Research Article

Further Evidence for Discrete-State Mediation in Recognition Memory

David Kellen, Henrik Singmann, Jan Vogt, and Karl Christoph Klauer

Albert-Ludwigs-Universität Freiburg, Freiburg i. Br., Germany

Abstract. The two high threshold model (2HTM) of recognition memory makes strong predictions regarding differences between receiver operating characteristics (ROC) functions across strength manipulations. Province and Rouder (2012) tested these predictions and showed that the 2HTM provided a better account of the data than a continuous signal detection model using an extended two-alternative forced-choice task. The present study replicates and extends Province and Rouder's findings at the level of confidence-rating responses as well as their associated response times. Model-mimicry simulations are also reported, ascertaining that the models can be well discriminated in this experimental design.

Keywords: recognition memory, mathematical models, discrete states, signal detection, thresholds

An ongoing discussion in the recognition-memory literature focuses on comparing models with continuous and discrete states, in particular the signal detection theory's (SDT) unequal-variance signal detection (UVSD) model and the two high-threshold model (2HTM; e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012). These models are usually compared by means of Receiver Operating Characteristics (ROC) data. ROCs plot the relative frequencies of "Old" responses for studied and non-studied items. ROCs can be obtained via response-bias manipulations or confidence-rating judgments. Both models are depicted in Figure 1. Figure 2 provides an example of a confidence-rating ROC.

The UVSD model assumes a continuous memory process, often termed familiarity, to describe the individuals' decisions based on memory information. Both old and new items evoke some degree of familiarity, with separate familiarity distributions for both item types. The ability to discriminate between the two kinds of items is inversely related with the overlap between the two distributions. According to UVSD, an item's familiarity is compared with an established response criterion (τ). If an item's familiarity is larger than the criterion, response "Old" is given; if the familiarity is lower than the criterion, then response "New" is given instead. The familiarity distributions are usually assumed to be Gaussian, with parameters $\{\mu_o, \sigma_o\}$ and $\{0, 1\}$ for old and new items, respectively. Responses in a confidence rating scale are produced by establishing several response criteria.

The 2HTM is a discrete-state model. It assumes that memory judgments are based on information (continuous or discrete, see Kellen & Klauer, in press; Rouder & Morey, 2009) that is *mediated* by "detect" and "guessing" states. When presented at test, an old item is detected with probability D_o, leading to an "Old" response. If the item is not detected, with probability $(1 - D_o)$, then a guessing state is entered: The status of the item is then guessed, with response "Old" occurring with probability g, and response "New" with probability $(1 - g)$. The true status of a new item is detected with probability D_n, leading to response "New." Similar to old items, when detection fails for new items (with probability $1 - D_n$), a guessing state is entered. Responses in a confidence rating scale are produced by establishing state-response mapping functions that determine how the detect (δ_o and δ_n) and guessing (γ_o and γ_n) states are mapped onto the scale.

The two models cannot be compared on the basis of *single* confidence-rating ROCs given that both can account for the ROC curvilinearity that is almost ubiquitously observed (e.g., Klauer & Kellen 2010; Malmberg, 2002). This situation has led some researchers to attempt to compare the two models on the basis of binary-response ROCs (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube et al., 2012; Kellen, Klauer, & Bröder, 2013). Despite their greater diagnostic value, other difficulties are present in comparisons based on binary-response ROCs: First, observed ROCs with a reliable shape require data from multiple study-test phases, usually leading to a small number of test trials per response-bias condition. Second, in order to reliably evaluate the shape of binary-response ROC data the response-bias manipulation needs to produce large differences in response bias, something which is not easy to accomplish (e.g., Cox & Dobbins, 2011). Third, one needs to assume that memory discriminability is unaffected by the response-bias manipulation (Rouder, Province, Swagman, & Thiele, 2014; Van Zandt, 2000). These difficulties suggest that new tests based on alternative properties are desirable.

In a recent study, Province and Rouder (2012) compared the 2HTM and the UVSD model by means of an

© 2014 Hogrefe Publishing

DOI: 10.1027/1618-3169/a000272

Revised galleys for publication clearance – not for distribution

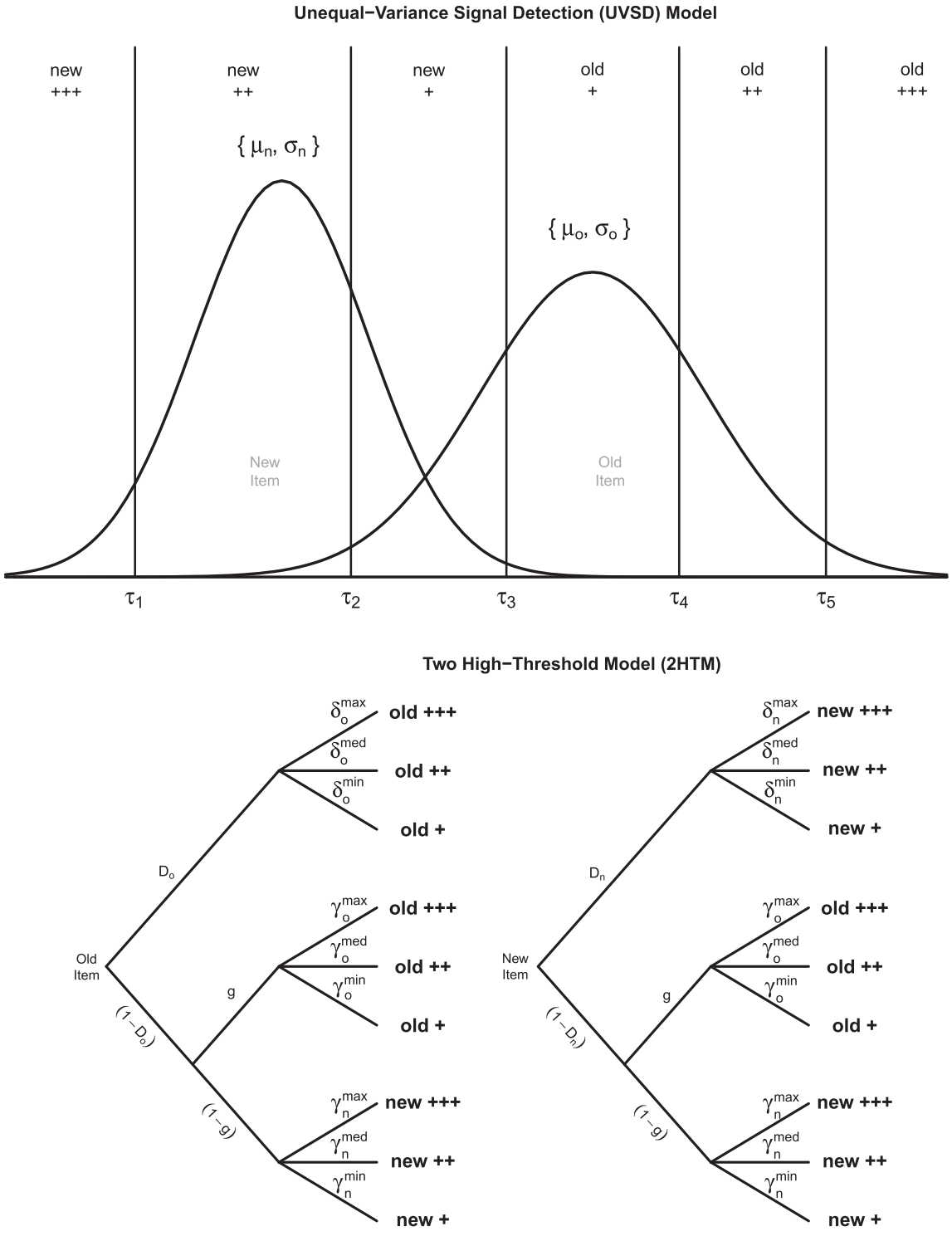

Figure 1. Graphical Representation of the 2HTM and UVSD for old/new judgments. Symbols "+," "++," and "+++" indicate minimum, medium, and maximum confidence, respectively. A description of model parameters can be found in the body of text.

unexplored property of discrete-state models – *conditional independence*. Specifically, the state-response mappings in the 2HTM are not a function of the probability of the different discrete memory states being reached, which means that study-strength manipulations should *only* affect the detection of studied items (i.e., D_o), but not the mapping of the different states on responses. In terms of the distribution of responses across a confidence-rating scale,

 © 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution

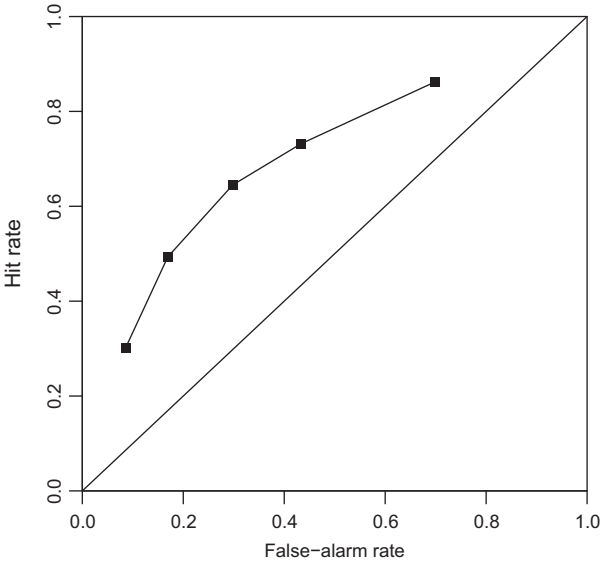

Figure 2. Example of a Receiver Operating Characteristic (ROC) function.

conditional independence implies that the component distributions (defined by the state-response mappings) should remain invariant under study-strength manipulations, with only the mixture weights (defined by the probabilities of entering the different detection states) being affected. Note that conditional independence does not introduce constraints on single confidence-rating ROCs, but across a set of ROCs. The change in the latent-mixture weights implied by conditional independence is inconsistent with the UVSD model given that the latter assumes that study-strength manipulations are captured by latent-distribution shifts.

In order to test this conditional independence, Province and Rouder (2012) used a two-alternative forced-choice task (2AFC) in which weak and strong items were paired with non-studied items. Additionally, test pairs in which both items were not studied (NEW-NEW trials) were also included. Although the testing of conditional independence does not require the use of a 2AFC task (see Rouder et al., 2014), the ability to pair different types of items in 2AFC trials facilitates it. This results from the fact that the predictions of conditional independence are imposed on the responses observed across a larger set of item types. For example, a two-level strength manipulation (weak vs. strong items) in a yes-no task results in three multinomial-distributed response vectors per individual, while in a 2AFC task it results in five response vectors.

The reported results from three experiments overwhelmingly supported the discrete-state model and conditional independence: The individual fits for the discrete-state model rarely indicated statistically significant model violations ($p < .05$) and were systematically better than the ones from a UVSD model. Finally, Province and Rouder reported an analysis of response times (RT) indicating that RT differences across the study-strength manipulation are also consistent with conditional independence.

The results reported by Province and Rouder (2012) are particularly interesting as they provide strong support (based on a new form of evidence) for a model that is traditionally considered to be based on inappropriate assumptions, namely on discrete detection and guessing states (e.g., Wixted, 2007; Yonelinas & Parks, 2007). However, the only evaluation of conditional independence available in the literature so far is the one reported by Province and Rouder. This state-of-affairs encourages attempts to replicate and extend these tests of conditional independence in order to better understand their robustness and generality (e.g., Pashler & Wagenmakers, 2012).

The present manuscript reports a replication of Province and Rouder (2012), along with important extensions: First, conditional independence is evaluated with both word and picture stimuli (in different sessions), a comparison that is important given recent results suggesting that the recognition of word and picture stimuli are associated to distinct retrieval processes (Onyper, Zhang, & Howard, 2010). Second, a manipulation of stimulus-response payoffs was introduced in order to test its selective influence on the 2HTM's state-response mapping parameters. The observation of selective influence attests to the validity of the model's characterization of the data (Schweickert, Fisher, & Sung, 2012) and introduces further constraints on the model.

The present work also extends the work of Province and Rouder (2012) in terms of the modeling analysis: First, the model comparisons are based on implementations of the 2HTM and UVSD that closely follow their postulated processes for old/new judgments (see Figure 1). As shown below, the use of different implementations is not a trivial matter and has a non-negligible impact in model performance. Second, the model fits obtained in the two sessions were further evaluated via a model-mimicry simulation (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). Model-mimicry simulations are particularly valuable in assessing the diagnosticity of individual datasets and how diagnosticity relates with observed model performance (e.g., Jang, Wixted, & Huber, 2011).

Method

Participants and Materials

Thirty-three participants (mean age = 21.42, ranging from 18 to 34) participated in individual sessions in exchange for course credits and/or a monetary prize for best performance (€50, €30, and €20 for first, second, and third place, respectively). Prize eligibility required participation in two experimental sessions (Word and Picture session) taking place at least one week part. The Word session was always the first. Thirty participants took part in both sessions.

The word stimuli consisted of 857 common German words ranging between 4 and 8 letters in length and were obtained from Lahl, Göritz, Pietrowsky, and Rosenberg (2009). According to the ratings obtained by Lahl et al., the words were all of medium valence (range: 3.5–6.5 on

© 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution

an 11-point scale) and low in arousal (range: 0.5–4.5 on an 11-point scale). The picture stimuli used in the Picture session were 525 black and white line drawings from Szekely et al. (2004).

Procedure

Participants were informed that they would participate in a memory study and that their accuracy would be used to evaluate their performance. The Word session consisted of four study-test blocks. In the study phase 82 different words were presented in a randomized order, each word being presented for 400 ms with a 200 ms inter-stimulus interval. Forty-six different studied words were presented once (weak words) and thirty-six words were studied four times (strong words). The first and last five words presented in the study list (which were presented only once) were fillers and thus not presented in the test phase. Between each study and test phase, participants engaged for 5 min in a mental arithmetic task. In each trial of the test phase (in a total of 90 trials), two items were presented side by side. The pair could include a studied item on the left (OLD-NEW), on the right (NEW-OLD), or none (NEW-NEW). Eighteen NEW-NEW trials were tested in each block. Following Province and Rouder (2012), participants were informed that there was always an old item in the pair. Features of the experimental design such as (a) the short item-presentation times and (b) the small amount of NEW-NEW trials (20% in total) served the purpose of minimizing the possibility of participants questioning the accuracy of the instructions.

In each trial, participants were requested to indicate which item was previously studied using a 6-option confidence-rating scale. Each option of the scale had a number corresponding to the number of points that could be gained in case the binary LEFT-RIGHT response (dichotomized at the scale midpoint) was correct. In case the response was incorrect an equivalent amount of points would be lost. In the low-risk scale condition the scale was [3, 2, 1, 1, 2, 3], while in the high-risk scale condition the scale was [9, 5, 1, 1, 5, 9]. Two study-test blocks were conducted in each scale condition, in a randomized but restricted order such that at least two changes of scale condition occurred for each participant.

The Picture session was virtually identical to the Word session, with the exception that it only consisted of two study-test blocks with the low-risk scale. In each block, 110 pictures were presented for 250 ms (200 ms ISI), with weak and strong pictures being presented one and three times respectively. The first and last five pictures presented were fillers. Twenty-five NEW-NEW trials (20% of the test trials) were included in each test phase.

Candidate Models

Because the models are fitted to 2AFC data (with OLD-NEW, NEW-OLD, and NEW-NEW trials), the specification of the model is somewhat different from the one depicted in Figure 1. A graphical representation of the 2HTM and UVSD model for the 2AFC task is provided in Figure 3.

The models used for the 2AFC data differ to a certain extent from the models used by Province and Rouder (2012). The comparisons reported by Province and Rouder focused on restricted cases of the SDT and 2HTM, versions which are usually found to be oversimplifications that do not account for the data at large (e.g., Yonelinas & Parks, 2007). For example, the equal-variance SDT model (when $\sigma_o = 1$) used by Province and Rouder is adequate for the symmetrical ROCs obtained in a traditional 2AFC task (where only OLD-NEW and NEW-OLD trials are considered), but cannot account for potential asymmetries between the response pattern in either the OLD-NEW or NEW-OLD trials on the one hand and the response patterns in the NEW-NEW trials on the other hand as further illustrated below. Similarly, Province and Rouder's (2012) restricted 2HTM enforced the restriction $D_n = 0$, which is also known to produce non-negligible misfits (e.g., Kellen, Klauer, & Bröder, 2013). It is important to note that Province and Rouder only focused on these models after an exhaustive comparison of different versions (Jeffrey N. Rouder, personal communication, December 5th, 2013), however these versions were not natural extensions (i.e., rely on the same parameters) of the versions most successful in accounting for old/new judgments in traditional paradigms.

The UVSD assumes that responses are produced by comparing the *difference in familiarity* between the two items in each test pair. The mean and standard deviations of the Gaussian distributions are, for each test-pair type: OLD-NEW: $\{\mu_o, \sqrt{1+\sigma_o^2}\}$, NEW-OLD: $-\mu_o, \sqrt{1+\sigma_o^2}\}$, and NEW-NEW: $\{0, \sqrt{2}\}$. Parameters μ_o^w and μ_o^s characterize the mean familiarity of weak and strong items, with σ_o as a common standard deviation parameter (see Dube et al., 2012). Thus, the model assumes that the differences between weak and strong items is captured by familiarity-distribution *shifts*. In this specification of the UVSD each distribution corresponds to the difference between the old and new-item familiarity distributions for the OLD-NEW and NEW-OLD trials, and in the case of the NEW-NEW trials, the difference between the new-item distribution with itself (see Wickens, 2002, Chap. 6).

The 2HTM assumes that in the OLD-NEW and NEW-OLD trials, individuals can independently detect the old item (with probability D_o) and the new item (with probability D_n). Parameters D_o^w and D_o^s denote the detection probabilities of weak and strong items, respectively. In the NEW-NEW trials, individuals can independently detect each of the two new items. When only one of the items is detected the other item is judged to be the studied one. When both new items are detected (with probability $D_n \times D_n$) it is assumed that individuals simply guess which item is old in the exact same way as they would when none of the items is detected. Similar to Province and Rouder (2012) the state-response mapping parameter sets δ_o, δ_n, γ_o, and γ_n were restricted in order to have a model with

 © 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution

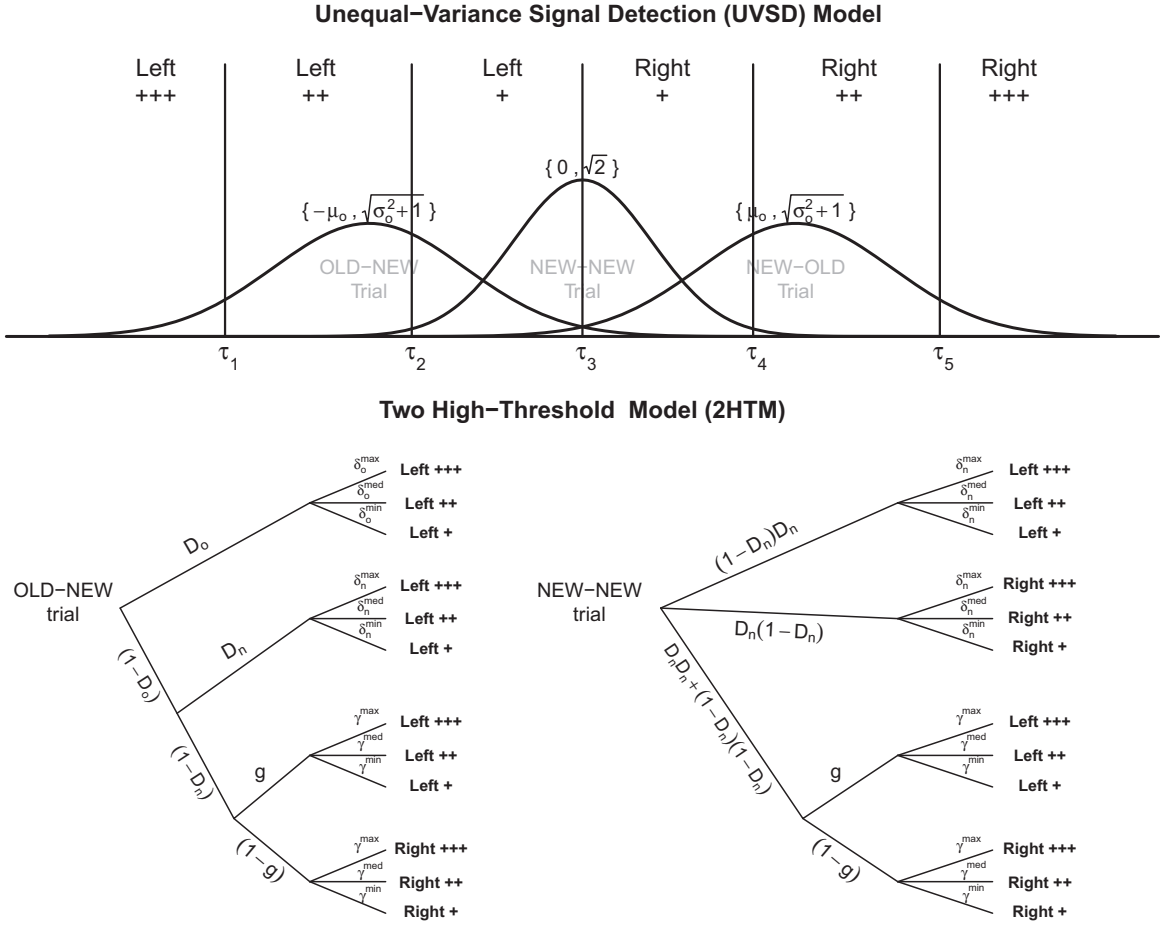

Figure 3. Graphical Representation of the 2HTM and UVSD for Province and Rouder's (2012) 2AFC task. Symbols "+," "++," and "+++" indicate minimum, medium, and maximum confidence, respectively. A description of model parameters can be found in the body of text. Note that for reasons of space, we do not depict the tree in the 2HTM concerning NEW-OLD trials. Also note that in the tree concerning NEW-NEW trials, the branches corresponding to both items being detected (with probability $Dn \times D_n$) and no item being detected (with probability $(1 - Dn) \times (1 - D_n)$) are collapsed given the assumption that they have the same state-response mapping.

the same number of free parameters as the UVSD (16 in the Word session and 8 in the Picture session). The restrictions imposed were based on the psychologically plausible response-mapping function previously used by Klauer and Kellen (2010) and the notion that confidence rating responses based on guessing are symmetrical in the 2AFC task ($\gamma = \gamma_o = \gamma_n$). The response-mapping function allows for different sets of parameters to be specified by varying a single compression parameter (e.g., the differences between the parameter sets δ_o and δ_n, and between δ_o and γ result exclusively from differences in compression parameters λ_δ and λ_γ, respectively). For reference, in the Word session the baseline 2HTM is comprised of the following *free* parameters in each of the two levels of the scale condition: Three detection parameters (D_o^w, D_o^s, D_n), one binary-response guessing parameter (g), and four state-response mapping parameters (compression parameters λ_δ, λ_γ, and set δ_o). Models were fitted to the data using MPTinR (Singmann & Kellen, 2013). Details on the R scripts and the response-mapping function are provided in the Electronic Supplementary Materials 2–9.

Results

Scale Manipulation

In the Word session, the baseline models (without restrictions across scale conditions) provided a generally good account of the individual data, with 2HTM and UVSD producing statistically significant misfits in 12% and 24% of the individual datasets, respectively. The effect of the scaling manipulation was tested via significance testing on different sets of parameter restrictions (Riefer & Batchelder, 1988). Restricting memory parameters to be equal across

© 2014 Hogrefe Publishing

Revised galleys for publication clearance – not for distribution

Table 1. Goodness of fit results and parameter estimates

Model	Word session			Picture session		
	Goodness of fit					
	G^2	$p < .05$	Win	G^2	$p < .05$	Win
	1435.56	15%	67%	530.26	10%	67%
	Mean parameters estimates					
2HTM	$D_o^w = .19$	$D_o^s = .46$ $g_H = .50$ $g_L = .53$ $\delta_o^H = \{.06, .13, .81\}$ $\delta_o^L = \{.07, .14, .79\}$ $\delta_n^H = \{.53, .21, .26\}$ $\delta_n^L = \{.55, .17, .28\}$ $\gamma^H = \{.59, .23, .18\}$ $\gamma_L = \{.57, .22, .21\}$	$D_n = .15$	$D_o^w = .28$	$D_o^s = .61$ $g = .53$ $\delta_o = \{.07, .15, .78\}$ $\delta_n = \{.60, .16, .24\}$ $\gamma = \{.75, .16, .09\}$	$D_n = .21$
	Goodness of fit					
	G^2	$p < .05$	Win	G^2	$p < .05$	Win
	1564.89	18%	33%	614.29	20%	33%
	Mean parameters estimates					
UVSD	$\mu_o^w = 0.77$	$\mu_o^s = 1.65$	$\sigma_o = 1.68$	$\mu_o^w = 1.28$	$\mu_o^s = 2.53$	$\sigma_o = 1.89$
	$\tau^H = \{-2.17, -1.36, 0.03, 1.36, 2.18\}$			$\tau = \{-2.62, -1.80, 0.12, 1.87, 2.65\}$		
	$\tau^L = \{-2.12, -1.35, 0.10, 1.27, 2.05\}$					

Note. G^2 corresponds to the summed goodness-of-fit results; $p < .05$ corresponds to the percentage of individual datasets with statistically significant misfits; ''win'' corresponds to percentage of individuals datasets for which this model provided the best fit. $\delta = \{\delta^{min}, \delta^{med}, \delta^{max}\}$ and $\gamma = \{\gamma^{min}, \gamma^{med}, \gamma^{max}\}$ are restricted by a compression function, reducing the number of free parameters (for additional details, see the Supplemental Material). Superscripts *H* and *L* correspond to the high- and low-risk scale conditions, respectively.

the scale manipulation seldom led to statistically significant misfits (6% and 9% of individual datasets for the 2HTM and UVSD, respectively). Also the summed misfits ($\Delta G^2(99) = 106.49$, $p = .29$ and $\Delta G^2(99) = 123.04$, $p = .05$, respectively) of both models failed to reach statistical significance, but approached significance for the UVSD. Still none was larger than the critical χ^2 value (χ^2_{crit}) of 146.88 obtained in a compromise power analysis ($\alpha = \beta$) when assuming a small effect size ($\omega = 0.1$; see Faul, Erdfelder, Lang, & Buchner, 2007). In contrast, the subsequent restriction on parameters governing scale usage (i.e., state-response mapping and response-criteria parameters) frequently led to significant misfits (51% and 39% of individual datasets for the 2HTM and UVSD, respectively). The summed misfits ($\Delta G^2(165) = 412.62$ and 383.60, respectively) were statistically significant and considerably larger than the compromise critical χ^2 value ($\chi^2_{crit} = 215.57$).

Mean parameter estimates of the memory-restricted models are reported in Table 1. For both models there are no visible differences in the mean response-mapping parameter estimates as a function of high-risk (*H*) versus low-risk (*L*) scale, which can be seen as somewhat at odds with the statistical tests on parameter restrictions reported above. These seemingly contradictory findings are explained by the fact that the direction of the effect of risk manipulation depends on each individual's *risk attitudes*. This contrasts with more traditional manipulations such as response-bias manipulations (e.g., Bröder & Schütz, 2009) where a consistent directional effect is expected (e.g., encouraging the response ''Old'' should lead to an increase of the latter's frequency). For example, Schmidt and Traub (2002) implemented a non-parametric method for assessing risk attitudes based on decisions involving potential gains and losses of the same magnitude (similar to the payoff schemes used here) and found that 33% of the participants were classified as ''loss averse,'' 24% as ''risk seeking'' and most participants (42%) exhibited ''loss neutrality.'' As a consequence of these individual differences in risk attitudes, effects average out at the level of mean parameter estimates, whereas they affect the parameter-restriction tests at the individual level reported above as summed misfits.

The overall results suggest that the scale manipulation had a selective influence on the parameters governing scale usage (although its direction varied between participants with presumably different risk attitudes). This result holds for both models, suggesting that the parameters of two the models are attempting to capture the same processes (see also Bröder, Kellen, Schütz, & Rohrmeier, 2013). Because of these results, the model evaluation and comparison reported below is made under the restriction that

 © 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution

memory parameters are equal across the scale manipulation (although the outcome of the comparisons reported below does not hinge on such restriction), a restriction that could be maintained empirically.

Model Comparison

Because the two candidate models have the same number of parameters, the comparison will be based on the models' goodness of fit results (quantified by the G^2 statistic). This approach is equivalent to the use of popular model-selection indices such as Akaike and Bayesian information criteria (AIC and BIC) because the latter rely on the number of parameters to quantify model flexibility. As shown in Table 1, the 2HTM provided a better account of the individual data in comparison to the UVSD, a result that was found to be systematic across individual datasets with a Wilcoxon test in both the Word session ($V = 131$, $p = .007$) and the Picture session ($V = 120$, $p = .02$). These differences are also found if one compares the joint goodness-of-fit of the Word and Picture sessions for the 30 individuals that took part in both sessions (73% of the participants were better described by the 2HTM; $V = 388$, $p < .001$). The superior performance of the 2HTM was also observed when analyzing the aggregate data: The G^2 values of the 2HTM and the UVSD were, respectively, 60.72 and 102.68 in the Word session ($\chi^2_{crit} = 79.27$), and 15.30 and 75.27 in the Picture session ($\chi^2_{crit} = 42.23$). Taken together, the present results replicate the preference for the discrete-state model reported by Province and Rouder (2012). Confidence-rating histograms for selected participants are presented in Figure 4.

One important aspect of the data is the form of the different ROCs. As shown in panels A and B of Figure 5, the aggregate-data ROCs for OLD-NEW (as ''signal'' trials) and NEW-OLD trials (as ''noise'' trials) are curvilinear and symmetrical, consistent with the equal-variance SDT model usually adopted for 2AFC data. However, the ROCs for OLD-NEW (as ''signal'' trials) and NEW-NEW trials (as ''noise'' trials; panels C and D) show a considerable asymmetry. The equal-variance SDT model adopted by Province and Rouder (2012) is unable to account for this asymmetry, potentially leading to severely inflated misfits. In the present data the restriction $\sigma_o = 1$ was rejected in 70% and 83% of the individual datasets in the Word and Picture sessions. However, the considerable improvements in fit brought by establishing σ_o as a free parameter are insufficient for the UVSD to provide a better account than the 2HTM. This result shows that the better performance of the 2HTM originally reported by Province and Rouder does not hinge on this unaccounted for asymmetry, and holds across different model specifications.

Given the improvements in model fit brought by setting parameter σ_o free, we also checked whether the use of different standard-deviation parameters for weak and strong items (σ_o^w and σ_o^s, respectively) would produce further improvements in the UVSD model, and whether these improvements affect the comparison with the 2HTM. In the Word session this extension of the UVSD model led to somewhat ambiguous results as the restriction $\sigma_o^w = \sigma_o^s$ is only rejected in 36% of the individual datasets. The average parameters of the extended UVSD model were $\mu_o^w = 0.75$, $\sigma_o^w = 1.49$, $\mu_o^s = 1.75$, and $\sigma_o^s = 1.95$. Regarding the comparison with the 2HTM, the latter still outperforms the extended UVSD model (summed $G^2 = 1456.83$), being preferred in 64% of individual datasets ($V = 171$, $p = .05$) according to AIC, and 85% ($V = 62$, $p < .001$) according to BIC. Regarding the Picture session, the restriction $\sigma_o^w = \sigma_o^s$ is rejected in only 20% of the individuals. The extended model's average parameters were $\mu_o^w = 1.25$, $\sigma_o^w = 1.70$, $\mu_o^s = 2.70$, and $\sigma_o^s = 2.20$. Again, the 2HTM outperforms the extended UVSD (summed $G^2 = 541.86$) in terms of AIC (63%, $V = 121$, $p = .02$) and BIC (90%, $V = 25$, $p < .001$). This preference is stronger when comparing fits to both sessions jointly (AIC: 83%; $V = 421$, $p < .001$, and BIC: 97%; $V = 464$, $p < .001$). The extended SDT model will not be considered any further.

Model Flexibility and Mimicry

It is well known in the model-selection literature that goodness-of-fit results do not fully characterize model performance as they overlook differences in *model flexibility* (e.g., Kellen et al., 2013). A successful model is one that provides a good account of the data because it is accurately characterizes the major underlying processes. A successful model is not a model that fits well data in general, regardless of the latter's origin. This notion encourages the search for the model that strikes the best tradeoff between goodness of fit and flexibility (Kellen et al., 2013). A related concept is *model mimicry*, which concerns the ability to discriminate between two models on the basis of given data (Wagenmakers et al., 2004; Jang et al., 2011).

An accurate quantification of the flexibility of the present models represents a formidable challenge that is beyond the scope of this paper (see Kellen et al., 2013). Still, much can be learned by means of the model-mimicry analysis method developed by Wagenmakers et al. (2004) as it evaluates the ability to discriminate between two models and detect asymmetries in mimicry in each individual dataset. The method developed by Wagenmakers et al. (2004) is implemented as follows: (1) A non-parametric bootstrap sample is obtained from a given individual dataset. (2) The bootstrapped data are then fitted by the two models, and their respective parameter estimates are in turn used to generate two parametric bootstrap samples (one from each model). (3) The sample generated by each model is then fitted by the two models and the resulting fit results are compared. (4) This process is repeated several times (in the present case 1,000 times) for each individual dataset.

The model-mimicry simulation provides an assessment of how well each model fits data depending on the model generating the latter. For example, in a case of high mimicry the fits of the two models will be very similar, irrespective of which of the two is the data-generating model. The model-mimicry simulation results can be evaluated in two different ways: One way consists of evaluating

© 2014 Hogrefe Publishing

Revised galleys for publication clearance – not for distribution

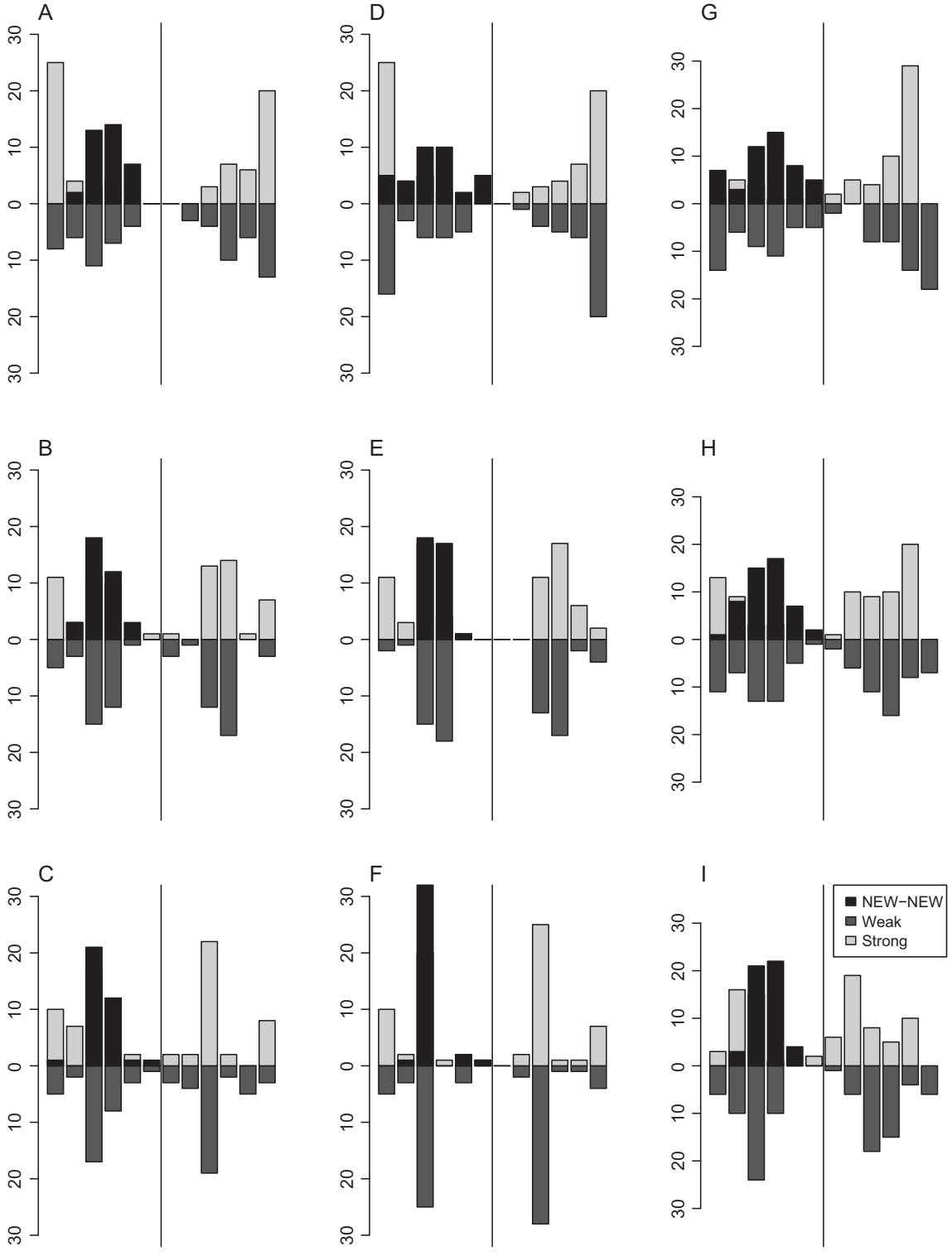

Figure 4. Histograms of confidence ratings across selected participants. Panels A–C: Individual datasets with, in order, the lowest, median, and maximum $G^2_{\text{2HTM}} - G^2_{\text{UVSD}}$ values in the Word session, high-risk scale condition. Panels D–F: Word session, low-risk scale condition. Panels G–I: Picture session. In each panel, responses to OLD-NEW/NEW-OLD trials are presented to the left/right of the vertical midline. Responses to NEW-NEW trials are presented to the left of the vertical midline

© 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution

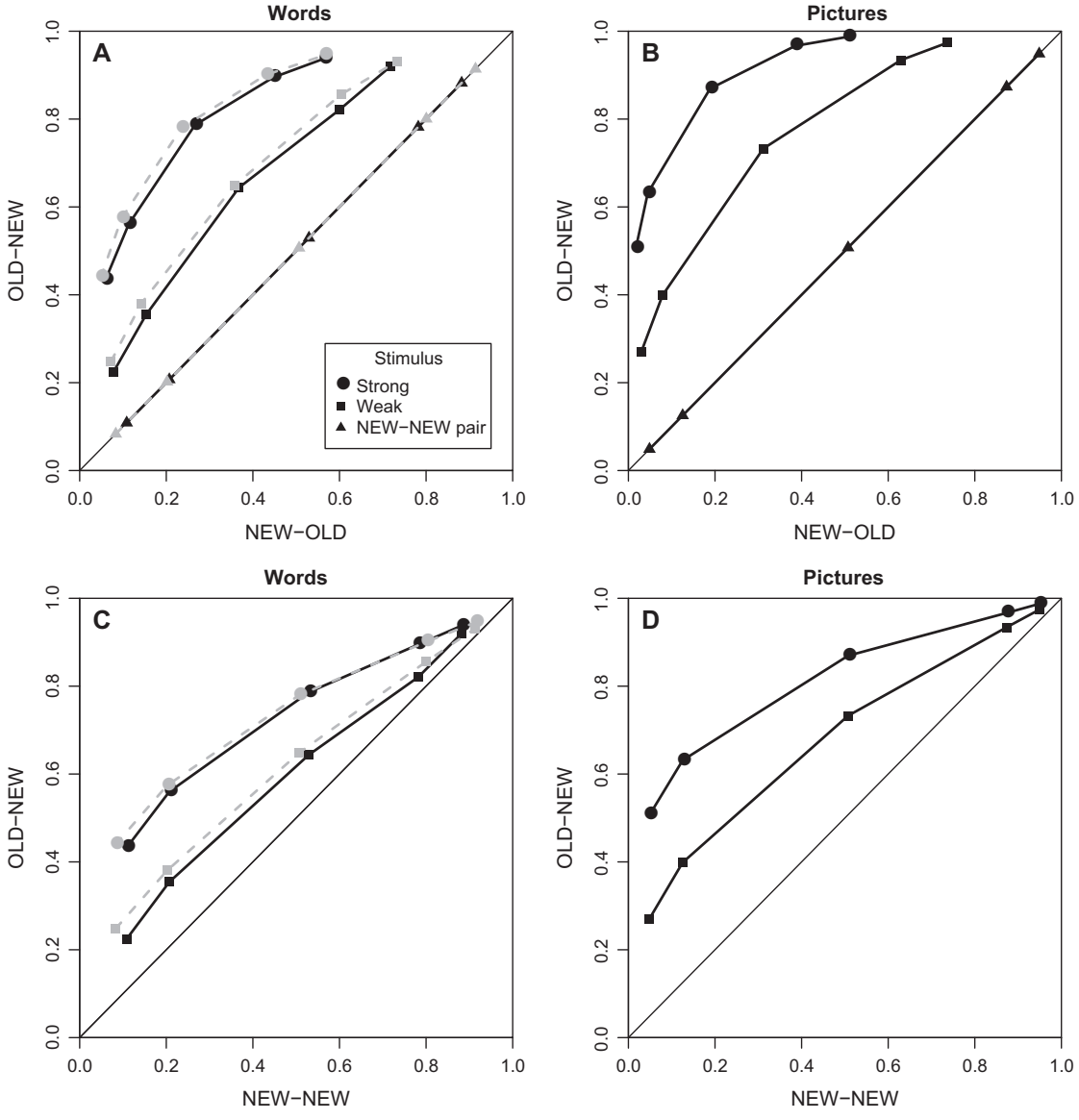

Figure 5. Receiver operating characteristic (ROC) functions from the Word and Picture sessions (Aggregate Data). The grey points correspond to the high-risk scale condition and the black points to the low-risk scale condition. In panels A and B the cumulative response proportions to NEW-NEW pairs are plotted on both axes for reference purposes.

the ability to distinguish both models in terms of summed individual fit results. Another way consists of evaluating the ability to distinguish both models for each individual dataset and relating this ability to model performance (Jang et al., 2011).

The results of the model-mimicry analysis are summarized in Figure 6. The two top panels show that in terms of summed individual fits the two models can be well discriminated from each other as recovery of the data-generating model was virtually perfect. No consistent differences in model mimicry were found as the median absolute summed G^2 differences ($|\Sigma G^2_{\text{2HTM}} - G^2_{\text{UVSD}}|$) in the Word session was somewhat larger when the UVSD was the data-generating model (432.42) than when the 2HTM generated the data (403.78) but a small difference in the opposite direction was observed in the Picture session (235.86 vs. 242.52). However, the observed G^2 differences were found to be very unlikely under both models (as according to these models, larger differences are expected). This discrepancy reflects the fact that both models being compared are mere approximations to an unknown data-generating process, models with several restrictions being imposed (e.g., $\sigma^w_o = \sigma^s_o$, or the response-mapping restrictions).

The average individual recovery rates (across both models) shown in the bottom panel of Figure 6 were in general very high. Moreover, the individual datasets that were better fitted by the 2HTM were associated with higher

© 2014 Hogrefe Publishing

Revised galleys for publication clearance – not for distribution

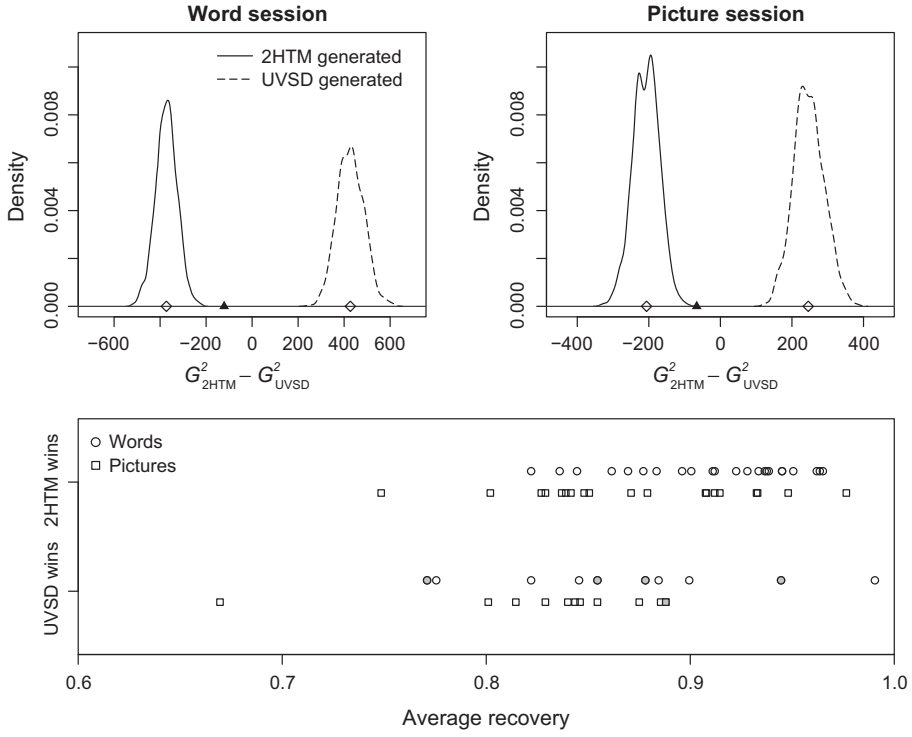

Figure 6. Model-mimicry simulation results. The top panels present the difference in summed individual fits for the two models. The empty diamonds indicate the median of each distribution and the black triangles correspond to the observed difference in summed individual fits. The bottom panel presents the average recovery rate of each individual dataset. Labels ''2HTM wins'' and ''UVSD wins'' indicate the individual datasets that were better fitted by the 2HTM and UVSD, respectively. The grey circles and squares correspond to the datasets whose best-performing model changes when using the estimated mimicry penalties.

recovery rates than the datasets better fitted by the UVSD ($r = 0.37$, and $r = 0.43$ for the Word session and the Picture session respectively, both $p < .05$). This pattern in individual recovery is somewhat suggestive of a greater flexibility of the UVSD, as model flexibility is known to have a greater impact in cases where the data are less diagnostic (e.g., Kellen et al., 2013; Myung, 2000). Still, note that mimicry simulations do not provide a precise assessment of model flexibility (see Wagenmakers et al., 2004, pp. 40–42), and therefore do not dismiss the need of a principled quantification of the latter (Kellen et al., 2013).

Following Wagenmakers et al. (2004) and Jang et al. (2011), the cutoff value that optimizes the recovery rate was computed for each individual dataset. The computed optimal cutoffs produced modest improvements in the average recovery rate (an average 2% improvement in both sessions) which is not surprising given that most recovery rates were already quite high. Nevertheless the use of optimal cutoffs in model selection was not inconsequential as it accentuates the preference for the 2HTM, which is then found to be best model in 79% and 70% of the individual datasets in the Word and Picture sessions respectively, in comparison to the 67% reported in Table 1. This difference results from six individual datasets that were now better accounted by the 2HTM (four in the Word session and two in the Picture session) and one individual dataset better accounted by the UVSD (Picture session). R scripts implementing the model mimicry analysis can be found in Electronic Supplementary Materials 3, 4, 6, 10, and 11.

Response Times

RTs decreased as study strength increased: In the Word session, mean RTs for the NEW-NEW pairs (no word was studied), weak pairs (old word studied once), and strong pairs (old word studied four times) were, in order, 2,733 ms, 2,601 ms, and 2,400 ms. This pattern was found in both high and low-risk scale conditions. A within-subjects ANOVA on mean RTs with ''study strength'' and ''scale condition'' as factors (using Greenhouse-Geisser corrected degrees of freedom) revealed only a significant effect of ''study strength'' $F(1.43, 45.75) = 23.08$, $\eta^2_G = .22$, $p < .001$. No other effect (or interaction) was found to be statistically significant (largest $F = 2.70$, $\eta^2_G = .03$, $p = .11$). In the Picture condition, the same effect was also found, $F(1.33, 38.52) = 27.21$, $\eta^2_G = .48$, $p < .001$.

Conditional independence in the 2HTM predicts that the speed of responses produced by the different states is independent of the probability of those states being entered.

© 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution

Accordingly, RT differences observed across test pairs result from differentially weighted mixtures of detection and guessing states. In order to evaluate conditional independence at the level of RTs, it is necessary to identify which RTs are more likely to result from responses produced by the guessing and detection states. This was achieved by calculating the conditional probability that a response was produced by a certain state, given (a) the type of trial, (b) the response, and (c) individual parameter estimates. Given that we are interested in the effect of the study-strength manipulation (which is expected to selectively affect D_o), we focus on the distinction between the detect-old state and the detect-new/guessing states. Figure 7 depicts the individual mean RTs with conditional detection-probability dichotomized at the .50 value; as can be seen, the overall mean RTs for the different states replicate the pattern reported by Province and Rouder (2012, Figure 3C). Next, these conditional probabilities will be used as a covariate in a linear mixed-model (LMM) analysis (Baayen, Davidson, & Bates, 2008). According to the 2HTM's conditional-independence prediction, *conditional detection-probability* should be able to account for the RT differences across study-strength conditions.

The 2HTM's account was evaluated by means of a single LMM on the individual RTs of the complete dataset (i.e., Word and Picture session) with both (1) the *study-strength* manipulation and (2) the *conditional detection-probability* that an RT was produced by the detect-old state specified as fixed effects. To account for the differences in mean RTs between the two sessions we estimated (a) a fixed effect for session as well as the interactions of this fixed effect with study-strength and conditional detection-probability. Furthermore, we estimated the maximal-random effects structure (Barr, Levy, Scheepers, & Tily, 2013) to account for the systematic variance introduced by the participants: A random participant intercept and a (crossed) random intercept for the Participant × Session interaction. For both random intercepts we also estimated random slopes for study-strength and conditional detection probability. The first of the random effects terms allowed individuals to have an idiosyncratic RT mean and idiosyncratic effects of study-strength and conditional detection-probability whereas the latter allowed the effect of session and the difference in study-strength and

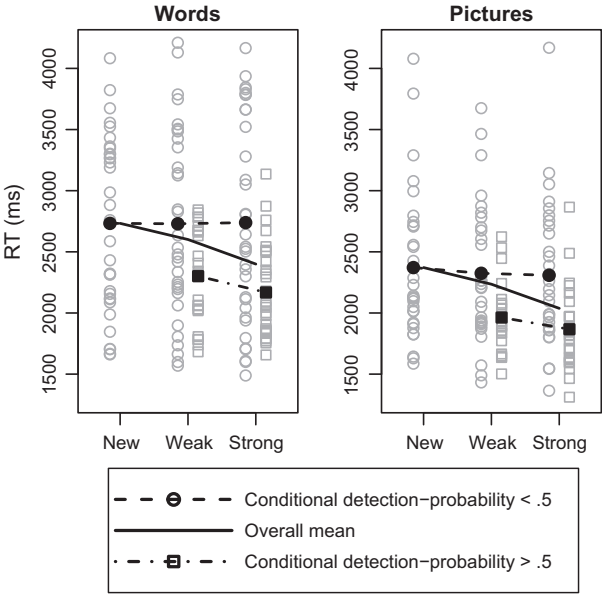

Figure 7. Individual mean RTs (nonfilled symbols) and their mean (filled symbols) of trials with conditional detection-probability < .50 and > .50. In addition, the overall mean RTs are depicted by a solid straight line (the individual overall RT means are not depicted).

conditional detection-probability between sessions to vary idiosyncratically across participants (see Electronic Supplementary Material 1 for more details).

To assess the statistical significance of fixed effects, we compared by means of likelihood-ratio tests the full model against restricted models in which one effect of interest was excluded. Results of the LMM, displayed in Table 2, were in line with the 2HTM's prediction of conditional independence. The effect of conditional detection-probability, when controlling for the effect of study-strength, was significant, $\chi^2 = 24.89$, $p = .002$. This effect was also large, as the difference in RT between conditional detection-probabilities of 0 and 1 was 648.68 ms. In contrast, when controlling for the effect of the conditional detection-probability, the effect of study-strength was non significant, $\chi^2 = 0.57$, $p = .74$, and near zero. The estimated difference between

Table 2. LMM analysis of RTs

Effect	Estimate	χ^2	*Δdf*	*p*
(Intercept)	2556.17	91.70	1	< .0001
Study-strength	[25.26]	0.57	2	.75/.74
Conditional detection-probability	−648.67	24.89	1	< .0001/.002
Session	206.91	14.23	1	.0002
Session × study-strength	[21.01]	0.57	2	.75/.78
Session × conditional detection-probability	−68.13	1.20	1	.27

Notes. Estimates are in milliseconds (ms). The estimates of the study-strength effects (in squared brackets []) correspond to the estimated maximal difference (i.e., between NEW-NEW trials and strong OLD-NEW/NEW-OLD trials). The *p* values reported after the "/" are based on 500 parametric-bootstrap samples. Further details can be found in the body of text and the Electronic Supplementary Material 1.

© 2014 Hogrefe Publishing

Revised galleys for publication clearance – not for distribution

the NEW-NEW word trials and trials including a strong word is only 25.26 ms. Furthermore, neither the interaction of session with study-strength nor with conditional detection-probability reached significance, indicating that, as expected, the conditional detection-probability alone captured the relevant variance.

Discussion

The present results replicate the findings reported by Province and Rouder (2012) at the level of confidence-rating responses and RTs, but also extend them in several ways. First, conditional independence was tested across word and picture stimuli. Furthermore, the models used were specified in a way that naturally extends the versions most successful in fitting old/new judgments in traditional paradigms, in particular in fitting implied ROC asymmetries. Moreover, the validity of the compared models received experimental support via the selective influence of a payoff manipulation on response-mapping/response-criteria. Finally, a model-mimicry simulation showed that both models can be well discriminated with these data, with simulation results running counter the possibility that the results are due to greater flexibility of the 2HTM relative to the UVSD. Altogether, the 2HTM's conditional-independence property provides a good account of the data.

Although the present data conform to the constraints imposed by conditional independence, it should be noted that there are cases in which this property is expected to be violated. These expected violations do not represent a failure of the 2HTM but merely its *underspecification*. Note that the 2HTM is a model of item recognition but also a core component of a larger model dealing with item and source memory (Kinchla, 1994; Klauer & Kellen, 2010), the two high-threshold source model (2HTSM). Conditional on old-item detection, the 2HTSM assumes that the source (or study context) of the item is remembered with probability d, and not remembered with probability $(1 - d)$. The 2HTSM assumes a state where only item memory is available, and another state where both item and source memory are available. These two states, which have distinct state-response mapping functions (Klauer & Kellen, 2010) are "aggregated" into a single memory state by the 2HTM (Kinchla, 1994). If instead of simple study-repetition manipulations, one manipulates item memory in a way that emphasizes contextual differences in the study phase (e.g., using different "levels of processing" in the study phase; Craik & Lockhart, 1972) then the different response mapping functions of the 2HTSM are likely to produce violations of conditional independence.

Discrete-state modeling is often dismissed in the literature in favor of continuous models such as SDT. The results reported show that discrete-state mediation can provide an extremely good account of complex and diagnostic data on recognition memory. We hope that these results will encourage researchers to consider the benefits of including discrete-state modeling in their data-analysis toolboxes along with continuous modeling approaches.

Acknowledgments

David Kellen and Henrik Singmann contributed equally to this manuscript. The research reported in this paper was supported by grant Kl 614/32-1 from the Deutsche Forschungsgemeinschaft to Karl Christoph Klauer. We thank Andrew Heathcote and Jeff Rouder for valuable comments.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at http://dx.doi.org/10.1027/1618-3169/a000272

ESM1: modeling_details.pdf

Provides detailed description of the 2HTM's state-response mapping function as well as details on the linear mixed modeling of the RTs.

ESM2: analysis-script.R

R script that performs the analyses reported in the manuscript and produces all values reported in the text, Table 1, and Figures 4 and 6. Requires files ESM3 to ESM 9 to run successfully.

ESM3: words-data.txt

Provides the data of the word session.

ESM4: pics-data.txt

Provides the data of the picture session.

ESM5: words.sdt.model

Model file containing the UVSD model for the word session.

ESM6: pics.sdt.model

Model file containing the UVSD model for the picture session.

ESM7: mpt-models.R

R file containing the 2HT models for word and picture session.

© 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution

ESM8: mm.words.rda

Provides the G^2 values obtained in the mimicry analysis of the word session.

ESM9: mm.pics.rda

Provides the G^2 values obtained in the mimicry analysis of the picture session.

ESM10: words.mimicry.r

R script that performs the model mimicry analysis for the word session. Requires ESM 3, ESM 5, and ESM 7 to run.

ESM11: pics.mimicry.r

R script that performs the model mimicry analysis for the picture session. Requires ESM 4, ESM 6, and ESM 7 to run.

References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. doi: 10.1016/j.jml.2012.11.001

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high threshold model for confidence rating data in recognition memory. *Memory, 8*, 916–944. doi: 10.1080/09658211.2013.767348

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. doi: 10.1037/a0015279

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities standard, distractor-free, and target-free recognition. *Memory & Cognition, 39*, 925–940. doi: 10.3758/s13421-011-0090-3

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684. doi: 10.1016/S0022-5371(72)80001-X

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

Dube, C., Starns, J. J., Rotello, C. M., & Ratliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and Language, 67*, 389–406. doi: 10.1016/j.jml.2012.06.002

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. doi: 10.3758/BF03193146

Jang, Y., Wixted, J. T., & Huber, D. E. (2011). The diagnosticity of individual data for model selection: Comparing signal-detection models of recognition memory. *Psychonomic Bulletin & Review, 18*, 751–757. doi: 10.3758/s13423-011-0096-7

Kellen, D., & Klauer, K. C. (in press). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review, 20*, 693–719. doi: 10.3758/s13423-013-0407-2

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

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. doi: 10.3758/PBR.17.4.465

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

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

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*, 190–204. doi: 10.1006/jmps.1999.1283

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. doi: 10.1037/a0018926

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530. doi: 10.1177/1745691612465253

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

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*, 318–339. doi: 10.1037/0033-295X.95.3.318

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

Rouder, J. N., Province, J. M., Swagman, A. R., & Thiele, J. E. (2014). *From ROC curves to psychological theory*. Manuscript submitted for publication.

Schmidt, U., & Traub, S. (2002). An experimental test of loss aversion. *Journal of Risk and Uncertainty, 25*, 233–249. doi: 10.1023/A:1020923921649

Schweickert, R., Fisher, D. L., & Sung, K. (2012). *Discovering cognitive architecture by selectively influencing mental processes*. New Jersey, NJ: World Scientific.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models with R. *Behavior Research Methods, 45*, 560–575. doi: 10.3758/s1342801202590

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., ... Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory*

© 2014 Hogrefe Publishing

Revised galleys for publication clearance – not for distribution

and Language, 51, 247–250. doi: http://dx.doi.org/10.1016/j.jml.2004.03.002

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

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*, 28–50. doi: 10.1016/j.jmp.2003.11.004

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

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

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

Received February 27, 2014
Revision received May 7, 2014
Accepted May 9, 2014
Published online July X, 2014

David Kellen

Institut für Psychologie
Albert-Ludwigs-Universität Freiburg
79085 Freiburg i. Br.
Germany
E-mail david.kellen@psychologie.uni.freiburg.de

© 2014 Hogrefe Publishing
Revised galleys for publication clearance – not for distribution