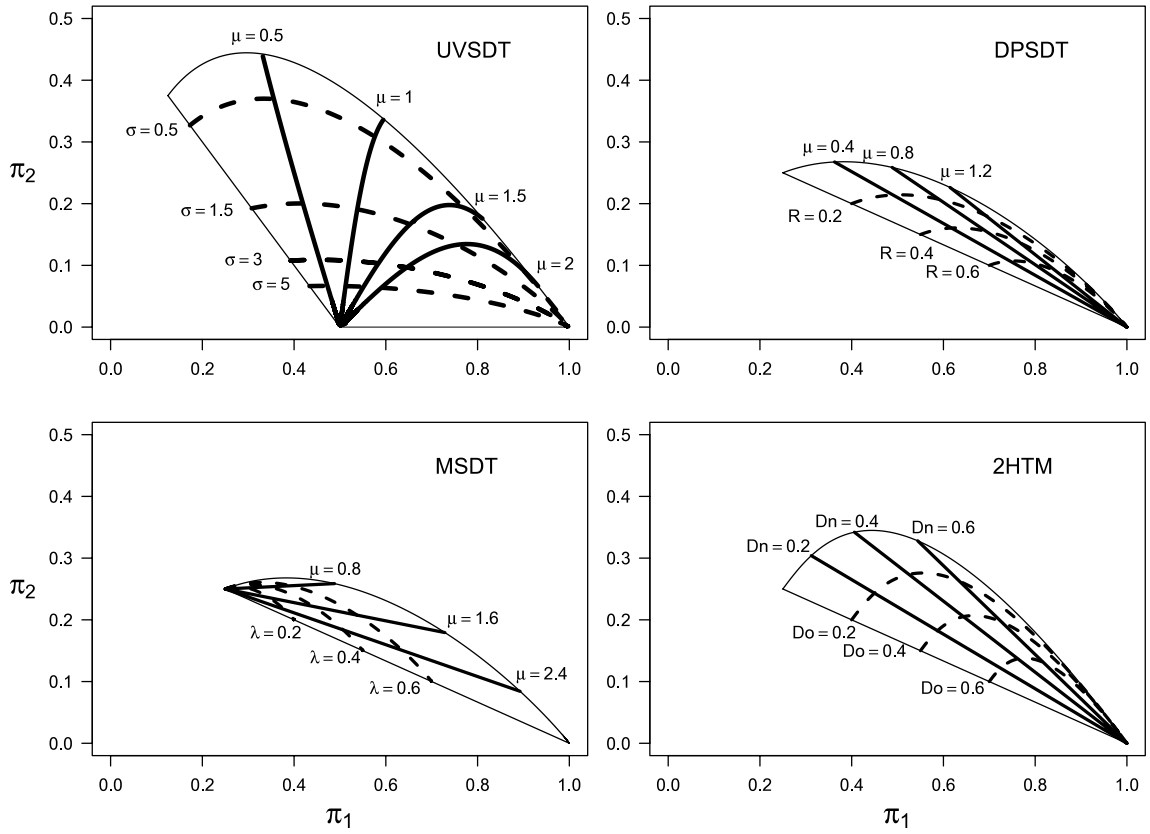

Fig. 3. Predictions for the two parameter models as a function of parameter values. The depicted lines represent changes in (π_1, π_2) with one parameter fixed, and the remaining parameter varying across its range. For example, in UVSDT, the solid line $\mu = 0.5$ corresponds to the predicted (π_1, π_2) values when μ is fixed to 0.5 and σ^2 varies.

Theorem 3. *The set of probabilities (π_1, π_2) described by the DPSDT model with $\mu \geq 0$ is $\{(\pi_1, \pi_2) : \frac{1}{4} \leq \pi_1 \leq 1$ and $\frac{1}{3}(1-\pi_1) \leq \pi_2 \leq e_2(e_1^{-1}(\pi_1))\}$, where e_1^{-1} is the inverse of e_1.*

Theorem 4. *The set of probabilities (π_1, π_2) described by the MSDT model with $\mu \geq \mu^* \geq 0$ for a fixed μ^* is $\{(\pi_1, \pi_2) : e_1(\mu^*) \leq \pi_1 < 1$ and $c_2(\mu^*)(1-\pi_1) < \pi_2 \leq e_2(e_1^{-1}(\pi_1))\} \cup \{(e_1(\mu^*), e_2(\mu^*))\}$.*

Under the identifiability restriction $\mu^* = 0$, $e_1(\mu^*) = \frac{1}{4}$ and $c_2(\mu^*) = \frac{1}{3}$ as is easy to see (see also the Appendix). A surprising outcome is thus that the MSDT model with $\mu^* = 0$ describes the same choice probabilities as the DPSDT model in the 4AFC-2R task, with the exception that DPSDT, but not MSDT, accommodates the extreme cases of $(\pi_1, \pi_2) = \left(\pi_1, \frac{1}{3}(1-\pi_1)\right)$ for $\pi_1 > 0.25$ although MSDT approximates these values arbitrarily closely. Note also that the MSDT model does not describe additional patterns of choice probabilities if μ^* is allowed to vary. This latter observation follows from the fact that $c_2(\mu)$ is non-decreasing in μ (see the Appendix).

For both models, the lower bound for π_2 and π_2^c is the curve predicted by 1HTM, whereas the upper bound is the curve predicted by EVSDT. These bounds point to the fact that both models' predictions are within the boundaries established by the respective predictions of EVSDT and 1HTM, meaning that the DPSDT and MSDT can account for exactly the same data patterns. Thus, MSDT and DPSDT are essentially the same model for the 4AFC-2R task parameterized differently, despite widely disparate underlying psychological assumptions. Note that this surprising case of model mimicry holds even if considering an unrestricted version of MSDT ($\mu^* \geq 0$). One implication is that the models cannot be distinguished in terms of goodness of fit. Nevertheless, given the theoretical interpretation of the parameters of each model in terms of psychological processes, their estimated values and relationships have diagnostic value and may still support a preference for one model over the other, as exemplified later.

By implementing selective parameter restrictions, the DPSDT and MSDT produce equivalent nested models. More specifically, the DPSDT with $\mu = 0$ becomes 1HTM, and with $R = 0$, it is EVSDT. The MSDT reduces to EVSDT for $\lambda = 1$, and to 1HTM as $\mu \to +\infty$.

Inspecting Fig. 2(b), it is seen that both lower and upper bounds for π_2^c are non-decreasing functions of π_1 for DPSDT and MSDT, suggesting positive correlations between π_1 and π_2^c, but depending on the trajectory realized by given data in the model space, negative correlations are of course not excluded.

Besides the models' areas of predicted (π_1, π_2) and (π_1, π_2^c) values, it is also important to consider how these response probabilities are parameterized by each model. Fig. 3 depicts the predicted probabilities as a function of specific parameters. For instance, although the DPSDT and MSDT models make the same predictions, the parametrization of these predictions is quite distinct. These differences will ultimately be reflected in the way these models account for different aspects of the data, such as the presence (or absence) of a correlation between π_1 and π_2^c as elaborated on later.

3.4. Relationships among the models

From Fig. 2 it is clear that the model space of UVSDT is not only larger than those of the other models, but all other models are included in that space. This means that UVSDT has a tremendous advantage in terms of goodness of fit, as it can account for all the possible predictions made by its competitors while making additional predictions. In addition, the models are seen to be ordered as follows: UVSDT is a supermodel of 2HTM, which in turn

contains DPSDT and MSDT as submodels, with DPSDT and MSDT being different parameterizations of basically the same model. 1HTM and EVSDT are boundary cases of DPSDT and MSDT. These orderings will be reflected in the measures of model complexity described in the next section.

Considering the experimental design implemented by Parks and Yonelinas (2009) and their data, the ability of each of the candidate models to account for the data hinges on two aspects: (1) the differences between single-choice and first-choice accuracy, and (2) the different joint values of first-choice and second-choice accuracy that each model can predict. Regarding (1), remember that Parks and Yonelinas implemented a condition with only one choice along with the 4AFC-2R condition for each participant. The assumption is that the probability of first-choice correct in the 4AFC-2R part of the experiment equals that of single-choice correct, providing an important check for the consistency of the results from the 4AFC-2R paradigm with a more traditional condition. Because the single-choice probability is predicted to be the same as the first-choice probability, the assumption basically leads to the prediction that observed relative frequencies should be the same for both single-choice correct and first-choice correct, irrespective of the candidate model considered. While this first potential source of misfit is thus bound to punish each model equally, not providing any diagnostic evidence, the second one is the sole basis for discriminating between models.[2] What is diagnostic for model comparison are the data from the two-choice condition alone, but as can be seen in Fig. 2, all models show considerable overlap in the range of first- and second-choice correct probabilities (π_1 and π_2) that they can account for. This renders taking into account model flexibility over and above goodness of fit imperative for the purpose of data-driven model selection.

Also important is the observed relationship between parameter estimates in each model, as they should reflect the theoretical notions that are associated to them. Given that the models parametrize choice probabilities differently (see Fig. 3), their account of distinct data patterns should lead to differences in the observed relationship between parameters estimates as well.

4. Assessing model complexity by means of minimum description length

One of the fundamental issues in model selection is the fact that more complex models tend to provide better fits to data in general, without necessarily providing an adequate description of the underlying processes. Given the goal of selecting the model that best describes the data, model complexity is an aspect that should always be taken into account. However, four out of the six models considered have the same number of parameters, which means that more common model selection measures that rely on the number of parameters to assess model complexity, like the Akaike Information Criterion (AIC Akaike, 1973) or the Bayesian Information Criterion (BIC Schwarz, 1978), have no informative value. Also, the theorems presented reveal large differences in the range of choice probabilities predicted by each model, a flexibility that should be quantified. This situation emphasizes the need to use measures of model complexity that take the flexibility of models into account.

An approach to model selection that naturally takes model flexibility into account is Minimum Description Length (MDL), a statistical framework based on algorithmic coding theory (for a comprehensive review of MDL, see Grünwald, 2007). In the MDL framework, both models and data are understood as codes that can be compressed. The goal of MDL is to assess models in terms of their ability to compress data. The greater the compression, the better the account of the underlying regularities that are present in the data. In the present context, an implementation of the principles of MDL is the Normalized Maximum Likelihood (NML; see Myung, Navarro, & Pitt, 2006), an extension of Maximum Likelihood that is defined (for discrete data) as:

$$\mathrm{nml}(x) = \frac{f(x|\hat{\theta}_x)}{\Sigma f(y|\hat{\theta}_y)}. \tag{9}$$

The numerator represents the likelihood for data set x observed in an experiment, given maximum likelihood parameter estimates $\hat{\theta}_x$, and the denominator corresponds to the sum of the likelihoods for all possible data (denoted by y) that could in principle be observed in such an experiment, given their respective maximum likelihood parameter estimates $\hat{\theta}_y$. The ability to fit arbitrary data sets reflects model complexity, as more complex models provide better fits in general than simpler ones: The more complex the model, the larger the denominator, resulting in a smaller normalized likelihood value. By considering a model's ability to fit observed data relative to its ability to fit any data in general, NML provides a principled implementation of Occam's Razor. The operationalization of model complexity provided by NML can be used as a penalty factor that is added to the model's goodness of fit, as in AIC and BIC. Despite being computationally intensive, this measure has advantages relative to AIC and BIC because it takes into account the functional form of the model equations, making it useful even in cases where models have the same number of parameters. A representation of NML for an arbitrary model ($\mathcal{M}_i$) is

$$\mathrm{NML}_{\mathcal{M}_i} = 2[\ln(f(x|\hat{\theta}_x, \mathcal{M}_0)) - \ln(f(x|\hat{\theta}_x, \mathcal{M}_i))] + 2\ln\left(\sum_y f(y|\hat{\theta}_y, \mathcal{M}_i)\right), \tag{10}$$

where $\mathcal{M}_1, \ldots, \mathcal{M}_k$ are the models to be compared and $\mathcal{M}_0$ corresponds to the saturated model. The first term of the equation corresponds to the likelihood-ratio test statistic (G^2) for a model, quantifying badness of fit, and the second term is the penalty factor for model complexity. Models with smaller NML values are preferred as providing the better compromise between fit and parsimony.

5. Goodness of fit and minimum description length results

The six candidate models were fitted to the data obtained from the two experiments by Parks and Yonelinas (2009), using the maximum likelihood method. The model fitting procedure was implemented in R (R Development Core Team, 2009). The NML values were obtained via exhaustive computation of model fits for each model across all possible datasets in each condition. This procedure was implemented in Fortran, using the NAG Library. R scripts and Fortran codes can be obtained from the authors.

Given that only individual response probabilities were available, the response frequencies had to be reconstructed. In each condition, 40 participants were tested. In the Item Condition, the recognition task consisted of 120 single-choice and 240 two-choice trials per participant. In the remaining conditions, the recognition task comprised 30 single-choice and 60 two-choice trials instead.

One common issue in fitting models to data is whether one should aggregate over participants or fit them individually. Although aggregation is well known to create severe distortions in the data and lead to wrong inferences (e.g., Estes & Maddox, 2005), it can still be advantageous in situations with a small number of trials per condition (Cohen, Sanborn, & Shiffrin, 2008). For this

[2] An exception would be the observation of below chance performance in first choice and/or second choice, which would be diagnostic in that it can only be accommodated by UVSDT, but by none of the other models.

Table 2
Goodness of fit and NML results for Parks and Yonelinas (2009) data.

Condition	Model	Summed (individual) G^2	Aggregated G^2	NML penalty factor	Summed NML
Item	EVSDT	278.70*	237.84*	2.855	392.92
	UVSDT	33.77	4.12*	4.709	222.12
	DPSDT	40.88	4.12*	3.688	188.40
	MSDT	40.88	4.12*	3.688	188.40
	1HTM	260.63*	115.40*	2.815	373.22
	2HTM	39.25	4.12*	3.937	196.72
Pair	EVSDT	186.47*	97.72*	2.209	274.85
	UVSDT	56.90*	0.53	3.509	197.28
	DPSDT	68.72*	0.53	2.679	175.90
	MSDT	68.72*	0.53	2.679	175.90
	1HTM	111.58*	8.55*	2.171	198.42
	2HTM	65.81*	0.53	2.858	180.14
Sentence	EVSDT	203.04*	153.07*	2.209	291.42
	UVSDT	43.25	0.65	3.509	183.63
	DPSDT	56.08*	0.65	2.679	163.26
	MSDT	56.08*	0.65	2.679	163.26
	1HTM	81.73	2.37	2.171	168.57
	2HTM	55.74*	0.65	2.858	170.07
Compound	EVSDT	151.81*	97.89*	2.209	240.19
	UVSDT	48.77	0.81	3.509	189.15
	DPSDT	55.35	0.81	2.679	162.53
	MSDT	55.35	0.81	2.679	162.53
	1HTM	98.80	12.00*	2.171	185.64
	2HTM	54.86	0.81	2.858	169.19

Note. For the goodness of fit test, both EVSDT and 1HTM have 80 degrees of freedom, summing the individual fit values, and 2 for the aggregated data.
All other models have 40 and 1 degrees of freedom, respectively.
Also, the NML penalty factors (for individual data) are the same in the Pair, Sentence and Compound conditions.
* $p < 0.05$.

reason, the models were fitted to both aggregated and individual data. The NML values are only reported for the individual data fits because for the aggregated data, the number of trials considered becomes too large and the computational costs of obtaining NML become rather prohibitive. Nevertheless, the rank order between the models' penalty factors should be preserved in the case of aggregated data.

The goodness of fit results in Table 2, reported in terms of the G^2 statistic, show that, with the exception of EVSDT, which is rejected in all experimental conditions (smallest $G^2(1) = 97.72$, $p < 0.01$), both with individual and aggregated data, the model fits are good. For the aggregated data, the UVSDT, MSDT, DPSDT, and 2HTM models perform equally well in the four experimental conditions. Considering individual fits, the summed G^2 results reveal some considerable advantage for UVSDT in both experiments, followed by the 2HTM, which is in turn followed by DPSDT and MSDT which are perfectly tied. This tie occurs because both models always fit equally well data sets in general, which means that they are not distinguishable within the present paradigm on the basis of goodness of fit results. Also, the ordering in goodness of fit values of the models is as expected on the basis of the inclusion pattern derived in the previous section and depicted in Fig. 2.

Surprisingly, 1HTM achieves better results than EVSDT in every experimental condition, even in the Item condition, a result that reinforces the idea that the sole reliance on summary statistics like correlation between choice accuracies can be misleading. Fig. 4 presents the observed (π_1, π_2^c) values in the Item condition, which is the experimental condition in which the highest correlation between single-choice and (conditional) second-choice correct responses was observed (see Table 1). The individual fits reveal that the 1HTM fits better 21 individual datasets, while EVSDT fits better the remaining 19. It can be seen that despite the fact that EVSDT implies a positive correlation between the two choices, the observed relationship between correct first and second choices seems not to be sufficient to cause a better fit for the EVSDT model relative to the 1HTM model which predicts no correlation at all. As shown in the next paragraphs, these results are not caused by

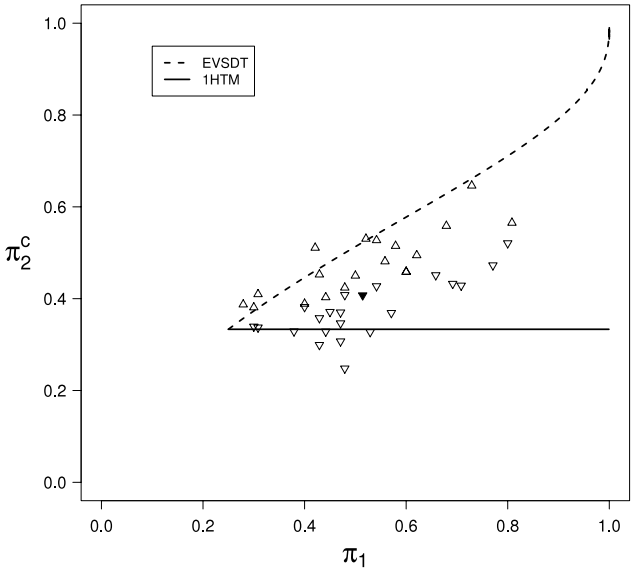

Fig. 4. Predictions of EVSDT and 1HTM for (π_1, π_2^c), and the observed data from the Item condition. The upward-pointing triangles correspond to the observed individual (π_1, π_2^c) values that are better fitted by EVSDT, while the downward-pointing triangles correspond to the values that are better fitted by 1HTM. The filled downward-pointing triangle corresponds to the aggregated data (which are fitted better by 1HTM). Note that there are only 18 visible upward-pointing triangles, given that two participants (participants 3 and 37 in the original dataset) produced exactly the same choice frequencies.

a greater flexibility of the 1HTM relative to the EVSDT. In fact, EVSDT is slightly more complex than 1HTM. Two conclusions can be retained from these results, that (1) summary statistics such as the correlation between responses do not provide an adequate assessment of model performance, and (2) neither the EVSDT nor the 1HTM provides satisfactory accounts of the observed data.

The advantage of UVSDT is mostly based on a superior fit to data patterns in which second-choice accuracy is below chance

level. UVSDT is the only model considered that can predict below chance second-choice accuracy, which occurs when the σ parameter assumes large values. In the remaining cases the fits of all models are virtually identical. Binomial tests were performed to check whether these cases of under chance performance are statistically different from chance performance. Across the four experimental conditions, only two participants were found to have second-choice accuracy significantly lower than chance ($p < 0.05$), suggesting that these are mere cases of sampling error which are being capitalized upon by an overfitting UVSDT, a flexibility that is taken into account in the NML results reported later.

The simpler models EVSDT and 1HTM can only predict first-choice and second-choice accuracies that are positively correlated or uncorrelated, respectively, because they assume a strong structure in the data. On the contrary, the more complex models like UVSDT are able to produce a wide range of different degrees of association between first- and second-choice accuracies as discussed in the previous section. Therefore, summary statistics like the correlations between first- and second-choice accuracy are not informative because all of the more complex models can account for them. It just depends on the relationships between each model's parameters, which in the absence of firm a priori expectations are best assessed by direct model fits.

When considering the NML results, also reported in Table 2, the differences between the candidate models change considerably. The rank order of computed penalty factors was (from more complex to less complex): UVSDT > 2HTM > DPSDT = MSDT > EVSDT > 1HTM. The calculated NML penalty factors for an individual data set in each condition are also reported in Table 2. Note that because they have the same number of trials, the Pair, Sentence, and Compound conditions have same the penalty factors. Regarding all the four datasets, the DPSDT and MSDT emerge as the most adequate models, having identical results (as they have to, given that they are equivalent models in the 4AFC-2R task). It is important to note that these results contradict the notion that hybrid models are in general more complex because they postulate additional cognitive processes. The better NML results for the hybrid models than for UVSDT demonstrate that conceptual complexity in psychological models does not necessarily translate into stochastic complexity in their corresponding statistical models.

The 2HTM model is the third most adequate model in terms of NML, except for the Sentence condition, where 1HTM performs better. The flexibility of the UVSDT is clearly punished by NML: In the Sentence and Compound conditions it only outperforms the worst model (EVSDT), while on the remaining conditions it also outperforms 1HTM.

6. Evaluation of parameter estimates

Regarding the parameter estimates, shown in Table 3, the estimates obtained with both aggregated and individual data seem to be consistent with each other, suggesting that no major and systematic distortions are produced by data aggregation. The estimates obtained are also roughly consistent with the ones reported in the literature on the basis of ROC analyses. Despite this consistency, the interquantile ranges are in some cases large, as in the case of the μ parameter from the MSDT, and the σ parameter from the UVSDT model, reflecting not only inter-participant variability, but also estimation error due to the fact that some response categories like response category 2 (first-choice incorrect *and* second-choice correct) occur infrequently, leading to poor estimates. Nevertheless, the goodness of fit results and the consistency with estimates obtained by means of ROCs suggest that the 4AFC-2R task constitutes a viable means of estimating different candidate models.

Table 3
Parameter estimates and corresponding correlations. Values on the left of the dashes (/) correspond to estimates based on aggregated data, while the values on the right correspond to the medians and the interquantile ranges of individual data estimates.

EVSDT	μ		
Item	0.83/0.76 (0.54)		
Pair	0.82/0.77 (0.51)		
Sentence	0.91/0.83 (0.80)		
Compound	0.88/0.93 (0.76)		
UVSDT	μ	σ	$r(\mu, \sigma)$
Item	0.93/0.89 (0.68)	1.48/1.43 (0.46)	0.42*
Pair	0.95/0.90 (0.93)	1.67/1.55 (1.20)	0.71*
Sentence	1.15/0.96 (1.59)	1.91/1.75 (0.96)	0.84*
Compound	1.03/1.18 (1.00)	1.67/1.56 (0.88)	0.60*
DPSDT	μ	R	$r(\mu, R)$
Item	0.34/0.38 (0.51)	0.27/0.23 (0.19)	0.06
Pair	0.19/0.14 (0.42)	0.32/0.29 (0.34)	−0.33*
Sentence	0.09/0.11 (0.35)	0.39/0.31 (0.43)	−0.10
Compound	0.22/0.25 (0.58)	0.33/0.31 (0.33)	−0.20
MSDT	μ	λ	$r(\mu, \lambda)$
Item	1.93/1.86 (1.12)	0.48/0.62 (0.47)	−0.58*
Pair	2.46/2.97 (6.79)	0.42/0.53 (0.45)	−0.63*
Sentence	3.11/3.12 (5.90)	0.43/0.55 (0.36)	−0.48*
Compound	2.39/2.83 (4.84)	0.45/0.64 (0.31)	−0.79*
1HTM	D_o		
Item	0.36/0.35 (0.23)		
Pair	0.36/0.35 (0.23)		
Sentence	0.41/0.36 (0.33)		
Compound	0.39/0.41 (0.33)		
2HTM	D_o	D_n	$r(D_o, D_n)$
Item	0.32/0.30 (0.21)	0.14/0.16 (0.18)	0.48*
Pair	0.35/0.32 (0.25)	0.08/0.06 (0.19)	−0.10
Sentence	0.40/0.35 (0.34)	0.04/0.05 (0.16)	0.12
Compound	0.36/0.41 (0.34)	0.10/0.11 (0.23)	0.15

* $p < 0.05$.

Considering the discrete-state models, 1HTM and 2HTM, the estimates of D_o are quite similar across conditions, and parameter D_n seems to have a larger role only in the Item and Compound conditions. Given that the 1HTM is nested within the 2HTM, it is possible to test whether the 2HTM model describes the data significantly better than the 1HTM through their differences in goodness of fit.

Note that the null hypothesis ($D_n = 0$) lies on the boundary of the alternative hypothesis ($0 < D_n \leq 1$). In these circumstances the sampling distribution of the likelihood-ratio test statistic ΔG^2 no longer corresponds to the χ^2 distribution with the appropriate degrees of freedom, but to a weighted mixture of χ^2 distributions with different number of degrees of freedom (see Self & Liang, 1987). More specifically, for the aggregated data, the sampling distribution corresponds to an equal mixture of χ^2 distributions with $df = 0$[3] and $df = 1$:

$$P(\Delta G^2|H_0) = 0.5P(\Delta G^2|\chi_0^2) + 0.5P(\Delta G^2|\chi_1^2). \quad (11)$$

For the summed individual fits, the sampling distribution corresponds to the sum of a mixture of χ^2 distributions with degrees of freedom ranging from 0 to 40 (the number of individuals):

$$P(\Delta G^2|H_0) = \sum_{i=0}^{40} (0.5)^{40} \binom{40}{i} P(\Delta G^2|\chi_i^2). \quad (12)$$

For the aggregated data, these differences were significant for all conditions (smallest $\Delta G^2 = 8.02$, $p < 0.01$), except for the

[3] The χ^2 distribution with $df = 0$ is a distribution that puts all probability mass on 0.

Sentence condition ($\Delta G^2 = 1.72, p = 0.09$). The summed individual data fits produced equivalent results, as the difference in goodness of fit is significant for every condition (smallest $\Delta G^2 = 43.94, p < 0.01$) except for the Sentence condition ($\Delta G^2 = 25.99, p = 0.19$). These results suggest that the prediction of Parks and Yonelinas (2009) that distractor detection should be absent in the three pair recognition conditions is not corroborated by the data, and they reinforce the notion that the D_n parameter is not solely modeling "recall-to-reject" judgments (see footnote 1 Rotello & Heit, 2000).

For the UVSDT model, the median σ estimates were between 1.43 and 1.75, and thus somewhat larger than the average of 1.25 that is usually estimated by means of ROCs (Yonelinas & Parks, 2007). One reason for these high σ estimates is the occurrence of under chance second-choice accuracy in some participants, which forces the UVSDT model to assume high values for the standard deviation of the signal distribution. The differences between UVSDT and EVSDT are always statistically significant (smallest $\Delta G^2 = 97.09, p < 0.01$). Also, both μ and σ estimates are positively correlated. This correlation is easily explained in the case of UVSDT by the plausible assumption that better encoding produces not only larger values of μ but also larger encoding variabilities. But still, the presence of a correlation of this magnitude is an interesting issue to be further explored because according to the first generation of globalist memory models, increases in μ should be accompanied by larger values of σ, a prediction which was subsequently contested through the use of ROCs (Ratcliff, Sheu, & Gronlund, 1992; but see Glanzer, Kim, Hilford, & Adams, 1999).

Regarding the DPSDT, estimates of parameter μ were lower for the Pair and Sentence conditions than for the Item and Compound conditions, which is consistent with the hypothesis that recollection is the sole mechanism underlying the pair recognition performance (except for the Compound condition). This hypothesis can be directly tested, by assessing the difference in goodness of fit when μ is constrained to be zero, a parameter constraint that produces a model equivalent to the 1HTM. Note that similarly to the comparison between 2HTM and 1HTM, the null hypothesis ($\mu = 0$) is again at the boundary of the alternative hypothesis ($\mu > 0$).

For the aggregated data, the difference was only non-significant for the Sentence condition ($\Delta G^2 = 1.72, p = 0.09$). In terms of the summed individual data fits, the differences in G^2 were significant in every condition (smallest $\Delta G^2 = 42.86, p < 0.01$) except for the Sentence condition ($\Delta G^2 = 25.65, p = 0.20$). Overall, these differences in goodness of fit are not inconsistent with the theoretical principles underlying the DPSDT.

Concerning the MSDT model parameter, μ, large median and interquantile range values are found across the different experimental conditions. While the interquantile ranges might just reflect large individual variability and/or sampling and estimation error, the high medians can be interpreted as a strong indication that the MSDT is trying to approximate the 1HTM model.[4]

Also, the negative correlation found between μ and λ seems to be consistent with this idea. This pattern can be considered as somewhat inconsistent with the notion that λ represents attentive processes, as it would mean that to the extent to which individuals paid more attention to some items during encoding (leading to a higher μ), they also tended to pay attention to a smaller proportion of items (reflected by smaller λ values). One possible argument against this supposed inconsistency is that these negative correlations occur due to overfitting sparse data. A problem with this argument is that a significant negative correlation is also present in the Item condition, in which a larger number items were tested.[5]

The MSDT model considered here is a restricted version, with μ^* set equal to zero (see DeCarlo, 2002) because the unrestricted version is not identified in the 4AFC-2R task. An unrestricted version like the some-or-none model (e.g., Onyper, Zhang, & Howard, 2010) could perhaps present a different relationship between parameter estimates. The characterization of MSDT in the Appendix shows that leaving μ^* free to vary does not add new predictions to the MSDT model, rendering this possibility unlikely but not impossible. Taken together, these findings illustrate that parameter estimates can be useful to differentiate between candidate models that are formally equivalent, as the DPSDT and the MSDT are in the 4AFC-2R task.

7. General discussion

The 4AFC-2R task was one of the original tasks used to pit EVSDT against 1HTM (Swets et al., 1961), and alongside the shape of ROCs and the relation between yes/no and 2AFC performance, has been considered so far one of the main sources of evidence for adequately distinguishing them. Parks and Yonelinas' (2009) study introduces the 4AFC-2R task to the recognition memory literature, representing an interesting addition in a well developed research field that is in need of new approaches. The present study builds on Parks and Yonelinas' (2009) work and extends it in several ways by (a) providing a thorough formalization and characterization of the models, (b) considering a wider choice of candidate models, (c) implementing model selection methods that provide an assessment of model complexity and adequacy, and (d) discussing how parameter estimates can be used for validation and comparison.

The results obtained are consistent with Parks and Yonelinas' (2009) conclusions, and suggest a preference for the hybrid models, namely the MSDT and DPSDT, for which model fit and model selection results are identical. Despite this similarity in goodness of fit and model selection indices, if the parameter estimates are also considered in the general assessment of the models, then it can be claimed that the DPSDT has the best overall performance. The DPSDT not only adequately describes the observed data, but also does so in a way that is not inconsistent with its theoretical underpinnings in terms of psychological processes. A model that performed quite poorly was the UVSDT, which is surprising given that it is one of the most important models considered in this field. The major reason for this was its lack of parsimony. Of course, these model characterization and model selection results are restricted to the 4AFC-2R task; in the future, approaches that

[4] In part, this reflects a boundary problem for the MSDT: Data with second-choice accuracy at or below chance correspond to the boundary case $\mu = \infty$, so that a finite maximum likelihood estimate of μ does not exist, leading the maximization routine to approximate infinity by choosing a μ value so large that larger values do not change the likelihood noticeably any more. However, this leads to unstable estimates for μ. To deal with this problem, we imposed an upper bound of $\mu = 10$, corresponding to second-choice accuracy approaching the lower bound by a margin of 10^{-12} or less and leading to numerically stable estimates and likelihood values that were numerically equal to that of the DPSDT model that differs from MSDT in terms of model space only in that it does not have this boundary problem.

[5] The negative correlations do not appear to go back to the boundary problem discussed in Footnote 4. To check this, we fitted the model with the μ parameter transformed via different transformations (hyperbolic tangent, logistic) that monotonically map the interval $[0, \infty]$ onto a finite interval such as [0, 1], effectively including $\mu = \infty$ as an admissible parameter value. The correlations between the estimates of the λ parameter and the transformed μ parameter remained significantly negative in each case. Furthermore, excluding participants with second-choice accuracy at or below chance (i.e., participants for which the boundary problem exists, namely 6, 17, 19, and 12 participants from, in order, the Item, Pair, Sentence, and Compound conditions) still resulted in negative correlations between λ and the untransformed μ estimates that were significant in each condition other than the Sentence condition ($r = -0.35, p = 0.125$) for which almost half of the participants had to be excluded.

try to integrate these results with other approaches, such as the ones obtained via ROCs should be considered. For instance, Wixted (2007) performed a model recovery simulation that pitted the UVSDT against the DPSDT in a confidence-rating ROC paradigm, obtaining results that indicated a greater flexibility for the DPSDT model, which is inconsistent with the present NML results and the model characterization results from Theorems 1 and 3. This inconsistency suggests that model flexibility might be paradigm specific, a possibility that should be further investigated.

Another important aspect is that the parameter estimates are consonant with the ones that are generally found in the literature. This aspect should be further exploited as it can be extremely informative in the study of relationships between a model's parameter estimates across experimental manipulations. Several studies have found that for some experimental manipulations, like memory strength (via study time), the slope of the zROC function (mapping of probit-transformed hits and false alarms proportions) does not seem to vary along with the function's intercept, while other manipulations like word frequency or level-of-processing during study tend to produce a positive correlation between slope and intercept (for a review see Yonelinas & Parks, 2007). Given that for SDT models the slope of the zROC corresponds to the ratio of the standard deviations of studied and distractor distributions, and the intercept corresponds to μ, this suggests that the mean and standard deviation of the signal distribution are dissociable. These findings have been used as an argument in favor for DPSDT, given that UVSDT has no principled argument for these dissociations (Yonelinas & Parks, 2007). Nevertheless, it is important to consider that the shape of ROCs, and consequently of zROCs, can be affected by several confounding factors like response mapping (e.g., Klauer & Kellen, 2010; Malmberg, 2002), or response criterion noise (Benjamin et al., 2009), which have the potential of distorting results. In the 4AFC-2R task, only the relative strength and/or detection of the alternatives is considered by the models, with no guessing biases or response criteria being postulated. This fundamental difference might be capitalized on by using this method as a way of cross-validating findings obtained via ROCs, leading to a better understanding and a less paradigm-specific assessment of the relationship between parameters.[6]

Nevertheless, the present results also point out limitations of the 4AFC-2R task for discriminating between different candidate models. Efforts in model selection should focus on experimental designs that can produce different predictions for each model, otherwise the evidence produced will be of little value (e.g., Roberts & Pashler, 2000). In the present case, the more complex models show considerable overlap in the range of data patterns that they can account for (see Fig. 2) which leads to model complexity being the major source of model discriminability (for a similar case in free recall research, see Howard, Jing, Rao, Provyn, & Datey, 2009). The fact that models based on distinct assumptions can predict similar results should be interpreted as a major shortcoming of this specific method. Nevertheless, these limitations are shared to some extent by other methods like ROCs based on confidence ratings, which have provided large bodies of evidence supporting distinct candidate models without any decisive conclusion reached so far (DeCarlo, 2010, Yonelinas & Parks, 2007, Wixted, 2007.

Even if these criticisms are taken into account, we believe that dismissing the 4AFC-2R task as a useless method for evaluating recognition memory models would be too harsh a verdict. The fact that the 4AFC-2R is the only known method besides the use of ROCs that allows for the estimation of parameters like σ, λ, or R in the more complex models should be sufficient to forestall its simple dismissal. This method provides an alternative and relatively inexpensive means to assess parameter values, their relationship, as well as their responsiveness to experimental manipulations. In this respect, several refinements can still be made, like the development of hierarchical models (e.g., Klauer, 2010; Rouder & Lu, 2005), which can potentially reach a compromise between the advantages and disadvantages of both individual and aggregated data modeling, or even the disentangling of potential item effects that can also distort parameter estimates (Freeman, Heathcote, Chalmers, & Hockley, 2010).

Other refinements worth exploring would be the inclusion of an additional condition in which participants are requested to indicate an alternative that they consider more likely to be incorrect (see Green & Swets, 1966, p. 110), or the implementation of a 4AFC task with three choices (4AFC-3R). The additional response category introduced by the third response would drastically reduce the overlap between the model spaces of the candidate models, contributing to the ability of the paradigm to dissociate between models.[7]

Also, it is important to consider that the adequacy of different models should not only be based on the goodness of fit in one specific paradigm, but on its generalizability to related tasks. In this perspective, the 4AFC-2R task might be used as a parallel task that permits the assessment of the robustness of parameter estimates across different operationalizations (e.g., Busemeyer & Wang, 2000; Jang, Wixted, & Huber, 2009). One of the main criticisms in the literature regarding recognition memory is its overreliance on ROCs (e.g., Ratcliff & Starns, 2009). This mono-operation bias is especially problematic given that the candidate models considered are assumed to be more than purely descriptive measurement models; they are also explanatory models whose parameters have specific meanings that are integrated in larger theoretical frameworks (Yonelinas & Parks, 2007). If the objective is to go beyond goodness of fit and understand if the model's parameters act according to their theoretical interpretations, an important step in cognitive modeling in general (e.g., Bayen, Murnane, & Erdfelder, 1996; DeCarlo, 2010; Voss, Rothermund, & Voss, 2004), then the 4AFC-2R task can be a valuable addition to the researchers' toolbox.

Acknowledgments

The research reported in this article was supported by grant SFRH/BD/48346/2008 from the Fundação para a Ciência e a Tecnologia to the first author, and by grant Kl614/31-1 from the Deutsche Forschungsgemeinschaft to the second author. We would like to thank Colleen Parks for providing the original data.

Appendix. Characterizing the models in terms of their predictions

In the Appendix, we provide proofs for Theorems 1–4. In doing so, we will also show that each model is identified.

[6] A similar approach, in the sense of assessing models' ability to account for phenomena in different tasks, was implemented in the field of visual perception (Solomon, 2007a,b), in which different candidate models had to account for data from second-choice responses in the 4AFC-2R and the psychometric function for 2AFC detection.

[7] The two response categories of the 4AFC-2R task allow parameter identifiability so that the probability of third-choice correct (π_3) can be expressed as function of the first-choice correct and second-choice correct probabilities π_1 and π_2. This means that model predictions are two-dimensional surfaces in a three-dimensional probability space generated by π_1, π_2, and π_3.

A.1. Characterizing UVSDT in the 4AFC-2R task

Regarding the Signal Detection model with unequal variances, UVSDT, remember that $\pi_1(\mu,\sigma) = \int F^3 f_{\mu,\sigma}$ and $\pi_2(\mu,\sigma) = 3\int F^2(1-F)f_{\mu,\sigma}$.

We need to consider the function $g(\sigma) = \int F^3(z) f_{(0,\sigma)}(z)\,\mathrm{d}z = \int F^3(\sigma z) f(z)\,\mathrm{d}z$. Note first that g is strictly increasing as a function of $\sigma > 0$. To see this, note that its derivative, $g'(\sigma)$ is positive:

$$\begin{aligned}
g'(\sigma) &= 3\int zF^2(\sigma z)f(\sigma z)f(z)\,\mathrm{d}z\\
&= 3\int_{-\infty}^{0} zF^2(\sigma z)f(\sigma z)f(z)\,\mathrm{d}z\\
&\quad + 3\int_{0}^{\infty} zF^2(\sigma z)f(\sigma z)f(z)\,\mathrm{d}z\\
&= 3\int_{0}^{\infty} (-z)F^2(-\sigma z)f(-\sigma z)f(-z)\,\mathrm{d}z\\
&\quad + 3\int_{0}^{\infty} zF^2(\sigma z)f(\sigma z)f(z)\,\mathrm{d}z\\
&= 3\int_{0}^{\infty} (-z)(1-F(\sigma z))^2 f(\sigma z)f(z)\,\mathrm{d}z\\
&\quad + 3\int_{0}^{\infty} zF^2(\sigma z)f(\sigma z)f(z)\,\mathrm{d}z\\
&= 3\int_{0}^{\infty} z(F^2(\sigma z) - (1-F(\sigma z))^2)f(\sigma z)f(z)\,\mathrm{d}z\\
&> 0,
\end{aligned}$$

because $F(\sigma z) > (1 - F(\sigma z))$ for $z > 0$.

Furthermore, the normal distribution with mean 0 and standard deviation σ converges in probability law to a point distribution ϵ_0 with all of its probability mass on 0, as $\sigma \to 0$. This can easily be verified by means of the moment generating functions of the normal distribution and the point distribution. It follows that $g(\sigma) \to F^3(0) = \frac{1}{8}$ as $\sigma \to 0$ and by the monotonicity of g that $g(\sigma) > \frac{1}{8}$.

On the other hand, $g(\sigma) \to \frac{1}{2}$ as $\sigma \to \infty$, implying that $g(\sigma) < \frac{1}{2}$. To see this, note that (a) $F^3(\sigma z)f(z) \to (1_{\{z>0\}}(z) + F^3(0)\,1_{\{z=0\}}(z))f(z)$ for all z as $\sigma \to \infty$, while $F^3(\sigma z)f(z)$ is dominated by $f(z)$ (i.e., $|F^3(\sigma z)f(z)| \le f(z)$ for all z). It follows from the dominated convergence theorem (Bartle, 1995, Chap. 5) that $g(\sigma) \to \int (1_{\{z>0\}}(z) + F^3(0)1_{\{z=0\}}(z))f(z)\,\mathrm{d}z = \frac{1}{2}$.

Taken together, $g(\sigma)$ is strictly increasing and ranges from $\frac{1}{8}$ to $\frac{1}{2}$. It follows that its inverse, g^{-1} exists on the interval $\left(\frac{1}{8}, \frac{1}{2}\right)$. We will prove the following theorem:

Theorem 1. *The set of probabilities (π_1, π_2) described by the UVSDT model with $\mu \ge 0$ and $\sigma > 0$ is $\{(\pi_1,\pi_2) : \frac{1}{8} < \pi_1 < 1$ and $\frac{1}{2} - \pi_1 \le \pi_2 < 3(\pi_1^{\frac{2}{3}} - \pi_1)$ for $\pi_1 < \frac{1}{2}$ and $0 < \pi_2 < 3(\pi_1^{\frac{2}{3}} - \pi_1)$ for $\pi_1 \ge \frac{1}{2}\}$.*

It is easy to see that $\pi_1(\mu,\sigma) = \int F^3 f_{\mu,\sigma}$ increases as μ increases for a fixed $\sigma > 0$ and that it approaches 1 arbitrarily closely as $\mu \to \infty$, establishing the upper bound for π_1. On the other hand, it follows that $\pi_1(\mu,\sigma) \ge \pi_1(0,\sigma) = g(\sigma)$. The lower bound on π_1 in the theorem follows from the above discussion of the function g.

To obtain the bounds for π_2, consider a fixed value of π_1, say π_1^0 with $\frac{1}{8} < \pi_1^0 < 1$. As already mentioned, the function $\pi_1(\mu,\sigma)$ is strictly increasing in μ for any fixed $\sigma > 0$. It ranges from $g(\sigma)$ to 1. It follows that for a fixed $\sigma > 0$, one and only one $\mu = \mu(\sigma)$ exists with $\pi_1(\mu,\sigma) = \pi_1^0$ if and only if $g(\sigma) \le \pi_1^0$, that is, if and only if π_1^0 is in the range of $\pi_1(\mu,\sigma)$ as μ ranges from 0 to infinity. This implies that for $\pi_1^0 \ge \frac{1}{2}$, $\mu(\sigma)$ is defined for each $\sigma > 0$ (because $g(\sigma) < \frac{1}{2}$ for all σ), whereas for $\pi_1^0 < \frac{1}{2}$, $\mu(\sigma)$ is defined for $0 < \sigma \le \sigma_0 = g^{-1}(\pi_1^0)$ (because g is strictly increasing). Note that indeed $\pi_1(0,\sigma_0) = \pi_1^0$, implying that $\mu(\sigma_0) = 0$.

It follows that the function $\mu(\sigma)$ implicitly defined by $\pi_1(\mu(\sigma),\sigma) = \pi_1^0$ exists, that is, there is one and only one $\mu(\sigma)$ with $\pi_1^0 = \pi_1(\mu(\sigma),\sigma)$ for each $\sigma > 0$ and for each σ with $0 < \sigma \le \sigma_0$ for $\pi_1^0 \ge \frac{1}{2}$ and $\pi_1^0 < \frac{1}{2}$, respectively. By standard theorems on implicit functions and the regularity of $\pi_1(\mu,\sigma)$, $\mu(\sigma)$ is continuously differentiable (Erwe, 1962, Chap. 5).

We will show that $\pi_2(\mu(\sigma),\sigma)$ is strictly decreasing as a function of σ. Note that this implies global identifiability of the UVSDT, because each point in the model space can be generated by only one pair (μ,σ) due to the monotonicity of $\pi_2(\mu(\sigma),\sigma)$. To show the monotonicity, we need the derivative of $h_n(\sigma) = \int F^n(z) f_{(\mu(\sigma),\sigma)}\,\mathrm{d}z = \int F^n(\sigma z + \mu(\sigma))f(z)\,\mathrm{d}z$ for $n = 2$ and $n = 3$ with respect to σ. It is given by

$$\begin{aligned}
h_n'(\sigma) &= n\int F^{n-1}(\sigma z + \mu(\sigma))f(\sigma z + \mu(\sigma))(z + \mu'(\sigma))f(z)\,\mathrm{d}z\\
&= n\int F^{n-1}(z)f(z)\left(\frac{z-\mu(\sigma)}{\sigma} + \mu'(\sigma)\right) f_{(\mu(\sigma),\sigma)}(z)\,\mathrm{d}z.
\end{aligned}$$

Note that $h_3(\sigma) = \pi_1(\mu(\sigma),\sigma) = \pi_1^0$ for all σ for which $\mu(\sigma)$ is defined, hence $h_3'(\sigma) = 0$. It follows that

$$\begin{aligned}
&\mu'(\sigma)\int F^2(z)f(z)f_{(\mu(\sigma),\sigma)}(z)\,\mathrm{d}z\\
&\quad = -\frac{1}{\sigma}\int F^2(z)f(z)(z-\mu(\sigma))f_{(\mu(\sigma),\sigma)}(z)\,\mathrm{d}z.
\end{aligned}$$

Multiplying the derivative of h_2 by $\alpha(\sigma) = \int F^2 f f_{(\mu(\sigma),\sigma)} > 0$, and using the above expression for $\mu'(\sigma)\alpha(\sigma)$ it follows that

$$\begin{aligned}
h_2'(\sigma)\alpha(\sigma) &= \frac{2}{\sigma}\left(\int F(z)f(z)(z-\mu(\sigma))f_{(\mu(\sigma),\sigma)}(z)\,\mathrm{d}z\right.\\
&\qquad \times \int F^2(y)f(y)f_{(\mu(\sigma),\sigma)}(y)\,\mathrm{d}y\\
&\qquad - \int F^2(z)f(z)(z-\mu(\sigma))f_{(\mu(\sigma),\sigma)}(z)\,\mathrm{d}z\\
&\qquad \left.\times \int F(y)f(y)f_{(\mu(\sigma),\sigma)}(y)\,\mathrm{d}y\right)\\
&= \frac{2}{\sigma}\iint (z-\mu(\sigma))(F(y)-F(z))F(y)f(y)\\
&\qquad \times F(z)f(z)f_{(\mu(\sigma),\sigma)}(y)f_{(\mu(\sigma),\sigma)}(z)\,\mathrm{d}y\,\mathrm{d}z.
\end{aligned}$$

Setting $m(y,z) = F(y)f(y)F(z)f(z)f_{(\mu(\sigma),\sigma)}(y)f_{(\mu(\sigma),\sigma)}(z)$ and observing that $m(y,z) = m(z,y)$, it is seen that $\int \mu(\sigma)F(y)m(y,z)\mathrm{d}y\,\mathrm{d}z = \int \mu(\sigma)F(z)m(y,z)\,\mathrm{d}y\,\mathrm{d}z$, so that we can drop $\mu(\sigma)$ in the above integral. Note furthermore that $(z-y)(F(y)-F(z)) < 0$ for $y \ne z$ because $F(x)$ strictly increases as a function of x. Hence,

$$\begin{aligned}
0 &< \iint (z-y)(F(y)-F(z))m(y,z)\,\mathrm{d}y\,\mathrm{d}z\\
&= \iint z(F(y)-F(z))m(y,z)\,\mathrm{d}y\,\mathrm{d}z\\
&\quad - \iint y(F(y)-F(z))m(y,z)\,\mathrm{d}y\,\mathrm{d}z\\
&= \iint z(F(y)-F(z))m(y,z)\,\mathrm{d}y\,\mathrm{d}z
\end{aligned}$$

$$+ \iint y(F(z) - F(y))m(y,z)\,dy\,dz$$
$$= 2h_2'(\sigma)\alpha(\sigma)\frac{\sigma}{2}.$$

It follows that h_2 is strictly decreasing as a function of σ. Note that $\pi_2(\mu(\sigma),\sigma) = 3(h_2(\sigma) - \pi_1^0)$. It follows that $\pi_2(\mu(\sigma),\sigma)$ is strictly decreasing in σ.

For $\pi_1^0 < \frac{1}{2}$, $\sigma \le \sigma_0 = g^{-1}(\pi_1^0)$ and as already noted, $\mu(\sigma_0) = 0$. It follows that $\pi_2(\mu(\sigma),\sigma) \ge \pi_2(0, g^{-1}(\pi_1^0))$. Furthermore $\pi_2(0, g^{-1}(\pi_1^0))$ simplifies to $\frac{1}{2} - \pi_1$. To see this note that for every $\sigma > 0$,

$$2(\pi_1(0,\sigma) + \pi_2(0,\sigma)) = 2\int_{-\infty}^{\infty} F^2(z)(3 - 2F(z))f_{(0,\sigma)}(z)\,dz$$
$$= \int_{-\infty}^{\infty} F^2(z)(3 - 2F(z))f_{(0,\sigma)}(z)\,dz$$
$$+ \int_{-\infty}^{\infty} F^2(-z)(3 - 2F(-z))f_{(0,\sigma)}(-z)\,dz$$
$$= \int_{-\infty}^{\infty} F^2(z)(3 - 2F(z))f_{(0,\sigma)}(z)\,dz$$
$$+ \int_{-\infty}^{\infty} (1 - F(z))^2(1 + 2F(z))f_{(0,\sigma)}(z)\,dz$$
$$= \int_{-\infty}^{\infty} ((3F^2(z) - 2F^3(z)) + (1 - 3F^2(z) + 2F^3(z)))f_{(0,\sigma)}(z)\,dz$$
$$= \int_{-\infty}^{\infty} f_{(0,\sigma)}(z)\,dz = 1.$$

For $\pi_1^0 \ge \frac{1}{2}$, on the other hand, $\pi_2(\mu(\sigma),\sigma) \to 0$ as $\sigma \to \infty$. To see this, choose σ so large that $F(\sigma\epsilon) > 1 - \epsilon$ and $F(-\sigma\epsilon) < \epsilon$ for a fixed $\epsilon > 0$. It follows that

$$0 \le \int F^2(1 - F)(\sigma z + \mu(\sigma))f(z)\,dz$$
$$= \int_{-\infty}^{-\mu(\sigma)/\sigma-\epsilon} F^2(1 - F)(\sigma z + \mu(\sigma))f(z)\,dz$$
$$+ \int_{-\mu(\sigma)/\sigma-\epsilon}^{-\mu(\sigma)/\sigma+\epsilon} F^2(1 - F)(\sigma z + \mu(\sigma))f(z)\,dz$$
$$+ \int_{-\mu(\sigma)/\sigma+\epsilon}^{\infty} F^2(1 - F)(\sigma z + \mu(\sigma))f(z)\,dz$$
$$\le \epsilon + 2\epsilon + \epsilon = 4\epsilon,$$

implying that $\pi_2(\mu(\sigma),\sigma)$ approaches 0 arbitrarily closely as $\sigma \to \infty$. This establishes the lower bound for π_2.

Turning to the upper bound for π_2, let us first show that $|F^n(\mu(\sigma)) - \int F^n f_{(\mu(\sigma),\sigma)}|$ converges to zero, as $\sigma \to 0$ for $n = 2$ and $n = 3$. To see this, note that

$$\left|\int F^n(\sigma z + \mu(\sigma))f(z)\,dz - F^n(\mu(\sigma))\right|$$
$$= \left|\int \Big(F^n(\sigma z + \mu(\sigma)) - F^n(\mu(\sigma))\Big)f(z)\,dz\right|$$
$$\le \int |F^n(\sigma z + \mu(\sigma)) - F^n(\mu(\sigma))|f(z)\,dz$$
$$= \int \left|\int_{\mu(\sigma)}^{\sigma z+\mu(\sigma)} F^{n-1}(y)f(y)\,dy\right| f(z)\,dz$$
$$\le \int \sigma|z|f(z)\,dz = \sigma\frac{2}{\sqrt{2\pi}},$$

the last equality following from $\int_0^\infty zf(z)\,dz = -\frac{1}{\sqrt{2\pi}}e^{-0.5z^2}\Big|_{z=0}^{z=\infty} = \frac{1}{\sqrt{2\pi}}$. It follows that for σ small, $|F^n(\mu(\sigma)) - \int F^n f_{(\mu(\sigma),\sigma)}|$ becomes small.

Because $\int F^3 f_{(\mu(\sigma),\sigma)} = \pi_1^0$, it follows that $F^3(\mu(\sigma)) \to \pi_1^0$, or equivalently that $F(\mu(\sigma)) \to (\pi_1^0)^{\frac{1}{3}}$, as $\sigma \to 0$. By continuity of the inverse of F, this implies that $\mu(\sigma) \to F^{-1}((\pi_1^0)^{\frac{1}{3}})$. Furthermore $|(\pi_1^0)^{\frac{2}{3}} - \int F^2 f_{(\mu(\sigma),\sigma)}| \le |(\pi_1^0)^{\frac{2}{3}} - F^2(\mu(\sigma))| + |F^2(\mu(\sigma)) - \int F^2 f_{(\mu(\sigma),\sigma)}|$. It follows that $\int F^2 f_{(\mu(\sigma),\sigma)} \to (\pi_1^0)^{\frac{2}{3}}$ as $\sigma \to 0$. This, together with the fact that $\pi_2(\mu(\sigma),\sigma)$ is monotonically decreasing in σ establishes the upper bound for π_2.

A.2. Characterizing 2HTM in the 4AFC-2R task

For the 2HTM, remember that $\pi_1(D_o, D_n) = D_o + (1 - D_o)G(D_n)$ and $\pi_2(D_o, D_n) = (1 - D_o)H(D_n)$ with $G(D_n) = \frac{1}{4}(1 + D_n + D_n^2 + D_n^3)$ and $H(D_n) = G(D_n) - D_n^3$. It follows that G ranges from $\frac{1}{4}$ to 1 and that it is strictly increasing in D_n. We will prove the following theorem:

Theorem 2. *The set of probabilities (π_1, π_2) described by the 2HTM model is $\{(\pi_1,\pi_2) : \frac{1}{4} \le \pi_1 \le 1$ and $\frac{1}{3}(1 - \pi_1) \le \pi_2 \le H(G^{-1}(\pi_1))\}$. Furthermore, setting $a = \sqrt{3 - 30\pi_1 + 81\pi_1^2}$ and $b = 10 - 54\pi_1 + 6a$, $H(G^{-1}(\pi_1))$ can be written explicitly as*

$$\frac{1}{6}\left(2 - 18\pi_1 + (-4 + 18\pi_1 + 2a)b^{\frac{1}{3}} + (-1 + 9\pi_1 + a)b^{\frac{2}{3}}\right).$$

For a given π_1, say π_1^0, with $\frac{1}{4} \le \pi_1 < 1$, D_o can be expressed as a function of D_n: $D_o(D_n) = \frac{\pi_1^0 - G(D_n)}{1 - G(D_n)}$. $D_0 \ge 0$ implies that $\pi_1^0 \ge G(D_n)$ or, equivalently, that $D_n \le G^{-1}(\pi_1^0)$. Furthermore $D_o(G^{-1}(\pi_1^0)) = 0$.

Note that

$$H(D_n) = \frac{1}{4}(1 - D_n)(1 + 2D_n + 3D_n^2) \quad \text{and}$$
$$1 - G(D_n) = \frac{1}{4}(1 - D_n)(3 + 2D_n + D_n^2).$$

Hence, the function $C(D_n) = \frac{\pi_2(D_o(D_n),D_n)}{1-\pi_1(D_o(D_n),D_n)} = \frac{H(D_n)}{1-G(D_n)}$ equals $\frac{1+2D_n+3D_n^2}{3+2D_n+D_n^2}$, and its derivative with respect to D_n: $\frac{4(1+4D_n+D_n^2)}{(3+2D_n+D_n^2)^2}$. It follows that $C(D_n)$ is strictly increasing in D_n. Note that this implies that the 2HTM is identified for all $(\pi_1,\pi_2) \ne (1, 0)$. For $(\pi_1,\pi_2) = (1, 0)$ only D_o is identified, that is it follows that $D_o = 1$, whereas D_n can take on any value. The monotonicity of $C(D_n)$ also implies that

$$\frac{1}{3} = C(0) \le \frac{\pi_2(D_o(D_n), D_n)}{1 - \pi_1(D_o(D_n), D_n)} \le C(G^{-1}(\pi_1^0)),$$

hence $\frac{1}{3}(1 - \pi_1^0) \le \pi_2 \le H(G^{-1}(\pi_1^0))$. For $\pi_1^0 = 1$, on the other hand, $G^{-1}(\pi_1^0) = 1$, and $H(1) = 0$, so that the above bounds for π_2 also hold for $\pi_1^0 = 1$.

π_2 also take on these bounds. This is trivial for $\pi_1^0 = 1$. For $D_n = 0$ and $\pi_1^0 < 1$, it follows that $D_o(D_n) = \frac{1}{3}(4\pi_1^0 - 1)$. Because $H(0) = \frac{1}{4}$, it follows that $\pi_2(D_o(D_n), D_n) = (1 - D_o(D_n))H(D_n) = \left(1 - \frac{1}{3}(4\pi_1^0 - 1)\right)\frac{1}{4} = \frac{1}{3}(1 - \pi_1^0)$. For the upper bound and $D_n = G^{-1}(\pi_1^0)$, this follows immediately from $D_o(G^{-1}(\pi_1^0)) = 0$ and the definition of π_2. It is left to the reader to verify the explicit expression for $H(G^{-1}(\pi_1))$.

A.3. Characterizing DPSDT and MSDT in the 4AFC-2R task

Remember that the DPSDT model is defined by $\pi_1(\mu, R) = R + (1 - R)e_1(\mu)$ and $\pi_2(\mu, R) = (1 - R)e_2(\mu)$ whereas the MSDT model is defined by $\pi_1(\mu,\lambda) = \lambda e_1(\mu) + (1 - \lambda)e_1(\mu^*)$ and $\pi_2(\mu,\lambda) = \lambda e_2(\mu) + (1 - \lambda)e_2(\mu^*)$.

The function $e_1(\mu)$ is strictly increasing in μ. This follows from $e_1(\mu) = \int F^3(x) f_{(\mu,1)}(x)\,dx = \int F^3(x+\mu) f(x)\,dx$ and the fact that $F(x+\mu)$ is strictly increasing in μ. Furthermore, for $\mu \geq 0$ it ranges from $0.25 = e_1(0)$ to 1, without, however, ever reaching 1. It does, however, approach 1 arbitrarily closely as μ becomes large as is easily seen using the dominated convergence theorem. Note for later reference that this implies in particular that $e_2(\mu)$ tends to zero as μ becomes large, because the four e_i have to sum to one.

For these reasons, the inverse of e_1, e_1^{-1} exists and is defined on the interval (0, 1), and it is strictly increasing on that interval and continuously differentiable (by standard theorems on implicit functions; Erwe, 1962, Chap. 5). For π_1 in [0.25, 1), let $\mu_0 = \mu_0(\pi_1) = e_1^{-1}(\pi_1)$. Set $\mu_0(1) := \infty$, $e_1(\infty) := 1$, and $e_2(\infty) := 0$. Also, remember that the conditional probability of second-choice correct, given an incorrect first choice, under the EVSDT model is given by $c_2(\mu)$, $c_2(\mu) = \frac{e_2(\mu)}{1-e_1(\mu)}$. We will show the following two theorems:

Theorem 3. *The set of probabilities* (π_1, π_2) *described by the DPSDT model with* $\mu \geq 0$ *is* $\{(\pi_1, \pi_2) : \frac{1}{4} \leq \pi_1 \leq 1$ *and* $\frac{1}{3}(1-\pi_1) \leq \pi_2 \leq e_2(\mu_0)\}$.

Theorem 4. *The set of probabilities* (π_1, π_2) *described by the MSDT model with* $\mu \geq \mu^* \geq 0$ *for a fixed* μ^* *is* $\{(\pi_1, \pi_2) : e_1(\mu^*) \leq \pi_1 < 1$ *and* $c_2(\mu^*)(1-\pi_1) < \pi_2 \leq e_2(\mu_0)\} \cup \{(e_1(\mu^*), e_2(\mu^*))\}$.

Note that the lower bound, $\pi_2 = \frac{1}{3}(1-\pi_1)$, corresponds to the line predicted by 1HTM, whereas the upper bound, $\pi_2 = e_2(e_1^{-1}(\pi_1))$, corresponds to the curve predicted by EVSDT. To prove these theorems, we will make use of the following properties of the function c_2: For $\mu \geq 0$, $c_2(\mu) \geq \frac{1}{3}$, and $c_2(\mu)$ is non-decreasing in μ. Furthermore, we will need that the ratio of the derivatives of e_2 and e_1, $\frac{e_2'(\mu)}{e_1'(\mu)}$, is strictly decreasing in μ. These properties will be proved at the end of the proof of Theorems 3 and 4. We already noted that $c_2(0) = \frac{1}{3}$, and that $e_2(\infty) = 0$ (i.e. $e_2(\mu)$ tends to zero as μ becomes large).

Consider first the DPSDT model. Because $e_1(\mu) \leq 1$, $\pi_1(\mu, R) \geq Re_1(\mu) + (1-R)e_1(\mu) = e_1(\mu)$. It follows from the monotonicity of e_1 that $\pi_1(\mu, R) \geq e_1(0) = 0.25$. π_1 also takes on this value because $\pi_1(0, 0) = 0.25$. Furthermore π_1 is trivially smaller than or equal to one, and it can assume this value (for $R = 1$). For $R = 1$, or equivalently $\pi_1 = 1$, it follows trivially that $\pi_2 = 0$, hence, $0 = \frac{1}{3}(1-\pi_1) \leq \pi_2 \leq e_2(\mu_0) = e_2(\infty) = 0$.

It remains to be shown that $\frac{1}{3}(1-\pi_1) \leq \pi_2 \leq e_2(\mu_0)$ for values of (μ, R) with $\pi_1(\mu, R) < 1$. For a fixed value of $\pi_1(\mu, R) < 1$, say π_1^0, the model equation for π_1 can be solved for R, $R(\mu) = \frac{\pi_1^0 - e_1(\mu)}{1-e_1(\mu)}$. $R(\mu) \geq 0$ implies that $\pi_1^0 \geq e_1(\mu)$ or, equivalently $\mu_0 = e_1^{-1}(\pi_1^0) \geq \mu$. In turn, for each $\mu \leq \mu_0$, $\pi_1(\mu, R(\mu)) = \pi_1^0$.

From the model equations it follows that $\frac{\pi_2(\mu, R(\mu))}{1-\pi_1(\mu, R(\mu))} = c_2(\mu)$, hence by monotonicity of c_2 and because $c_2 \geq \frac{1}{3}$:

$$\frac{1}{3} \leq c_2(0) \leq \frac{\pi_2}{1-\pi_1^0} \leq c_2(\mu_0) = \frac{e_2(\mu_0)}{1-\pi_1^0},$$

remembering that $e_1(\mu_0) = \pi_1^0$. This implies that $\frac{1}{3}(1-\pi_1) \leq \pi_2 \leq e_2(\mu_0)$. π_2 also takes on these boundary values which can be seen by letting $\mu = 0$ and $R = \frac{\pi_1^0 - 0.25}{1-0.25}$ for the lower bound (note that $e_2(0) = 0.25$), and $\mu = \mu_0$ (implying $R(\mu) = 0$) for the upper bound. This completes the proof of Theorem 3.

Note that c_2 is in fact strictly increasing as shown below implying that the DPSDT model is identified for all $(\pi_1, \pi_2) \neq (1, 0)$. For $(\pi_1, \pi_2) = (1, 0)$, only R is identified, that is it follows that $R = 1$, whereas μ can take on any value.

Turning next to the MSDT model, set $e_1^* = e_1(\mu^*)$ and $e_2^* = e_2(\mu^*)$. It follows from the monotonicity of e_1 and from $\mu \geq \mu^*$ that $\pi_1(\mu, \lambda) \geq e_1^*$ and from $e_1(\infty) = 1$ that although $\pi_1 < 1$, π_1 can approximate the upper bound of 1 arbitrarily closely. This establishes the stated bounds for π_1.

For a fixed value of $\pi_1(\mu, \lambda)$, say π_1^0, with $\pi_1^0 > e_1^*$, it follows that $e_1(\mu) \geq \pi_1^0 > e_1^*$, hence $\mu \geq \mu_0 = e_1^{-1}(\pi_1^0) > \mu^*$, and that $\lambda = \frac{\pi_1^0 - e_1^*}{e_1(\mu) - e_1^*}$ can range from $\frac{\pi_1^0 - e_1^*}{1-e_1^*}$ to one, without taking on the lower bound, however. Solving the model equation for π_1^0 for e_1, it can be seen that $e_1(\mu) = \frac{1}{\lambda}(\pi_1^0 - (1-\lambda)e_1^*)$, hence $\mu = \mu(\lambda) = e_1^{-1}\left(\frac{1}{\lambda}(\pi_1^0 - (1-\lambda)e_1^*)\right)$ and thus, $\pi_2(\lambda) = \pi_2(\mu(\lambda), \lambda) = \lambda e_2(\mu(\lambda)) + (1-\lambda)e_2^*$.

The function $\pi_2(\lambda)$ is strictly increasing in λ. To see this, we take the derivative of $\pi_2(\lambda)$ with respect to λ. Note that the derivative of e_1^{-1}, $(e_1^{-1})'$ with respect to μ exists and is given by $\frac{1}{e_1'(\mu)}$ and that the derivative of $\mu(\lambda)$ with respect to λ is therefore $\frac{1}{e_1'(\mu(\lambda))\lambda^2}(e_1^* - \pi_1^0) = -\frac{1}{e_1'(\mu(\lambda))\lambda}(e_1(\mu(\lambda)) - e_1^*)$, because $\lambda = \frac{\pi_1^0 - e_1^*}{e_1(\mu) - e_1^*}$. Hence, the derivative of $\pi_2(\lambda)$ is

$$\begin{aligned}\pi_2'(\lambda) &= e_2(\mu(\lambda)) + \lambda \frac{e_2'(\mu(\lambda))}{e_1'(\mu(\lambda))} \frac{1}{\lambda^2}(e_1^* - \pi_1^0) - e_2^* \\ &= e_2(\mu(\lambda)) - e_2^* - \frac{e_2'(\mu(\lambda))}{e_1'(\mu(\lambda))}(e_1(\mu(\lambda)) - e_1^*) \\ &= \int_{e_1^*}^{e_1(\mu(\lambda))} \left(\frac{e_2'(e_1^{-1}(p))}{e_1'(e_1^{-1}(p))} - \frac{e_2'(\mu(\lambda))}{e_1'(\mu(\lambda))} \right) dp.\end{aligned}$$

The last equation follows because $e_2(\mu(\lambda)) - e_2^* = e_2(e_1^{-1}(e_1(\mu(\lambda)))) - e_2(e_1^{-1}(e_1^*))$ and because the derivative of $e_2(e_1^{-1}(p))$ with respect to p is $\frac{e_2'(e_1^{-1}(p))}{e_1'(e_1^{-1}(p))}$. Because $\frac{e_2'(\mu)}{e_1'(\mu)}$ is strictly decreasing in μ (as shown below), the term under the integral is positive, hence the integral is positive. It follows that $\pi_2(\lambda)$ is strictly increasing in λ. As already noted, for a given $\pi_1^0 > e_1^*$, λ can range from $\frac{\pi_1^0 - e_1^*}{1-e_1^*}$ to one, and it is easy to see that it cannot take on the lower bound although it approximates it arbitrarily closely with $\mu(\lambda)$ approaching infinity. Hence, as λ approaches the lower bound, $\pi_2(\lambda)$ approaches from above $c_2(\mu^*)(1-\pi_1)$, remembering that $e_2(\mu)$ goes to zero as μ goes to infinity. For the maximum value of λ, $\lambda = 1$, it follows that $e_2(\mu(1)) = e_2(\mu_0)$, because $\mu(\lambda) = \mu_0$ for $\lambda = 1$. Finally, for $\pi_1^0 = e_1^*$, it is easy to see that either $\lambda = 0$ or $\mu = \mu^*$, hence $\pi_1^0 = e_1^*$, whereas $\pi_2(\mu, \lambda) = e_2^*$. This completes the proof.

Note that the monotonicity of $\pi_2(\lambda)$ implies that the MSDT model is identified for all $(\pi_1, \pi_2) \neq (e_1^*, e_2^*)$. For $(\pi_1, \pi_2) = (e_1^*, e_2^*)$, either $\lambda = 0$ or $\mu = \mu^*$.

It remains to be shown that

1. for $\mu \geq 0$, $c_2(\mu) \geq \frac{1}{3}$,
2. $c_2(\mu)$ is non-decreasing in μ, and
3. $\frac{e_2'(\mu)}{e_1'(\mu)}$, is strictly decreasing in μ.

Point 1 follows from Point 2 and $c(0) = \frac{1}{3}$. For Point 2, note that $c_2 = \frac{e_2}{1-e_1} = \frac{e_2}{e_2+e_3+e_4} = \frac{1}{1+\alpha^{-1}}$ with $\alpha = \frac{e_2}{e_3+e_4}$. It is therefore sufficient to show that $\alpha(\mu)$ is non-decreasing in μ. Note that

$$\begin{aligned}\alpha(\mu) &= \frac{3\int F^2(1-F) f_{(\mu,1)}}{\int 3(1-F)^2 F + (1-F)^3 f_{(\mu,1)}} \\ &= \frac{3\int F^2(1-F) f_{(\mu,1)}}{\int (1-F)^2(1+2F) f_{(\mu,1)}}.\end{aligned}$$

Hence, the derivative of α with respect to μ is given by

$$\begin{aligned}\alpha'(\mu) = {} & \frac{3}{\left(\int (1-F)^2(1+2F)f_{(\mu,1)}\right)^2} \\ & \times \left(\int F^2(z)(1-F(z))(z-\mu)f_{(\mu,1)}(z)\,\mathrm{d}z \right. \\ & \times \int (1-F(y))^2(1+2F(y))f_{(\mu,1)}(y)\,\mathrm{d}y \\ & - \int F^2(z)(1-F(z))f_{(\mu,1)}(z)\,\mathrm{d}z \\ & \left. \times \int (1-F(y))^2(1+2F(y))(y-\mu)f_{(\mu,1)}(y)\,\mathrm{d}y \right).\end{aligned}$$

It remains to be shown that the last term in the big parentheses, $n(\mu)$, is non-negative. It is equal to

$$\begin{aligned}n(\mu) = {} & \iint F^2(z)(1-F(z))(1-F(y))^2(1+2F(y))(z-\mu) \\ & \times f_{(\mu,1)}(y)f_{(\mu,1)}(z)\,\mathrm{d}y\,\mathrm{d}z \\ & - \iint F^2(z)(1-F(z))(1-F(y))^2(1+2F(y))(y-\mu) \\ & \times f_{(\mu,1)}(y)f_{(\mu,1)}(z)\,\mathrm{d}y\,\mathrm{d}z.\end{aligned}$$

The two integrands in the upper and lower rows differ only in that $(z-\mu)$ occurs in the upper and $(y-\mu)$ in the lower. For this reason, we can drop $-\mu$. Switching the integration variables z and y in the lower row, it is seen that $n(\mu)$ equals

$$\begin{aligned}& \iint z\Big(F^2(z)(1-F(z))(1-F(y))^2(1+2F(y)) \\ & \quad - F^2(y)(1-F(y))(1-F(z))^2(1+2F(z))\Big) \\ & \quad \times f_{(\mu,1)}(y)f_{(\mu,1)}(z)\,\mathrm{d}y\,\mathrm{d}z.\end{aligned}$$

The function in the big parentheses is equal to $(F(z)-F(y))(1-F(y))(1-F(z))(F(y)+F(z)+F(z)F(y))$. Furthermore, due to the monotonicity of F, $(z-y)(F(z)-F(y)) > 0$ for $z \neq y$. Let $g(y,z) = (1-F(y))(1-F(z))(F(y)+F(z)+F(z)F(y))f_{(\mu,1)}(y)f_{(\mu,1)}(z)$ and note that $g(y,z) = g(z,y)$. Hence,

$$\begin{aligned}0 < {} & \iint (z-y)(F(z)-F(y))g(y,z)\,\mathrm{d}y\,\mathrm{d}z \\ = {} & \iint z(F(z)-F(y))g(y,z)\,\mathrm{d}y\,\mathrm{d}z \\ & - \iint y(F(z)-F(y))g(y,z)\,\mathrm{d}y\,\mathrm{d}z \\ = {} & 2n(\mu),\end{aligned}$$

the last equation following by interchanging the roles of z and y in the right integral in the row above it. Note that we have in fact shown the stronger result that c_2 is not only non-decreasing, but also strictly increasing (because its derivative is not only non-negative, but positive).

It remains to be shown that (3) $\frac{e_2'(\mu)}{e_1'(\mu)}$, is strictly decreasing in μ. To see this note that

$$e_1(\mu) = \int F^3(z)f_{(\mu,1)}(z)\,\mathrm{d}z = \int F^3(z+\mu)f(z)\,\mathrm{d}z \quad \text{and}$$

$$e_2(\mu) = 3\int (F^2(z+\mu) - F^3(z+\mu))f(z)\,\mathrm{d}z.$$

Hence,

$$\begin{aligned}e_1'(\mu) = {} & 3\int F^2(z+\mu)f(z+\mu)f(z)\,\mathrm{d}z \\ = {} & 3\int F^2(z)f(z)f_{(\mu,1)}(z)\,\mathrm{d}z \quad \text{and}\end{aligned}$$

$$e_2'(\mu) = 3\int (2F(z) - 3F^2(z))f(z)f_{(\mu,1)}(z)\,\mathrm{d}z$$

whereas

$$e_1''(\mu) = 3\int (z-\mu)F^2(z)f(z)f_{(\mu,1)}(z)\,\mathrm{d}z \quad \text{and}$$

$$e_2''(\mu) = 3\int (z-\mu)(2F(z) - 3F^2(z))f(z)f_{(\mu,1)}(z)\,\mathrm{d}z.$$

By a similar argument as for Point 2, it is sufficient to show that $m = e_2''e_1' - e_1''e_2' < 0$ for each μ. By a similar argument as for Point 2,

$$\begin{aligned}m(\mu) = {} & \iint z\Big(3(2F(z) - 3F^2(z))3F^2(y) - 3(2F(y) - 3F^2(y)) \\ & \times 3F^2(z)\Big)f(y)f_{(\mu,1)}(y)f(z)f_{(\mu,1)}(z)\,\mathrm{d}y\,\mathrm{d}z \\ = {} & 18\iint z(F(y)-F(z))F(y)F(z)f(y)f_{(\mu,1)}(y) \\ & \times f(z)f_{(\mu,1)}(z)\,\mathrm{d}y\,\mathrm{d}z.\end{aligned}$$

Because $(z-y)(F(y)-F(z)) < 0$ for $z \neq y$, it follows by a similar argument as for Point 2 that $m(\mu) < 0$ for each μ.

As pointed out by an anonymous reviewer, alternative proofs for the 2HTM, DPSDT, and MSDT could be based on the observation that these models are linear combinations of singletons and convex curves. For example, the 2HTM is a linear combination of a singleton given by $(\pi_1, \pi_2) = (1, 0)$ and the curve given by $(G(D_n), H(D_n))$. From this it follows that the model space is a cone generated by lines connecting the singleton and points on the curve. This implies the lower and upper bounds stated in Theorem 2, once it is shown that $H(G^{-1}(\pi_1))$ is a convex function. It is interesting to note that the UVSDT is also a cone with singleton $(0.5, 0)$, corresponding to $(\mu, \sigma) = (0, \infty)$, and curve generated by $\sigma = 0$ although neither the singleton itself nor the curve is in the UVSDT's model space.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Bartle, R. G. (1995). *The elements of integration and Lebesgue measure*. New York: Wiley.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564.

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*, 197–215.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: applications to recognition memory. *Psychological Review*, *116*, 84–115.

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America*, *53*, 129–160.

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

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171–189.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, *15*, 692–712.

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

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*, 304–313.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2008). The effects of unitization on familiarity-based source memory: testing a behavioral prediction derived from neuroimaging data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 730–740.

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: threshold theory or detection theory? *Journal of Experimental Psychology: General*, *127*, 83–96.

Erwe, F. (1962). *Calculus of differentiation and integration*, *Differential- und integralrechnung*. Mannheim: Hochschultaschenbücher-Verlag.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory & Language*, *62*, 1–18.

García-Pérez, M. A. (1990). A comparison of two models of performance in objective tests: finite states versus continuous ditributions. *British Journal of Mathematical and Statistical Psychology*, *43*, 73–91.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 500–513.

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

Grünwald, P. (2007). *The minimum description length principle*. Cambridge, Mass.: MIT Press.

Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design tacit: sensitivity to the structure of memory lists. *Quarterly Journal of Experimental Psychology: Section A*, *50*, 199–215.

Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., & Datey, A. V. (2009). Bridging the gap: transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 391–407.

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

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: a latent-trait approach. *Psychometrika*, *75*, 70–98.

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

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

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

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

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 380–387.

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

Mandler, G. (1980). Recognizing: the judgment of previous occurrence. *Psychological Review*, *87*, 252–271.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: an explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*, 465–494.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179.

Myung, J. I., & Pitt, M. A. (2004). Model comparison methods. *Methods in Enzymology*, *383*, 351–366.

Onyper, S., Zhang, Y., & Howard, M. W. (2010). Some-or-none recollection: evidence for item and source memory. *Journal of Experimental Psychology: General*, *139*, 341–362.

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

R Development Core Team, (2009). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*, 59–83.

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

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

Rotello, C. M., & Heit, E. (2000). Associative recognition: a case of recall-to-reject processing. *Memory & Cognition*, *28*, 907.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

Solomon, J. A. (2007a). Intrinsic uncertainty explains second responses. *Spatial Vision*, *20*, 45–60.

Solomon, J. A. (2007b). Contrast discrimination: second responses reveal the relationship between the mean and variance of visual signals. *Vision Research*, *47*, 3247–3258.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: metacognitive and presuppositional strategies. *Journal of Memory and Language*, *33*, 203–217.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, *99*, 100–117.

Swets, J., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340.

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

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: an empirical validation. *Memory & Cognition*, *32*, 1206–1220.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford: Oxford University Press.

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

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

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