MEMORY, 2013
http://dx.doi.org/10.1080/09658211.2013.767348

Validating a two-high-threshold measurement model for confidence rating data in recognition

Arndt Bröder[1], David Kellen[2], Julia Schütz[3], and Constanze Rohrmeier[3]

[1]School of Social Sciences, Department of Psychology, University of Mannheim, Mannheim, Germany
[2]Department of Psychology, University of Freiburg, Freiburg, Germany
[3]Department of Psychology, University of Bonn, Bonn, Germany

Signal Detection models as well as the Two-High-Threshold model (2HTM) have been used successfully as measurement models in recognition tasks to disentangle memory performance and response biases. A popular method in recognition memory is to elicit confidence judgements about the presumed old/new status of an item, allowing for the easy construction of ROCs. Since the 2HTM assumes fewer latent memory states than response options are available in confidence ratings, the 2HTM has to be extended by a mapping function which models individual rating scale usage. Unpublished data from 2 experiments in Bröder and Schütz (2009) validate the core memory parameters of the model, and 3 new experiments show that the response mapping parameters are selectively affected by manipulations intended to affect rating scale use, and this is independent of overall old/new bias. Comparisons with SDT show that both models behave similarly, a case that highlights the notion that both modelling approaches can be valuable (and complementary) elements in a researcher's toolbox.

Keywords: Recognition memory; Signal detection; Threshold models; Multinomial modelling.

In recognition memory tests, previously presented (“old”) items have to be distinguished from “new” items that had not been presented before. A less-than-perfect memory performance results in different kinds of errors, namely “misses” (denoting old items as new) and “false alarms” (denoting new items as old). As early as 1909, Schulze acknowledged that “it is hard to reconcile both types of errors under one common viewpoint” (Schulze, 1909, p. 188, translated from German by AB). Since not only memory performance, but also response strategies under uncertainty determine the correct and wrong responses, measurement models have been developed which strive to disentangle both processes and to provide process pure measures of memory performance which are corrected for strategic guessing or response biases. Two prominent models for recognition memory are the Signal Detection Theory model (SDT; Green & Swets, 1966) and the Two-High-Threshold model (2HTM; Snodgrass & Corwin, 1988). For a recent debate on the relative appropriateness of these model classes in recognition memory, see Bröder and Schütz (2009), Dube and Rotello (2012), Kellen and Klauer (2011), Klauer and Kellen (2010, 2011a, 2011b), and Province and Rouder (2012).

Although both models (outlined later) entail quite different psychological assumptions, the

Address correspondence to: Arndt Bröder, University of Mannheim, Schloss Ehrenhof Ost, D-68131 Mannheim, Germany. E-mail: broeder@uni-mannheim.de

This research was supported by the Deutsche Forschungsgemeinschaft (DFG), grants BR 2130/3-1 and BR 2130/7-1.

conclusions obtained from both methods are strikingly similar (Bröder & Schütz, 2009; Klauer & Kellen, 2010, 2011a), and systematic construct validation studies have identified both methods as valid measurement tools in standard recognition tests (Snodgrass & Corwin, 1988). However, the 2HTM's description of memory and decision processes is not readily applicable to confidence rating data which have gained popularity in recognition tests. Here, participants provide graded ratings about how confident they are concerning each of their "old" or "new" responses. Whereas SDT handles these data by establishing multiple decision criteria along a memory strength dimension, the 2HTM has to be supplemented by a state–response mapping function which models the translation of memory states onto the confidence scale, taking into account individually varying response styles. We present such an extended model in the spirit of ideas by Erdfelder and Buchner (1998), Klauer and Kellen (2010, 2011a), and Malmberg (2002). We report two unpublished data sets and three new experiments which demonstrate the validity of the model in showing that (1) the response format (binary or confidence rating) does not affect the core model parameters for estimating sensitivity and bias and (2) the response mapping parameters adequately capture experimentally manipulated response styles, leaving the estimates of core parameters largely unaffected. In this respect, we follow the tradition of parameter validation studies across different response modalities that has also been used in SDT research (e.g., Markowitz & Swets, 1967; Swets, 1959).

Note that the model presented here is not intended to compete with process models of recognition that aim at describing the processes and dynamics of memory retrieval (e.g., Malmberg, 2008; Ratcliff & Starns, 2009). Rather, it is a measurement model, intended to provide useful indices to quantify memory performance and response bias for researchers striving to disentangle both processes successfully while using confidence judgements as a preferred response mode.

This manuscript is organised as follows: First, we will distinguish between measurement models and process models. Second, we briefly introduce SDT and the 2HTM as well as the extension of the latter to confidence ratings. Third, we will analyse two unpublished experimental conditions from Bröder and Schütz (2009), demonstrating the equivalence of the parameter estimates from binary and rating responses. Fourth, we report three new experiments with different learning materials which manipulated the old/new response bias as well as the presumed scale use of participants, showing that only the model parameters representing these processes are affected. Finally, we discuss the results and their implications for "continuous" and "discrete" views of (re)cognition.

An important issue should be made clear at this point: The present manuscript is primarily concerned with the validation of the 2HTM. Comparisons with SDT in this context only serve to demonstrate the similarity of the accounts provided by both models, and not to pit each model against each other in a model-selection competition. A sensible model-selection analysis should rely on data that are diagnostic for the distinction of model assumptions (e.g., Province & Rouder, 2012) as it hinges on critical properties emerging from these assumptions (Birnbaum, 2011; Roberts & Pashler, 2000; see also Jang, Wixted, & Huber, 2011), as well as on adequate model-selection indices (e.g., Klauer & Kellen, 2011b). The data reported in the present work are suitable for the validation of the models—observe how parameter estimates differ across experimental conditions—but not necessarily for a direct competition between the models: First, there is no critical property of any of the models than can be observed in the data (both models predict curvilinear ROCs). Second, the common model-selection indices such as the Akaike and Bayesian Information Criteria (AIC and BIC; Burnham & Anderson, 2002) are known to be problematic as they fail to capture differences in model flexibility due to functional form (see Klauer & Kellen, 2011b). Kellen and Klauer (2011) and Klauer and Kellen (2011b) have shown that the 2HTM is considerably less flexible than the SDT model for different tasks and data (e.g., binary-response ROCs), although AIC and BIC consider both models to be equally flexible (in these tasks, the models have the same number of parameters). Also, BIC has been shown to grossly overpenalise models with larger number of parameters (see Jang et al., 2009; Klauer & Kellen, 2011b). We will therefore not place any emphasis on AIC and BIC results, although we report them for reference purposes.

MEASUREMENT MODELS VERSUS PROCESS MODELS

Formal modelling in cognitive psychology serves two interrelated, but distinct goals which we characterise as *epistemic* and *pragmatic*, respectively. Epistemic models strive at formulating process theories which adequately describe the cognitive processes and representations underlying encoding, storage, and retrieval in memory tasks. For example, so-called global memory models formalise assumptions about representations (e.g., traces as feature vectors; Flexser & Tulving, 1978) and retrieval processes (e.g., activation of traces resulting from vector comparisons defined by some similarity metric), which often lead to specific qualitative or even quantitative predictions about recognition performance. Examples for such models are REM (Shiffrin & Steyvers, 1997, 1998), SAM (Raaijmakers & Shiffrin, 1980, 1981), MINERVA (Brandt, 2007; Hintzman, 1984), and TODAM (Murdock, 1982, 1997, 2006a, 2006b), or the dynamic REM-framework proposed by Malmberg (2008). These formal models are theories about the structure and mechanics underlying memory, and they are evaluated according to the usual standards of theory testing: Their predictions are tested in empirical investigations, and the goal is to sort out the theory which explains memory phenomena most successfully. A theory like Ratcliff's diffusion model (Ratcliff, 1978; Ratcliff & Starns, 2009) may be viewed as a mid-level theory that models the general aspect of retrieval dynamics as evidence accumulation without committing itself to assumptions about a specific memory architecture, or representational format of memory traces.

In contrast, the *pragmatic* goal of modelling operates at the "surface" level of empirical investigations in trying to derive useful measures of memory performance from empirical observations. Since it is obvious that observed hit rates (HRs) and false alarm rates (FARs) are influenced by nonmnemonic strategic factors (response biases), these measurement models strive to describe the interplay between the output of some mnemonic process and a response strategy and to map the result on the data space. Hence, it is a miscategorisation to view models like SDT or the 2HTM as memory models. They are rather more general decision theories formalising how the output of some memory process is integrated with strategic processes to form memory judgements. Hence, they formalise how the output of some memory process is mapped onto observable responses. We guess that some controversies about the appropriateness of different measurement models result from the lack of acknowledging the distinction between the epistemic and the pragmatic modelling goals. Since measurement models differ in the way they conceptualise the *output* of the mnemonic processes (e.g., continuous or discrete), eventually identifying a superior measurement model is not inconsequential for underlying process theories.

Of course, measurement models also have to be scrutinised for their validity. This construct validation entails (1) their ability to describe actual data (by means of model fit) and (2) the demonstrations that parameters intended to reflect a cognitive process are selectively affected by manipulations targeting at the process they represent. In the following section, we will introduce SDT and the 2HTM as potential measurement models as well as 2HTM's extension to confidence data.

SDT, THE 2HTM, AND THE EXTENDED 2HTM

According to SDT, the output of the memory system when confronted with a test stimulus entails a single value on a continuous decision dimension which can be termed "memory strength" or sometimes "familiarity". The probability distributions of such values for old and new items are assumed to be Gaussian, with old items eliciting higher memory strength values on average (see Figure 1). The curves typically overlap, and judgements are therefore uncertain to some degree, implying occasional errors (false alarms and misses). The participant has to place a decision criterion that divides the continuum into an "old" and a "new" response region, respectively. The placement of the criterion is determined by various strategic influences. This model can be easily extended to capture confidence ratings as well if it is assumed that the expressed degrees of confidence directly reflect memory strength (but see Province & Rouder, 2012; Rouder, Pratte, & Morey, 2010). In this case, it is simply assumed that the participant responding on a k-bins rating scale places $k-1$ criteria along the memory strength continuum which determine the use of response bins given a

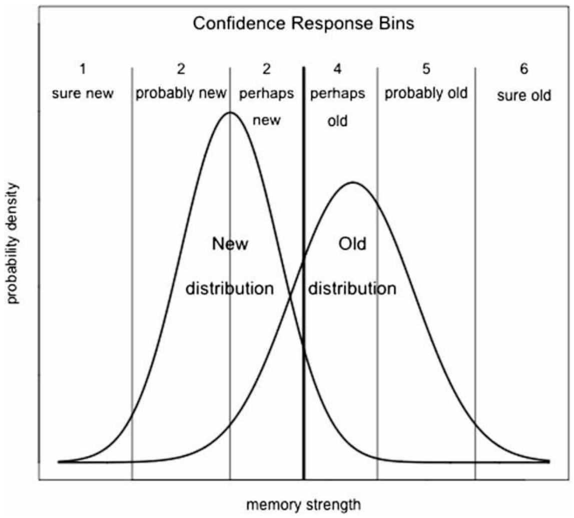

Figure 1. Signal Detection model of recognition for confidence rating data. Each rating bin is delimited by criteria placed on the memory strength decision variable.

specific strength value elicited by a test item. Note that this reasoning necessarily assumes that confidence ratings are merely graded ''old/new'' judgements, an assumption that is rarely made explicit and which is taken for granted by SDT theorists.

In contrast to SDT, the 2HTM entails the assumption that the cognitive representation of the memory system's output is discrete (Figure 2). A participant may reach one of two detect states: An old item may pass a threshold to be detected as old with probability p_o. A distractor may also be identified as such with probability p_n. Many processes may contribute to the latter state, for example an extreme lack of familiarity, several unsuccessful retrieval attempts, the mismatch of an item to the gist of a list, or other metacognitive assessments (''I would remember *this* word if it had been in the list''; see Strack & Bless, 1994). Hence, many reasons and arguments may contribute to a firm and valid distractor rejection. In fact, Ratcliff's (1978) diffusion model as one of the most prominent process models of recognition assumes an active sampling of information which speaks *against* the item being old. Information for or against an item being old is integrated over time in a diffusion process that eventually hits either an ''old'' threshold or a ''new'' threshold. With respect to the detect states, the 2HTM can be viewed as a simple measurement variant of the diffusion model's accuracy data with the additional constraint that new items never cross the old threshold, and old ones never reach the new boundary.

If an item passes neither of both detection thresholds, the participant is in a state of uncertainty, and she has to guess. The guessing probability b reflects a tendency to guess ''old'' and is determined by the participants' response strategy which may be affected by expectations, payoffs, or the expected base rate of old items in the test. In summary, the model assumes that three discrete memory states can result in a recognition test.

Rouder and Morey (2009) elaborated upon the difference between continuous models like SDT and models based on a classical Fechnerian conception of a threshold. The psychological difference in case of recognition models may be

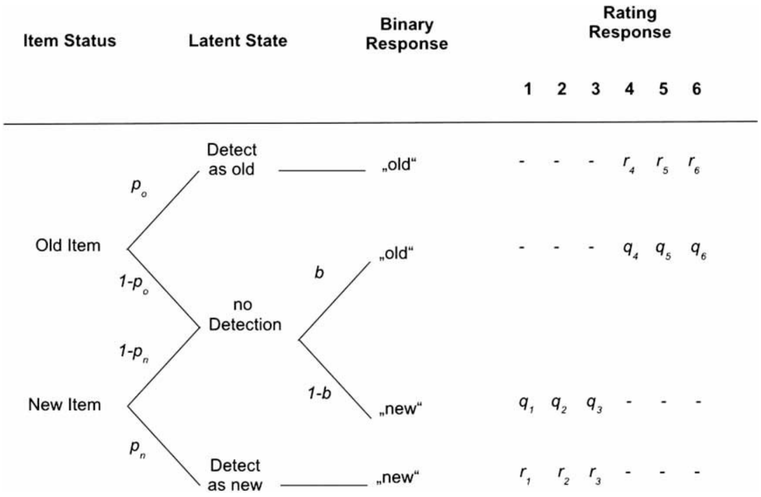

Figure 2. The original Two-High-Threshold model for binary responses is depicted as a processing tree on the left. The parameters on the right side model the response distributions across a 6-point confidence scale in detect states (r_i) and the uncertain state (q_i).

illustrated as follows: In case of SDT, the memory system's output is a continuous evidence strength variable to which the participant has conscious access entirely. This output value is compared to the criterion (or different criteria in the case of ratings) as an internal standard, and a response is generated. Hence, each response is influenced by a specific value of the graded internal variable. In contrast, discrete-state models assume that the evidence strength variable is divided by thresholds into regions that form equivalence classes of strength values with respect to the response made. Although the underlying strength variable may be continuous, the participant only has conscious access to the region in which a value falls which in case of the 2HTM is experienced as a "detect new", a "detect old", or an "uncertain" state. Different evidence values within one region are functionally equivalent with respect to the response distribution. Recently, Province and Rouder (2012) tested this bold prediction following from discrete-state models as applied to confidence rating responses: If the uncertain state is strictly uncertain without a further differentiation of memory strengths within the state, response distributions on a rating scale should show *conditional independence*, meaning that the response distributions across the rating scale that result from a certain mnemonic state should be identical, regardless how this state was reached. Province and Rouder manipulated mnemonic strength of items and estimated the response distribution of the guessing state in a 2AFC task involving pairs of lures (that imply a guessing state). For the majority of participants, conditional independence held, and actual response distributions could be well described as mixtures from the latent states. The discrete-state approach outperformed latent strength competitors as well as a dual process alternative model.

Formally, both kinds of measurement models (2HTM and SDT) therefore assume different response distributions conditional on internal strength values. Psychologically, this reflects different assumptions about the degree of precision in the conscious accessibility of the internal strength variable. Several process models of recognition memory explicitly assume "privileged" regions of evidence strength that lead to confident an quick correct "old" and "new" judgements (Atkinson & Juola, 1974; Malmberg, 2008; Ratcliff, 1978) which may be seen functionally as detect states.

Extension of the 2HTM to confidence rating data

SDT's assumption that the observer has conscious access to a graded internal memory signal is often supplemented by the idea that graded confidence judgements directly reflect this memory strength. Hence, confidence response bins are defined by multiple response criteria aligned in a monotonically increasing fashion along the evidence variable. Note, that SDT per se does not make any predictions about the relative spacing of the criteria or the width of the bins except for this ordering. Nevertheless, the ordering implies the often-observed intraindividual correlation between confidence and accuracy (Mickes, Wixted, & Wais, 2007). Note that the assumption of graded responses reflecting graded memory strength has long been questioned (Krantz, 1969).

Unlike assumed for SDT, an extension of the 2HTM to account for confidence judgements obtained with a k-bin rating scale is neither trivial nor model inherent. In fact, the specification of how latent states are mapped onto confidence rating judgements can be done in infinitely many ways (e.g., Klauer & Kellen, 2010). Since in most cases $k > 3$, one has to make assumptions about how the three latent states are mapped onto a finer graded response scale. It is well known that scale-usage in terms of giving extreme or moderate rating-responses are associated to personality traits and response styles (Hamilton, 1968; Tourangeau, Rips, & Rasinksi, 2000). Furthermore, other factors such as intraindividual variations (e.g., Haubensak, 1992), sequential dependencies (Malmberg & Annis, 2012), or even simple random errors (Rieskamp, 2008) also affect the mapping of the latent states onto observed rating responses.

Additionally, task characteristics or instructions as well as the exact wording of the confidence bins may caution participants from using extreme responses or encourage them to do so. Arbitrary anchoring values affect the use of scales (Schwarz, Knauper, Hippler, & Neumann, 1991). Also, demand characteristics may play a role: If the researcher asks participants to use the whole range of the scale, they may do so for strategic reasons rather than express their actual inner states (even if they were able to do so). The scale itself may be a strong demand cue to provide graded responses.

To summarise, there may be many variables besides memory strength possibly affecting the response distribution across a confidence rating scale. We would consider these influences as *nuisance variables* reflecting response style rather than memory or strategic guessing. In a model with only three distinct memory states and $k > 3$ rating bins, a mapping function must be formulated that allows the capture of these nuisance influences without affecting measures for memory performance. We therefore extend the original 2HTM—which we will denote as the core model—with response mapping parameters which we denote as the response model (see Figure 2, for an example extended to a 6-point rating scale). This extension is in line with ideas formerly expressed by Erdfelder and Buchner (1998) as well as by Klauer and Kellen (2010, 2011), but alternative implementations could be conceived.

In a nutshell, the core model still determines which half of the rating scale is used (corresponding to ‘‘old'' and ‘‘new'' judgements in the binary case). Hence, collapsing the halves of the rating scale to binary old/new judgements yields the original 2HTM. However, in each of the memory states, free parameters allow for idiosyncratic spreading of responses across the response bins. In the detect states, parameters r_i denote the probability of using response bin i in the detect old and detect new states. Note that the detect-old state is restricted to responses within the ‘‘old'' range of the scale, whereas the detect-new state is restricted to the range of ‘‘new'' responses. In the uncertainty state, however, recognition-memory judgements are based on guessing so all responses are possible, parameterised by probabilities q_i for choosing response i. The model equations for the core model and the extended model for six rating bins are provided in Table 1.

The equations of the extended model do not restrict the distributions of responses across the scale in either memory state with the exception that the rating response in a given detect state is consistent with that state's binary response. At this point there are no further theoretical assumptions restricting the model as we assume that the response mapping is determined by many nonmnemonic nuisance variables like strategies, response preferences, or personality. Also, we share Rouder et al.'s (2010) scepticism about graded judgements being valid indicators of latent variables. However, despite the many nonmnemonic influences on ratings, there will probably be a correlation between the memory state and the degree of confidence. If there is at least some correlation of confidence and accuracy, one would—ceteris paribus—expect more extreme ratings on average resulting from detect states rather than the uncertain state (e.g., Klauer & Kellen, 2010; Province & Rouder, 2012), which implies parameter inequalities such as $r_1 > q_1$ and

TABLE 1

Model equations of the original 2HTM, the extended model for 6-point rating scales, and the binary reparameterisation of the extended model

Original 2HTM	*Extended 2HTM*	*Binary reparameterisation*
$p(\text{«old»}\mid\text{old}) = p_o + (1 - p_o)*b$	$p(\text{«6»}\mid\text{old}) = p_o*r_6 + (1 - p_o)*b*q_6$	$p(\text{«6»}\mid\text{old}) = p_o*r'_6 + (1 - p_o)*b*q'_6$
$p(\text{«old»}\mid\text{new}) = (1 - p_n)*b$	$p(\text{«5»}\mid\text{old}) = p_o*r_5 + (1 - p_o)*b*q_5$	$p(\text{«5»}\mid\text{old}) = p_o*(1 - r'_6)*r_5 + (1 - p_o)*b*(1 - q'_6)*q'_5$ [b]
	$p(\text{«4»}\mid\text{old}) = p_o*r_4 + (1 - p_o)*b*q_4$[a]	$p(\text{«4»}\mid\text{old}) = p_o*(1 - r'_6)*(1 - r_5) + (1 - p_o)*b*(1 - q'_6)*(1 - q'_5)$
	$p(\text{«3»}\mid\text{old}) = (1 - p_o)*(1 - b)*q_3$	$p(\text{«3»}\mid\text{old}) = (1 - p_o)*(1 - b)*(1 - q'_1)*(1 - q'_2)$
	$p(\text{«2»}\mid\text{old}) = (1 - p_o)*(1 - b)*q_2$	$p(\text{«2»}\mid\text{old}) = (1 - p_o)*(1 - b)*(1 - q'_1)*q'_2$
	$p(\text{«1»}\mid\text{old}) = (1 - p_o)*(1 - b)*q_1$	$p(\text{«1»}\mid\text{old}) = (1 - p_o)*(1 - b)*q'_1$
	$p(\text{«6»}\mid\text{new}) = (1 - p_n)*b*q_6$	$p(\text{«6»}\mid\text{new}) = (1 - p_n)*b*q'_6$
	$p(\text{«5»}\mid\text{new}) = (1 - p_n)*b*q_5$	$p(\text{«5»}\mid\text{new}) = (1 - p_n)*b*(1 - q'_6)*q'_5$
	$p(\text{«4»}\mid\text{new}) = (1 - p_n)*b*q_4$	$p(\text{«4»}\mid\text{new}) = (1 - p_n)*b*(1 - q'_6)*(1 - q'_5)$
	$p(\text{«3»}\mid\text{new}) = p_n*r_3 + (1 - p_o)*(1 - b)*q_3$	$p(\text{«3»}\mid\text{new}) = p_n*(1 - r'_1)*(1 - r'_2) + (1 - p_n)*(1 - b)*(1 - q'_1)*(1 - q'_2)$
	$p(\text{«2»}\mid\text{new}) = p_n*r_2 + (1 - p_n)*(1 - b)*q_2$	$p(\text{«2»}\mid\text{new}) = p_n*(1 - r'_1)*r'_2 + (1 - p_n)*(1 - b)*(1 - q'_1)*q'_2$
	$p(\text{«1»}\mid\text{new}) = p_n*r_1 + (1 - p_n)*(1 - b)*q_1$	$p(\text{«1»}\mid\text{new}) = p_n*r'_1 + (1 - p_n)*(1 - b)*q'_1$

[a] $r_4 = (1 - r_6 - r_5)$, $q_4 = (1 - q_6 - q_5)$, $r_3 = (1 - r_1 - r_2)$ and $q_3 = (1 - q_1 - q_2)$. [b] The original parameters can be calculated from the estimates of the reparameterised model in the following way: $r_5 = (1 - r'_6)*r'_5$, $r_4 = (1 - r'_6)*(1 - r'_5)$, $q_5 = (1 - q'_6)*q'_5$, $q_4 = (1 - q'_6)*(1 - q'_5)$, $r_2 = (1 - r'_1)*r'_2$, $r_3 = (1 - r'_1)*(1 - r'_2)$, $q_2 = (1 - q'_1)*q'_2$, and $q_3 = (1 - q'_1)*(1 - q'_2)$.

$r_6 > q_6$. However, the model does not make further assumptions about the response distribution in the different states. Due to the many idiosyncratic factors potentially affecting scale usage, we simply do not know how participants map the different latent states onto the scale, and it is unclear if a monotonic relationship between memory strength can be assumed.

Altogether, the unrestricted model has three core parameters (p_o, p_n, b) and $2*(k-2)$ freely varying response mapping parameters q_i and r_i, resulting in, for example, 11 free parameters for a response scale with $k=6$. With only $2*(k-1) = 10$ free categories, the model is thus oversaturated and nonidentifiable. In order to overcome these problems, parameter restrictions can be imposed: For example, response mapping can be restricted to be conditionally symmetric for the middle categories (e.g., $r'_3 = r'_4$ and $q'_3 = q'_4$; see Table 1 and Appendix B). Hence, the extension of the 2HTM to rating data with the accompanying increase in data *df*s has two advantages: First, a testable model is generated which does not trivially fit the data. Second, separate estimates of p_o and p_n are possible. In Experiments with only one HR-FAR pair, p_o and p_n have to be assumed to be equal a priori in order to achieve identifiability (see Bröder & Schütz, 2009). This restriction might represent a valid assumption in some cases (see Snodgrass & Corwin, 1988, for a justification), but this is not necessarily the case. With the extended rating model, the assumption becomes a testable hypothesis. Third, the correlation between accuracy and confidence that is normally observed (e.g., Mickes, Wais, & Wixted, 2009) can be accounted by an extended 2HTM if one simply relaxes the assumption that participants invariably and deterministically map detect states onto maximum-confidence responses.

In a now-classic paper dealing with measurement models for the *binary* response format, Snodgrass and Corwin (1988) compared various models and their respective indices of memory performance and bias with respect to measurement validity. Their set contained SDT (with different bias measures and distributions), the 2HTM, and the nonparametric performance measure A' in combination with the bias measure B'' (Grier, 1971). Interestingly, the latter measures as well as SDT's bias measure β failed an initial test of independence (in the sense that a decrease in the discrimination parameter also decreases the range of values attainable by the bias measure) and were not further considered.[1] Only SDT (with the location criterion parameter as bias measure) and the 2HTM passed the test of mutually independent indices of performance and bias. In two subsequent studies, both SDT's and 2HTM's memory measures were affected by mnemonic variables (imageability of words, diagnostic group), and the bias measures were affected by different payoffs. 2HTM's parameters even showed superior sensitivity to the manipulations in comparison to the parameters of SDT. Snodgrass and Corwin therefore conclude that both models qualify as valid measurement models in recognition, showing convergent as well as discriminant validity of the parameters. The positive results have been replicated for the 2HTM that had been extended to source memory paradigms (Bayen, Murnane, & Erdfelder, 1996; Meiser & Bröder, 2002).

However, it is not self-evident that the 2HTM retains its good performance as a measurement model when extended to the confidence rating procedure. The processes introduced by using this response format might be different from those captured by the nuisance parameters, and core parameters of the model might be compromised. Hence, the validity of the extended model has to be demonstrated. First, we use two older data sets to demonstrate that parameter estimates from the rating and the binary procedure are comparable. Second, we report three additional experiments to demonstrate that a variable intended to affect scale use was reflected in the response mapping parameters exclusively.

BINARY AND RATING DATA COMPARED: ADDITIONAL DATA FROM BRÖDER AND SCHÜTZ (2009)

Bröder and Schütz (2009) conducted two experiments in which the response bias of participants was manipulated experimentally between subjects by changing the base rates of old items in the memory test in five steps (10%, 30%, 50%, 70%,

[1] Concerning the "nonparametric" measures, Snodgrass and Corwin (1988, p. 40) go on: "It has sometimes been assumed that the distribution-free model is preferable on the grounds that it does not make assumptions about the form and nature of the process underlying generation of hits and false alarms. However, because it determines a unique isomemory and isobias function for any possible H, FA pair, we believe it constitutes as strong a model of recognition memory as any other theoretical approach."

90% in Experiment 1 and 10%, 25%, 50%, 75%, and 90% in Experiment 2). Participants responded with a binary old/new format in these experiments. However, in both experiments, additional randomised groups (not reported in the original paper) received an identical study phase and a test with 50% old items. These groups responded on a 6-point confidence rating scale extending from 1 = "very sure new" to 6 = "very sure old". Hence, the only difference between these confidence rating groups and the 50% bias manipulation group was the response format, whereas the structure of the learning phase and the retention interval were identical. Consequently, if the Confidence-2HTM is a proper extension of the original 2HTM for binary responses, one would expect similar estimates of the core parameters that capture memory processes and the old/new bias. The equivalence of the parameters was tested by simultaneously estimating the model parameters from all conditions.

Data sets

In Experiments 1 and 2 of Bröder and Schütz (2009), participants were presented with a list of 60 words or 150 simple line drawings of objects, respectively. After filler tasks (3 minutes and 20 minutes, respectively), they received a recognition test containing 60 words (Experiment 1) or 150 line drawings (Experiment 2). In the five bias manipulation groups (with $n = 15$ each and $n = 10$ each in Experiments 1 and 2, respectively), participants were truthfully informed that a certain percentage of items in the test were "old", ranging from 10% to 90%. They responded "old" or "new" to each stimulus. A sixth (randomised) rating condition group ($n = 14$ and $n = 9$, respectively) received the 50% test, but they had to denote each item on a 6-point confidence scale. Further details about the materials and the procedure are reported in Bröder and Schütz.

Data analysis and results

Both experiments' raw frequencies of ratings can be obtained from Table A1 in Appendix A. First, the rating conditions were analysed by fitting the Confidence-2HTM to each of the two data sets using MPTinR (Singmann & Kellen, in press) and MultiTree (Moshagen, 2010). For these analyses, we assumed symmetric use of the confidence scale in the state of uncertainty conditional on old/new guessing (i.e., $q'_1 = q'_6$, $q'_2 = q'_5$ and $q'_3 = q'_4$; see Table 1). However, this symmetry assumption is not plausible for the detect states since confidence judgements are metacognitive assessments which may be subject to justification processes. Hence, "detect old" and "detect new" states may lead to different rating scale use because they rely on different information. Hence, we let r_1 and r_6 unrestricted.[2] The model has eight free parameters with 10 freely varying data categories, resulting in 2 *df*. The model fitted the data of both experiments very well ($G^2_{(2)} = 2.83$ and $G^2_{(2)} = 1.05$, respectively, both $p > .20$). The right column of Figure 3 shows the estimates of the response mapping parameters r_i and q_i. As can be easily seen, the extreme ratings "1" and "6" were used in the vast majority of cases in "detect" states in both experiments (parameters r_6 and r_1), whereas the middle categories (r_2 to r_5) were used much less frequently in these detect states. In contrast, responses were more evenly scattered across the scale in the uncertain state (q_1 to q_6), although there still is some overall preference for avoiding the middle categories. The response distributions clearly differ in both memory states (both $\Delta G^2_{(2)} > 59.42$, $p < .001$, if the restrictions $r_1 = q_1$ and $r_6 = q_6$ are introduced), which clearly makes sense psychologically.

The more interesting analysis, however, entails all experimental conditions of both experiments and compares the core memory and bias parameters across response formats (see left column of Figure 3). Since the whole learning and testing procedure was identical and groups were truly randomised, we would expect similar estimates of the respective measures. The bias parameter b of the rating condition should be equivalent to the bias parameter of the neutral bias condition in the binary response format (since there were also

[2] Setting both parameters equal leads to a significantly worse model fit in both cases. To achieve an identifiable model, the r parameters cannot be completely unrestricted. We analysed a stochastically equivalent reparameterised binary tree model with $r_1 = r_1'$, $r_2 = (1-r_1')*r_2'$, $r_3 = (1-r_1')*(1-r_2')$, $r_6 = r_6'$, $r_5 = (1-r_6')*r_5'$, and $r_4 = (1-r_6')*(1-r_5')$, see Table 1. This reparameterisation leaves the number of parameters unchanged. In this model, we introduced the restriction $r_3' = r_4'$, which yields an identifiable submodel. Parameter estimates of the original model are obtained by applying the reparameterisation equations.

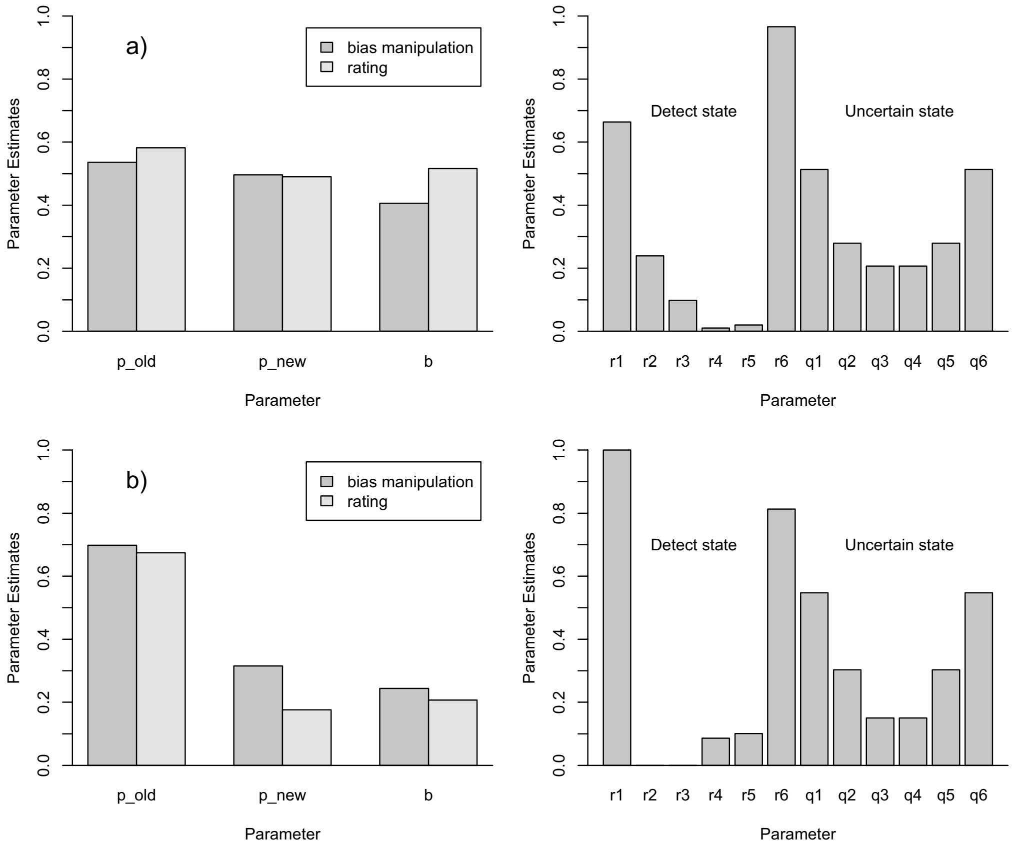

Figure 3. Parameter estimates of two experiments from Bröder and Schütz (2009). The left column depicts comparisons between estimates from binary conditions with bias manipulation and respective rating conditions in Experiments 1 and 2, respectively. The right column depicts the response distribution parameters of the rating conditions, conditional on the detect vs. uncertain states.

50% old items in this condition). The memory parameters p_o and p_n for reaching the detection states should also be equivalent to the estimates from the binary conditions. The five binary conditions and the rating model were analysed simultaneously with one supermodel which entailed two parameters for detect states in the binary conditions and the rating condition, respectively (p_{ob}, p_{nb}, p_{or}, p_{nr}), five different bias parameters b_{ib} in the bias manipulation conditions, one bias parameter b_r for the rating condition, and the response mapping parameters with the same restrictions as detailed earlier, leaving 15 parameters with 20 free data categories, hence, $df = 5$. The model fitted the data very well in both experiments ($\Delta G^2_{(5)}$>5.39 and $G^2_{(5)}$= 6.51, respectively, both p >.25). The left column of Figure 3 shows the core parameter estimates from the binary and rating conditions which converge almost perfectly. If their difference is tested via equality constraints, neither difference reaches any conventional significance level (largest $\Delta G^2_{(2)}$< 3.31, p = .13).

Discussion

The Confidence-2HTM is well able to fit the two data sets from a standard recognition paradigm with a confidence rating response format. The mapping parameters which reflect scale use in different memory states show sensible behaviour: The tendency to use the endpoints of the scale is clearly different in "detect" and "uncertain" states. However, although the vast majority of responses in detect states map to the extreme ends of the scale, a sizeable minority of responses also falls into the less extreme rating bins. As has been shown before (Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Malmberg, 2002), this response behaviour can elicit curved ROCs with confidence-based data, even if the underlying memory states are discrete. More importantly, the memory and bias measures derived from the extended Confidence-2HTM are equal to those elicited with a binary response format under identical learning and testing conditions. Hence, the response mapping parameters appear to adequately capture the differences that are induced by a different judgement method, leaving the measures for memory and strategic performance untouched.

The results obtained with the unpublished data by Bröder and Schütz (2009) indicate that an extended measurement model provides the same core parameter estimates as its already well-validated counterpart for binary data (see Bröder & Schütz, 2009; Snodgrass & Corwin, 1988). Still, further demonstrations of the measurement model's construct validity are desirable. This includes the demonstration that certain parameters can be intentionally controlled by manipulating variables which are expected to affect them selectively. We will present such a demonstration: Three further experiments were conducted which intended to affect the response mapping parameters by a different labelling of the confidence scale endpoints.

AFFECTING THE RESPONSE MAPPING PARAMETERS: THREE EXPERIMENTS

In all experiments, an 8-point confidence rating scale was used in order to provide more ''space'' for the intended variation in scale use. The Confidence-2HTM can easily be extended to this situation by adding parameters r_i and q_i ($i = 7$ and 8), leaving the core parameters untouched. In both experiments, we strove to manipulate the caution with which participants use the endpoints of the rating scale by a different labelling of the extreme categories (see later). If the model is valid and if the response mapping parameters reflect only response style of using a rating scale, this manipulation should affect the corresponding response parameters exclusively and leave the core parameters of the model untouched. In Experiments 2 and 3, we also introduced a bias manipulation by varying the base rates of old items in the test (30% vs. 70%) to rule out that potential complex interactions of scale use and old/new bias might distort the estimates of the memory parameters. In all experiments, we also assessed the fit and the parameter validity of the SDT model in order to check whether there is a convergence in both model accounts.

The models will be evaluated using both individual and aggregate datasets. It is well known that data aggregation can lead to distortions in the results (e.g., Estes & Maddox, 2005; Rouder & Lu, 2005), a situation that encourages the use of individual datasets. Still, there are circumstances in which data aggregation can be advantageous, such as when a small number of trials is collected per individual (e.g., Chechile, 2009; Cohen, Sanborn, & Shiffrin, 2008). Also, the evaluation of the 2HTM using aggregate data complements previously published validation studies on extensions of the model (e.g., to account for source-memory data; see Bayen et al., 1996), extensions that are traditionally used on aggregate data sets (e.g., Klauer & Kellen, 2010).

EXPERIMENT 1

The aim of the experiment was to demonstrate the validity of the response mapping function which was added to the 2HTM in order to make it appropriate for the analysis of confidence rating data.

Method

Participants. Seventy-two students of the University of Bonn and employed persons (54 female, age 24.44, $SD = 6.9$) volunteered to participate in the study for candy and a certificate of participation (psychology students). Two data sets were destroyed in hard disk crashes, two further participants were excluded because they apparently did not respond seriously (FAR > HR). Hence, data from 68 participants entered the analysis.

Design, materials, and procedure. Two randomised groups were compared who received differently worded response scales. In the *strong wording* condition, the end categories of the scale were denoted as 1 [8] = ''absolutely sure new [old]'', whereas in the *weak wording* condition, the labels were 1 [8] = ''pretty sure new [old]''. Furthermore, the endpoints were visually emphasised by big red exclamation marks in the strong condition. In the learning phase, each participant saw 120 simple line drawings of objects for 1.5 s each, taken from Snodgrass and Vanderwart (1980) and Szekely et al. (2004). Bizarre stimuli (e.g., a genie in a bottle) were excluded as well as pictures from one of the sets that contained the same object as one in the other set. The set of learning pictures was randomly drawn from 240 pictures for each participant. An unrelated filler task (mathematical puzzles) followed for 10 minutes. In the test, all 240 stimuli were presented, half of them new and half of them from the learning phase. In the strong wording condition, the instruction noted that participants should use extreme ratings only in the case of absolute certainty, whereas the participants in the weak condition were encouraged to use the complete scale. Participants received 4 points for each correct old/new decision (correct half of the scale), and 4 points were subtracted for each wrong decision. Every 80 items, they were informed about their account of points won so far. The points could be exchanged for confectionery afterwards.

Results

Overview of analyses. For each experiment, we will first report analyses of the aggregated frequency data summed across all participants. We will report model fit for both a 2df extended 2HTM and an unequal variance SDT with nine parameters (5df). Technical details about the model specification and the restrictions imposed can be found in Appendix B. Second, tests of parameter validity for the 2HTM will be reported, followed by modelling analysis at the individual level. Figure 4 shows the ROCs of all experiments for an illustration.

Manipulation check. First, we checked the efficiency of our scale manipulation at the descriptive level by calculating the proportion of extreme confidence rating responses for each participant. The proportion of extreme ratings was larger in the weak scale condition than in the strong wording condition (.66 vs. .51, respectively), $t(66) = 2.06$, $p = .02$, one-tailed, demonstrating less extreme responding with the latter scale as expected.

Table A2 in Appendix A provides the raw frequencies of ratings given in both conditions of the experiment.

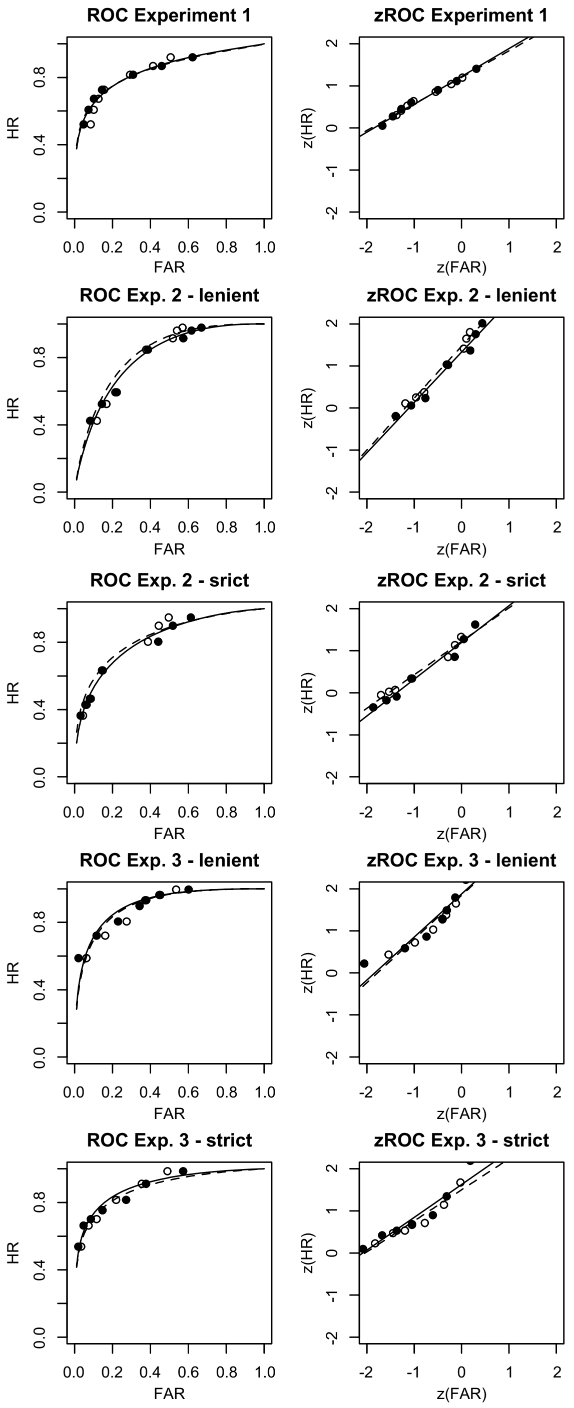

Figure 4. ROCs and zROCs from Experiments 1 to 3. Filled circles are from the strong scale wording conditions, open circles from the weak scale wording conditions. Lines depict best-fitting theoretical SDT curves.

Model fits of aggregated data. Both models SDT and 2HTM were fitted to the data according to the Maximum-Likelihood method, using MPTinR (Singmann & Kellen, in press). The 2HTM $G^2_{(2)}$ values were 1.22 ($p = .54$) and 1.87 ($p = .39$) in the strong and weak scale conditions, respectively. Hence, the 2HTM fit the data very well. According to a compromise power analysis, the critical χ^2-value to detect small model deviations ($w = .10$) with equal α and β probabilities is

22.62. The SDT model, in comparison yielded $G^2_{(5)}$ values of 50.18 and 48.05 (both $ps < .0001$), respectively. AIC values were 25.22 and 25.87 for the 2HTM, whereas they were 68.18 and 66.05 for the SDT model. The corresponding BIC values were 108.95/110.30 for the 2HTM and 130.97/129.38 for SDT.

Parameter values and parameter tests of the 2HTM for aggregated data. Figure 5 shows the parameter estimates of the 2HTM in the top panel with core parameters p_o, p_n, and b in the left and response mapping parameters in the right panel. Dark bars denote the strong wording condition; light bars denote the weak

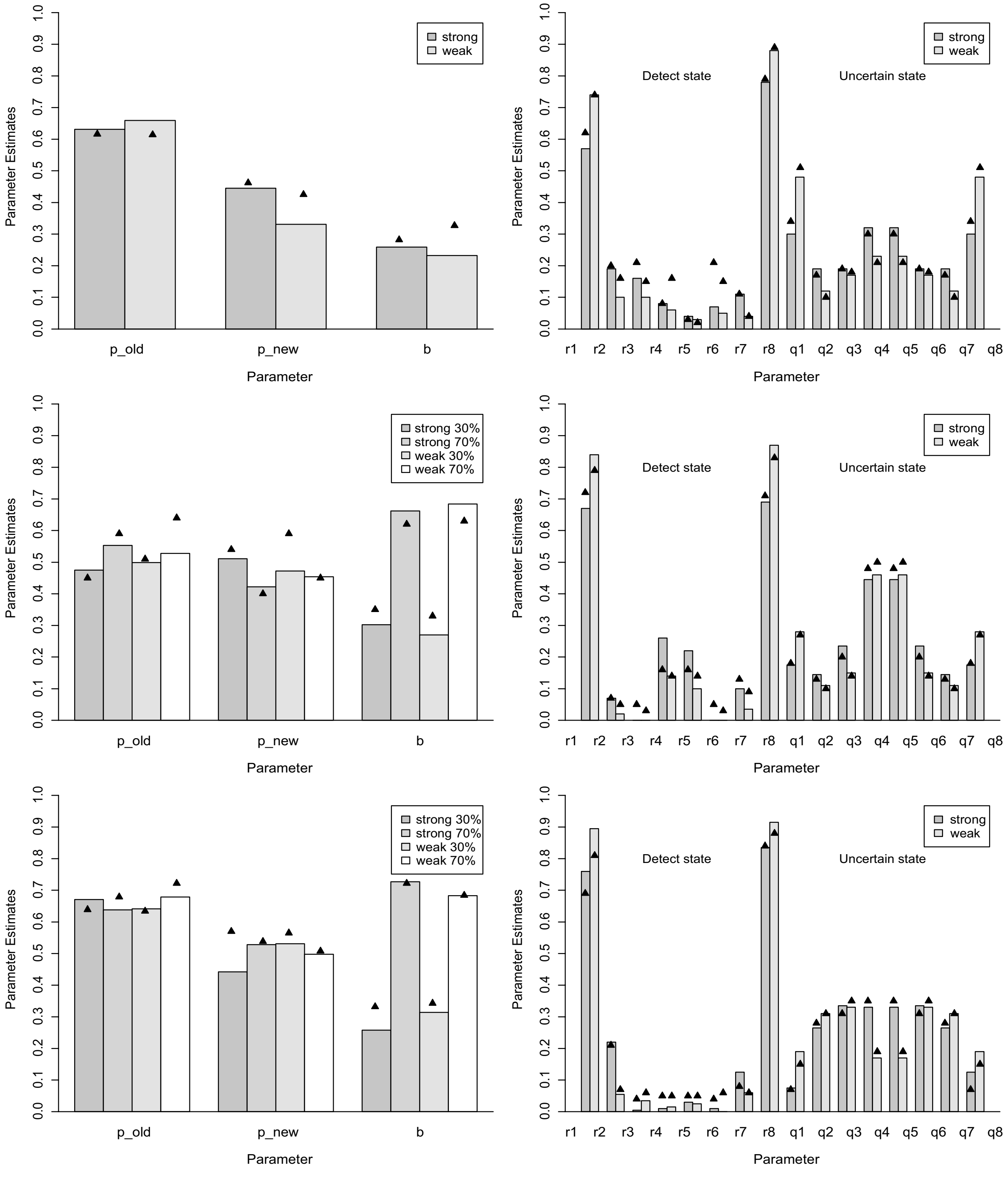

Figure 5. Memory parameter estimates (left column) and response mapping parameter estimates of the extended 2HTM for Experiments 1 to 3. Triangles represent the respective mean estimates from the individual model fits.

wording condition. The core parameter values are similar, and a statistical test confirms that they are not significantly different: Implementing the parameter restrictions $p_{\text{o\_strong}} = p_{\text{o\_weak}}$ and $p_{\text{n\_strong}} = p_{\text{n\_weak}}$, yields a nonsignificant increase in misfit $\Delta G^2_{(2)} = 1.78, p = .41$. The bias parameter b was also unaffected, $\Delta G^2_{(1)} = 0.31$, $p = .58$. Hence, no core memory or strategy parameter was affected by the scale wording manipulation.

The right top panel of Figure 5 shows the response-mapping parameters r_{i} and q_{i} for the detect and the uncertain states, respectively. Descriptively, both distributions are quite different: Whereas the detect state is primarily (but not exclusively) associated with extreme ratings, there are considerably less extreme ratings in the uncertain state. Comparing both scale wording conditions, the probabilities of using extreme ratings are affected in the detect state ($r_{1\_\text{strong}} = r_{1\_\text{weak}}$ and $r_{8\_\text{strong}} = r_{8\_\text{weak}}$), $\Delta G^2_{(2)} = 83.96$, $p < .0001$ as well as in the uncertain state ($q_{1\_\text{strong}} = q_{1\_\text{weak}}$ and $q_{8\_\text{strong}} = q_{8\_\text{weak}}$), $\Delta G^2_{(2)} = 110.97$, $p < .0001$. Finally, the probabilities of extreme responses are much higher in the detect than in the uncertain state, $\Delta G^2_{(2)} = 692.20$, $p < .0001$, corroborating a sensible validity requirement.

To summarise, the parameter estimates of the 2HTM behave largely as expected: Core parameters are unaffected by the rating scale wording, but this manipulation massively affects the response mapping parameters as expected. Also, the response mapping is plausible for the assumed different memory states: The detect state is associated predominantly with extreme ratings, but not exclusively so. In the uncertain state, middle ratings are more numerous.

Individual data analyses: Model fits. The models were also fitted to the data of each individual participant. According to a conventional significance level of .01, both models were rejected for seven (10%) out of 68 participants. The fits of the two models were found to be positively correlated, $r = .30$, $p < .02$. For 20 participants (29%), the 2HTM showed superior AIC values to SDT. BIC values were better for 2HTM than SDT in only two cases (3%). The median AIC and BIC values for the 2HTM are 26.26 and, 68.02, respectively; for SDT the median values are 23.28 and 54.61.

Individual analyses: Parameter values of the 2HTM. The means of the individual parameter estimates are depicted as triangles in Figure 5. As one can see, there is a tendency to overestimate p_{n} and b if the aggregated analysis is viewed as the reference. However, the means of individual estimates resemble the pattern of the aggregated analyses, particularly with respect to the response mapping parameters. The estimates of p_{o}, p_{n}, b, r_1, r_8, q_1 and q_8 were each compared across scale-labelling conditions via independent samples t-tests. Like in the aggregated data analyses, there was no effect on any of the core parameters p_{o}, p_{n}, or b, largest $t(66) = .58$, $p = .57$, whereas the response mapping parameters r_1, r_8, q_1, and q_8 were affected in the expected direction, smallest $t(66) = 2.00$, all ps $< .03$, one-tailed. Paired samples t-test comparing the extreme rating probabilities across memory states (e.g., r_1 vs. q_1 and r_8 vs. q_8) verify the expected patterns, smallest $t(67) = 4.09$, $p < .001$, one-tailed. Hence, the analysis of individual data sets confirms the parameter validity of the 2HTM.

Parameter values of SDT. Table 2 shows the estimated parameter values of the SDT for the aggregated and the individual analyses. Both μ and σ are estimated to be higher than in the aggregated analyses. Also, some individual parameter estimates appear to be somewhat extreme, which led us to use nonparametric tests. The result patterns are similar in both analyses. In particular, the individual estimates of the memory parameters μ and σ are similar across conditions, whereas the spread of criteria is descriptively larger in the strong wording condition which is expected given a supposed

TABLE 2
Parameter estimates for SDT in Experiment 1

	Strong scale labelling		*Weak scale labelling*	
	Aggregated	*Individual (medians)*	*Aggregated*	*Individual (medians)*
Memory parameters				
μ	1.85	2.11	1.91	2.26
σ	1.51	1.65	1.59	1.59
Criteria				
k1	−0.31	−0.44	−0.01	−0.05
k2	0.11	0.15	0.23	0.40
k3	0.49	0.59	0.54	0.60
k4	1.01	1.18	0.96	1.12
k5	1.23	1.52	1.12	1.31
k6	1.46	1.79	1.28	1.64
k7	1.75	1.94	1.41	1.73

reluctance to use the extreme rating bins. For the aggregate data, the restriction of μ and σ across conditions does not lead to an increase in misfit, $\Delta G^2_{(2)} > 1.26$, $p = .53$. The individual estimates of μ and σ were compared across conditions via Wilcoxon tests, which yielded no significant differences, largest $W = 622$, $p = .59$. To assess the effect on the response criteria, the range of the latter (the difference between the two extreme criteria, $k_7 - k_1$) was calculated for each participant. According to a Wilcoxon test, the range of the response criteria was smaller in the weak labelling condition as expected, $W = 747$, $p = .02$, one tailed. Also as expected, the scale wording had no effect on the response criterion determining binary yes/no responses (k_4), as $W = 619$, $p = .61$. Hence, the SDT parameters also reflect the manipulation as expected: Whereas the distribution parameters reflecting sensitivity were unaffected by the labelling of the response bins, the criteria were set narrower for the weakly worded than the strongly worded scale.

Discussion

The results of this experiment show that the response mapping parameters of an extended 2HTM are heavily influenced by a manipulation intending to change response style, while at the same time core parameters are not affected. These results provide some degree of validation to the extended 2HTM as a measurement tool since obviously neither the detection and response-bias parameters are distorted by an irrelevant task feature which may vary across different experiments. Both compared models fared equally well: The scale manipulation did not affect core memory parameters, but the response mapping parameters in the 2HTM as well as the criterion parameters in SDT behaved in a psychologically plausible manner. Altogether, the patterns of aggregated parameter estimates closely followed the mean estimates of individual analyses for both models.

In a second experiment, we both manipulated the old/new bias and the extremity of the scale labels. If complex interactions between scale use and old/new guessing strategies occurred, this would presumably result in a misfit of the model and/or compromised parameter estimates.

EXPERIMENT 2

The aim of this experiment was to replicate the finding that manipulating the scale use should only affect the response strategy parameters. However, it might be possible that scale use interacts in complicated ways with the overall tendency to respond “old” rather than “new”. If this were the case, distortions of core parameter estimates might result. Hence, we used a response bias manipulation orthogonal to the scale manipulation in order to test for such effects.

Method

Participants. Fifty-one participants (four employed, 47 students, 39 female, age 23.55, *SD* = 3.75) volunteered to attend the study which was conducted at the University of Bonn.

Design. Like in Experiment 1, the scale labels and instructions were manipulated between subjects. However, there was an additional base rate manipulation: Participants received two recognition tests with different percentages of old items (30% vs. 70%). The order of the tests was counterbalanced within the scale labelling conditions.

Materials. In order to minimise cross-interference with other memory studies in our lab, a total of 240 black-and-white portrait photos with emotionally neutral expressions were used as stimuli in this experiment (Phillips, Moon, Rizvi, & Rauss, 2000).

Procedure. For each participant, 120 photos were drawn randomly from the pool and presented for 1.5 s each with an ISI of 0.5 s. In the 20 minute retention interval, participants filled in an unrelated personality questionnaire and worked on word puzzles. Subsequently, they received a recognition test consisting of two parts. One part consisted of 30% old items and 70% new items, the other part of 70% and 30%, respectively. The sequence was counterbalanced across participants. Participants were informed before each test part about the percentage of old items in this section. They were motivated by a linear payoff scheme which added/subtracted 1 to 4 points to/from their account for correct/incorrect answers (depending on the extremity of the confidence judgement). The points could be

TABLE 3
Fit values for aggregated data in Experiment 2

	2HTM					*SDT*				
Condition	*G^2 (df = 2)*	*w*	*p*	*AIC*	*BIC*	*G^2 (df = 5)*	*w*	*p*	*AIC*	*BIC*
Strong, 30%	4.31	.04	.12	28.31	100.39	104.81	.19	<.0001	122.81	176.87
Strong, 70%	8.53	.05	.01	32.53	104.60	103.95	.19	<.0001	121.95	176.01
Weak, 30%	5.89	.04	.05	29.89	102.44	60.06	.14	<.0001	78.06	132.47
Weak, 70%	8.42	.05	.01	32.42	104.97	59.60	.14	<.0001	77.60	132.01

G^2 = likelihood ratio statistic, w = effect size (Cohen, 1988), AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

exchanged for confectionery afterwards. After the memory test, another unrelated personality test followed, then participants were thanked and debriefed.

Results

For a manipulation check, we compared the individual proportions of extreme responses across scale wording conditions. The strong wording elicited fewer extreme responses (43%) than the weak wording (56%), $t(49) = -3.12, p = .002$, one-tailed.

Aggregated data: Model fits. The same model as in Experiment 1 was fit to all four data sets separately. As depicted in Table 3, the 2HTM fitted the "strong, 30%" condition quite well, but it reached conventional significance criteria in the other conditions. Note, however, that the power to detect even small model violations (w = .10) is larger than .99 in all cases. The estimated effect sizes are much smaller than Cohen's (1988) convention for "small" effects. In comparison, SDT fits the aggregated data worse with respect to each fit criterion, and the estimated effect sizes are in between "small" and "medium" effects.

Aggregated data: Parameter values and parameter tests of the 2HTM. The left middle panel of Figure 5 shows the core parameter estimates for the 2HTM. Descriptively, neither the scale manipulation nor the bias manipulation seem to have an effect on the memory parameters, whereas b appears to be affected massively by the bias manipulation. The parameter tests reported in Table 4 confirm this impression with the exception of the strong labelling condition, which yielded a conventionally significant effect of the bias manipulation on both p_o and p_n. The old detection parameter p_o is estimated somewhat higher in the 70% condition, whereas the opposite pattern holds for $p_{n.}$ This effect is absent in the weak wording condition. The estimated effect size is only w = 0.05, however. The bias parameter b is not affected at all by the scale manipulation, but it shows the expected massive effect of the bias manipulation.

Furthermore, the estimates of b are close to the actual base rates of the tests in both bias conditions (see Figure 5). Actually, fixing b to .30 and .70 in the 30% and 70% conditions, respectively, has no significant effect on the model fit. This result is suggestive in showing that participants exhibit almost perfect *probability matching* behaviour which has been reported as a standard result in binary choice under uncertainty (Herrnstein, 1961; Koehler & James, 2009). Organisms tend to distribute their responses according to reinforcement proportions of two options rather than to maximise (i.e., always choose the more probable option). Note, however, that this behaviour is

TABLE 4
2HTM core parameter tests on aggregated data in Experiment 2

Parameter restriction	*Condition*	*ΔG^2*	*df*	*p*
$p_{0\_weak} = p_{0\_strong}$	30% old	0.89	2	.64
$p_{n\_weak} = p_{n\_strong}$				
	70% old	1.00	2	.61
$p_{o\_30} = p_{o\_70}$	strong	7.77	2	.02
$p_{n\_30} = p_{n\_70}$				
	weak	0.32	2	.85
$b_{\_weak} = b_{\_strong}$	30% old	1.17	1	.28
	70% old	0.64	1	.42
$b_{\_30} = b_{\_70}$	strong	45.70	1	<.0001
	weak	57.47	1	<.0001
$b_{30} = .30,\ b_{70} = .70$	strong	2.46	2	.29
	weak	**0.95**	2	.62

TABLE 5
Parameter estimates for SDT in Experiment 2

	30% old				*70% old*			
	Strong scale labelling		*Weak scale labelling*		*Strong scale labelling*		*Weak scale labelling*	
	Aggr.	*Indiv.*	*Aggr.*	*Indiv.*	*Aggr.*	*Indiv.*	*Aggr.*	*Indiv.*
μ	1.37	1.34	1.54	1.60	1.11	1.23	1.19	1.33
σ	1.14	1.20	1.26	1.08	0.84	0.97	0.81	0.68
k1	−0.30	−0.32	0.00	0.17	−0.47	−0.57	−0.20	−0.18
k2	−0.04	−0.06	0.14	0.30	−0.31	−0.38	−0.11	−0.09
k3	0.20	0.23	0.33	0.48	−0.10	−0.17	0.01	0.04
k4	1.01	0.98	1.07	1.19	0.27	0.27	0.34	0.41
k5	1.42	1.41	1.41	1.54	0.88	0.95	0.86	0.91
k6	1.57	1.70	1.50	1.65	1.06	1.20	0.98	1.08
k7	1.80	2.10	1.63	1.75	1.28	1.52	1.11	1.24

Aggr = parameter estimates based on aggregate data; Indiv. = median parameter estimates obtained with individual data.

measured by b, conditional on the estimated latent state of uncertainty. The open responses do not exhibit probability matching. Hence, the pattern is consistent with the interpretation that people probability match if they are uncertain, but they rely on their memory if they are in a state of detection.

The right middle panel of Figure 5 shows the response mapping parameters r_{i} and q_{i} for the detect and the uncertain states, respectively. Descriptively, both distributions are quite different: Whereas the detect state is primarily (but not exclusively) associated with extreme ratings, there are considerably fewer extreme ratings in the uncertain state. Comparing both scale wording conditions, the probabilities of using extreme ratings are affected in the detect state ($r_{1\_\mathrm{strong}} = r_{1\_\mathrm{weak}}$ and $r_{8\_\mathrm{strong}} = r_{8\_\mathrm{weak}}$), $\Delta G^2_{(2)} = 82.45$, $p <$ .0001, as well as in the uncertain state ($q_{1\_\mathrm{strong}} = q_{1\_\mathrm{weak}}$ and $q_{8\_\mathrm{strong}} = q_{8\_\mathrm{weak}}$), $\Delta G^2_{(2)} = 38.96$, $p < .0001$. Finally, the probabilities of extreme responses are much higher in the detect than in the uncertain state, $\Delta G^2_{(4)} = 1041.31$ and 1088.85 across scale or bias, respectively, both ps < .0001.

Individual data analyses: Model fits. The models were again fitted to the data of each individual participant separately for both bias conditions. According to a conventional significance level of .01, SDT was rejected for 46 of 102 data sets (45.1%), whereas the 2HTM was only rejected in 35 cases (34.3%). As in the previous experiment, the models' goodness-of-fit results were positively correlated, $r = .74$, $p < .001$. According to AIC, 2HTM was the better model in 58 cases (56.9%). According to BIC, however, the 2HTM outperformed SDT in only 13 cases (12.7%). The median AIC and BIC values for the 2HTM are 29.88 and, 63.33, respectively; for SDT the median values are 31.37 and 56.46.

Parameter values of the 2HTM. The means of individual parameter estimates are depicted as triangles in Figure 5. Again, a tendency to overestimate p_{n} is apparent which is pronounced in the strict bias conditions. Mixed model ANOVAs were computed for the parameters p_o, p_n, b with bias as a within participants factor. Focusing first on the core parameters, the factor scale affected none of them significantly, neither did the interaction with the factor bias, all Fs(1, 49) < 0.7, all ps > .41, all η^2_{G}s < .001. However, the manipulation of bias affected the estimates of both memory parameters p_{o} and p_n, both Fs(1, 49) > 9.36, both ps < .01, both η^2_{G}s > .06. The main effect on the bias parameter b was considerably larger, however, $F(1, 49) = 83.34$, $p < .001$, $\eta^2_{\mathrm{G}} = .47$.

The response-mapping parameters r_1 and r_8 denote the tendency to use the extreme rating bin on either half of the rating scale in the detect states. A mixed model ANOVA with parameter and bias as within subjects factors yielded the expected effect of scale, $F(1, 49) = 4.97$, $p < .05$, $\eta^2_{\mathrm{G}} = .04$. As in the previous experiments, the estimates of r_1 and r_8 were smaller in the strong-scale condition (.66 and .69) than in the weak-scale condition (.77 and .78). There was also a Bias × Parameter interaction, $F(1, 49) = 6.53$, $p = .008$, $\eta^2_{\mathrm{G}} = .03$: A strict bias (30%) reduced the tendency for extreme "new" responses ($r'_1 = .76$ and .66), whereas it increased the tendency for extreme "old" responses ($r'_8 = .70$ and .76). No other effect reached significance.

In a similar analysis for the parameters q_1 and q_8, the scale factor was significant, $F(1, 49) = 4.39$, $p = .04$, $\eta^2_G = .05$: Again, strong-scaling reduced the tendency for extreme "new" responses (q_1 = .22 and .14) as well as extreme "old" responses (q_8 = .33 and .21). No effects of bias as a within-subject factor was found for q_1 and q_8 ($F < .13$). Finally, the r parameters for extreme responding in the detect states are all larger than their counterparts in the uncertain states, all ts(50) > 9.57, all ps <.001.

Hence, in contrast to the aggregated analyses, there were some unexpected effects: Although the effect of bias on the b parameter was the largest as expected, there was also some effect on the memory parameters p_o and p_n. The effects of the scale manipulation on the response-mapping parameters found in the previous experiment were observed here as well.

Parameters of SDT. The validity issues observed for the 2HTM should be compared with the results from the SDT model in order to check whether these issues are exclusive to 2HTM or not. Parameter estimates are provided in Table 5. Here, the mixed model ANOVA also yields a significant effect of bias on the memory parameter μ, $F(1, 49) = 11.26, p < .01$, $\eta^2_G = .06$ (both other Fs <1). This effect of bias is also existent with respect to the estimated σ parameters, $F(1, 49) = 22.77$, $p < .001$, $\eta^2_G = .16$. The estimated σ parameters are also affected by the rating scale wording, $F(1, 49) = 4.95$, $p = .04$, $\eta^2_G = .06$. We also analysed the range and the median of the estimated criterion locations. The range was affected by the bias manipulation, $F(1, 49) = 14.19$, $p < .001$, $\eta^2_G = .03$, as well as by the scale manipulation, $F(1, 49) = 4.89$, $p = .03$, $\eta^2_G = .08$. The spread of criterion locations was larger under lenient (70%) than strict (30%) bias (2.15 vs. 1.82) and larger with strongly worded rather than weakly worded scales (2.28 vs. 1.70). As expected, the location of the middle criterion separating "new" from "old" responses was massively affected by the bias manipulation, $F(1, 49) = 150.68$, $p < .001$, $\eta^2_G = .49$ (other Fs <1).

Discussion

The second experiment was conducted in order to replicate the finding from Experiment 1 with other stimulus materials. Second, we tested whether complex interactions between old/new bias and scale use might distort other parameter estimates, particularly those of b that had indicated conservative responding in Experiment 1. Given the high statistical power, the fit was satisfactory, pointing to at most small model violations. Again, the endpoint labelling of the rating scale affected the response mapping parameters exclusively in the aggregated data analysis. This is desirable for a measurement model which intends to measure latent memory states independently of the peculiarities of the response format used. Second, the response scale manipulation did not interact in strange ways with the tendency to respond "old" under uncertainty, and consequently, estimates of b were presumably not distorted. Parameter b in both bias conditions almost perfectly matched the base rate of old items in the test. Hence, in the aggregate, our participants showed near to perfect probability matching which has been demonstrated in many other domains as well (Koehler & James, 2009; Shanks, Tunney, & McCarthy, 2002). This exact matching was probably fostered by our use of a within-subjects manipulation of bias. We consider it rather unlikely that the "real" guessing rates were different, whereas a potentially distorting effect of scale use on parameter estimation would exactly compensate for over- or undermatching of the participants. Rather, the exact matching of task probabilities corroborates one assumption of the 2HTM that deviates from other models like SDT: In the state of uncertainty, there is no mnemonic information available to the participant, and he/she has to guess strategically. Here, the guesses seem to be influenced only by the task structure. In contrast, the continuous memory dimension of SDT causes *every* memory judgement to be determined by the underlying memory strength variable. It is unclear how a near to perfect probability matching could be achieved in such a model. Note that this probability matching does *not* occur at the observable level (e.g., reflected in false alarm rates); rather, it occurs *conditional* upon one of three latent states, namely uncertainty. Hence, an SDT strategy simply to adjust the criterion to a certain false alarm rate would not be helpful to achieve this result.

The individual parameter estimates of the 2HTM did not behave as expected, as both memory parameters p_o and p_n were affected by the bias manipulation. This is problematic, given the explicit goal of recognition measurement models to disentangle memory and bias. However, the potential competitor SDT provided a similar picture: Both the location parameter μ and the standard deviation σ were affected by the bias manipulation, the latter one also by the scale manipulation. If one assumes that (1) the bias manipulation used here does not affect sensitivity and (2) confidence ratings reflect accuracy, *both* models are obviously misspecified when they are fitted to individual data, and SDT even more so. Parameter estimates were somewhat biased as compared to the aggregated analysis which may be due to sparse data and many zero cells in the individual analyses.

One might argue that face recognition is ‘‘special’’ (see e.g., Bruce, 1991), and our results were limited to this specific material. Hence, we added a replication experiment, using more standard recognition materials, namely word lists.

EXPERIMENT 3

In order to generalise the findings from Experiment 2, the aim of this experiment was a conceptual replication with more standard materials (word lists).

Method

Participants. Forty-eight volunteers attended the study, which was conducted at the University of Bonn. The data sets of two participants were excluded because they misunderstood the instructions, hence $n = 46$ (40 students, five employees, one unemployed, 32 female, mean age = 22.5).

Design. Like in Experiment 2, scale labelling (strong vs. weak) was varied between participants, whereas old/new bias (30% vs. 70% old items in test) was varied within participants with the order of conditions counterbalanced.

Materials and procedure. To create a big pool of at least 320 concrete nouns as learning and testing material we generated a list of concrete nouns by naming the pictures provided by Snodgrass and Vanderwart (1980) and Szekely et al. (2004).

Procedure. One hundred and sixty words were drawn randomly from the pool and presented for 3.5 s each with an ISI of 0.5 s. After a retention interval of 15 minutes during which the participants had to solve several picture puzzles (finding small discrepancies between two otherwise identical pictures), they absolved a recognition test consisting of two parts, one with 30% old items and one with 70% old items. Each part consisted of 160 nouns. Like in Experiment 2, the participants were informed about the percentage of old and new items in each test and the sequence of both parts was counterbalanced. Correct answers were rewarded by adding 4 points to a virtual account; incorrect answers were punished by subtracting 4 points from the virtual account. The accumulated points could be exchanged for confectionery afterwards.

Results

For a manipulation check, we again compared the individual proportions of extreme responses across scale wording conditions. The strong wording elicited fewer extreme responses (51%) than the weak wording (60%), $t(44) = -2.64$, $p < .01$, one-tailed, as expected.

Aggregated data: Model fits. Again, the two models were fit to the four experimental conditions separately. Table 6 shows that the 2HTM was now less successful in fitting the data, especially for the strong scale labelling conditions.

TABLE 6
Fit values for aggregated data in Experiment 3

	2HTM					*SDT*				
Condition	*G^2 (df = 2)*	*w*	*p*	*AIC*	*BIC*	*G^2 (df = 5)*	*w*	*p*	*AIC*	*BIC*
Strong, 30% old	49.87	.11	<.001	73.87	148.91	137.70	.19	<.001	155.70	211.98
Strong, 70% old	5.75	.04	.06	29.75	104.79	96.18	.16	<.001	114.19	170.47
Weak, 30% old	27.73	.09	<.001	51.73	125.73	113.93	.18	<.001	131.93	187.43
Weak, 70% old	2.13	.02	.35	26.13	100.12	135.10	.20	<.001	153.10	208.60

G^2 = likelihood ratio statistic, w = effect size (Cohen, 1988), AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

TABLE 7
2HTM core parameter tests on aggregated data in Experiment 3

Parameter restriction	*condition*	ΔG^2	*df*	*p*
$p_{0\_weak} = p_{0\_strong}$ $p_{n\_weak} = p_{n\_strong}$	30% old	5.64	2	.06
	70% old	0.26	2	.88
$p_{o\_30} = p_{o\_70}$ $p_{n\_30} = p_{n\_70}$	Strong	0.33	2	.85
	Weak	5.79	2	.06
$b_{\_weak} = b_{\_strong}$	30% old	4.34	1	.03
	70% old	0.05	1	.82
$b_{\_30} = b_{\_70}$	Strong	76.95	1	<.001
	Weak	83.93	1	<.001
$b_{30} = .30$, $b_{70} = .70$	Strong	6.34	2	.04
	Weak	7.88	2	.02

Note, however, that the estimated effect sizes only approach Cohen's (1988) conventions for "small" effects (w = .10). Again, SDT fares much worse with the aggregated data sets independent of the fit criterion used.

Aggregated data: Parameter estimates and tests. The bottom panel of Figure 5 depicts the parameter values of the 2HTM. Again, there does not seem to be a effect of bias or scale on the memory parameters, whereas the bias parameter *b* is massively influenced by the bias manipulation. The parameter tests depicted in Table 7 largely confirm this impression. There were no significant effects of either bias or scale on the memory parameters. The effects close to conventional *p* levels in Table 7 can be attributed almost entirely to the unusually low estimate of p_n in the "strong wording, 30%" condition. As already apparent in Figure 5, *b* is impacted on mainly by the bias manipulation, but much less so by the scale manipulation. Although the parameter estimates of *b* are again numerically close to .30 and .70 in the respective bias conditions (.36 and .73 for data aggregated across scale conditions), their fixation to those values leads to a significant misfit due to the high power.

The right lower panel of Figure 5 shows the response mapping parameters r_i and q_i for the detect and the uncertain states, respectively. Both distributions are quite different: Whereas the detect state is primarily (but not exclusively) associated with extreme ratings, there are considerably fewer extreme ratings in the uncertain state. Comparing both scale wording conditions, the probabilities of using extreme ratings are affected in the detect state ($r_{1\_strong} = r_{1\_weak}$ and $r_{8\_strong} = r_{8\_weak}$), $\Delta G^2_{(2)} = 42.49$, $p < .0001$, as well as in the uncertain state ($q_{1\_strong} = q_{1\_weak}$ and $q_{8\_strong} = q_{8\_weak}$), $\Delta G^2_{(2)} = 72.63$, $p < .0001$. Finally, the probabilities of extreme responses are much higher in the detect than in the uncertain state, $\Delta G^2_{(4)} = 2590.92$ and $\Delta G^2_{(4)} = 2607.86$ across scale or bias, respectively, both *p*s <.0001.

Individual data analyses: Model fits. The 2HTM and SDT were fitted to individual data sets. According to a conventional $\alpha = .01$, the 2HTM was rejected in 10 out of 92 cases (10.87%), whereas SDT was rejected 20 times (21.7%). As in the two previous experiments, the models' goodness-of-fit results were positively correlated, $r = .54$, $p < .001$. SDT outperformed the 2HTM slightly with 51 (55.4%) better AIC values, and SDT was markedly better than the 2HTM concerning BIC (88 cases, 75.7%). The median AIC and BIC values for the 2HTM are 27.53 and, 64.43, respectively; for SDT the median values are 27.52 and 55.20.

Individual analyses: Parameter values of the 2HTM. The means of individual parameter estimates are depicted as triangles in the lower panel of Figure 5.

The core parameters p_o, p_n, and *b* were analysed in a mixed model analysis with scale wording as a between subjects factor and bias as a within subjects factor. Parameter p_o was influenced by bias, $F(1, 44) = 4.97$, $p = .03$, $\eta^2_G = .02$, but not by scale or an interaction with scale (*F*s < 1). Parameter p_n was unaffected by either factor, all *F*s(1, 44) <1.14, all *p*s >.29. Finally, there was a considerably larger effect of bias on the parameter *b*, $F(1, 44) = 197.16$, $p < .001$, $\eta^2_G = .77$. Factor scale had no effect, (both *F*s <1). Hence, the disentangling of response bias and sensitivity was not perfect in the individual 2HTM analyses, but certainly, the bias parameter was affected much more than the sensitivity parameter p_o.

A mixed model ANOVA on parameters r_1 and r_8 with parameter and bias as within-subjects factors yielded the expected effect of scale, $F(1, 44) = 5.85$, $p < .05$, $\eta^2_G = .05$. As in the previous experiment, the estimates of r_1 and r_8 were smaller in the strong-scale condition (.68 and .82) than in the weak-scale condition (.80 and .86). No other effect reached significance.

In a similar analysis for the parameters q_1 and q_8 the scale factor was significant, $F(1, 44) = 11.48$, $p < .001$, $\eta^2_G = .12$: Again, strong-scaling reduced the tendency for extreme "new" responses ($q_1 = .12$ and .04) as well as extreme "old" responses ($q_8 = .17$ and .08). An effect of bias as a within-subject factor was also found for q_1 and q_8, $F(1,\ 44) = 18.60$, $p < .001$, $\eta^2_G = .06$: When there is a response tendency to recognise items (70% old condition), parameters q_1 and q_8 assume values (.06 and .10), that are lower than when the response tendency is in the opposite direction (.11 and .16). Finally, the r parameters for extreme responding in the detect states are all larger than their counterparts in the uncertain states, all ts(50) >16, all ps <.001.

Parameter values of SDT. The small but significant effect of the bias manipulation on p_o in the individual analyses challenges the 2HTM's discriminant validity as the bias manipulation is not expected to affect sensitivity measures. However, SDT does not fare better in this respect (see Table 8): The bias manipulation had a significant effect, both on the estimates of μ, $F(1, 44) = 4.57, p = .04$, $\eta^2_G = .02$, and on σ, $F(1,\ 44) = 24.48$, $p < .001$, $\eta^2_G = .20$. Neither μ nor σ were affected by the scale factor (both Fs <.1). The individual range of criterion values was impacted on by both the bias manipulation, $F(1,\ 44) = 8.27$, $p < .01$, $\eta^2_G = .05$, and the scale manipulation, $F(1, 44) = 10.90$, $p <$ .01, $\eta^2_G = .15$. Analogous to the results obtained with q_1 and q_8, strong wording of the scale and a strict bias (30%) both lead to smaller criterion ranges, but the effect was additive (interaction $F = .006$). The location of the middle criterion (median of criteria) was only affected by the bias manipulation, $F(1,\ 44) = 72.77$, $p < .001$, $\eta^2_G = .31$ (all other Fs <1) in the expected direction.

Hence, like the 2HTM, SDT showed a weakness in attributing a bias difference partly to sensitivity differences, but also like in the 2HTM, this effect was very small compared to the corresponding huge effect in the response tendency parameters.

Discussion

The memory parameters of the 2HTM were again unaffected by the bias and scale manipulations in the aggregate analyses. When fitting the models at the individual level, p_o was influenced by the bias. However, the same was true for parameters μ and σ of the SDT model.

Individual measurement and data aggregation. In the data analyses of the three experiments, we showed that the extended 2HTM works quite well with data aggregated across participants. Hence, it can be used to test hypotheses about the impact of (quasi)experimental variables on mnemonic or judgemental processes if groups of participants in different conditions are compared. This is an important result, given that extensions of the 2HTM are frequently used to fit aggregate data in fields where reliable individual datasets are usually not available (e.g., Bayen et al., 1996; Klauer & Kellen, 2010). These models are commonly used to compare population parameters estimated from different groups or conditions, similarly to other methods such as t-tests, ANOVAs, etc.

In terms of parameter validity, both the extended 2HTM and SDT showed the same strengths and weaknesses: In the individual

TABLE 8
Parameter estimates for SDT in Experiment 3

	30% old				*70% old*			
	Strong scale labelling		*Weak scale labelling*		*Strong scale labelling*		*Weak scale labelling*	
	Aggr.	*Indiv.*	*Aggr.*	*Indiv.*	*Aggr.*	*Indiv.*	*Aggr.*	*Indiv.*
μ	2.08	2.23	2.04	2.27	1.83	2.10	1.77	1.92
σ	1.28	1.30	1.36	1.39	0.98	0.95	0.96	0.95
k1	−0.19	−0.20	0.01	0.04	−0.30	−0.25	−0.13	−0.26
k2	0.33	0.40	0.40	0.51	0.15	0.23	0.17	0.15
k3	0.67	0.75	0.84	1.03	0.37	0.52	0.40	0.40
k4	1.07	1.10	1.05	1.18	0.53	0.64	0.49	0.52
k5	1.35	1.35	1.22	1.38	0.91	1.05	0.71	0.79
k6	1.58	1.61	1.40	1.49	1.23	1.32	1.04	1.14
k7	1.99	2.09	1.75	1.91	1.64	1.76	1.37	1.53

Aggr = parameter estimates based on aggregate data; Indiv. = median parameter estimates obtained with individual data.

analyses, there was an effect of the bias manipulation on memory parameters in both models. However, as expected, the bias parameters in both models were affected much more.

Aggregation across participants with different parameter values in modelling may lead to compromised overall parameter estimates or goodness-of-fit statistics (Estes, 1956; Klauer, 2006; Malmberg & Xu, 2006; Morey, Pratte, & Rouder, 2008). According to the model fits reported here, the SDT model seems to suffer more from aggregation across participants with different parameter values. The 2HTM, on the other hand, is less successful with individual data. As a speculation, the latter may reflect the vulnerability of MPT models to sparse data with many zero cells which may lead to parameter estimates close to the parameter space boundaries, thus inflating misfit values as well as compromising the sampling distributions of the goodness-of-fit statistics (see Davis-Stober, 2009). Since no SDT parameter is confined within a bounded space, this problem cannot arise with this model.

GENERAL DISCUSSION

Confidence rating scales have become an increasingly popular method to elicit responses in recognition memory research. In comparison with binary responses, they are believed to provide more information about the underlying processes, and they can be used to construct ROCs from single participants' data. As the participants with more extreme ratings than correct responses show, however, it is naive to view those ratings as valid and reliable reports of internal variables. The ROCs built on ratings were used to argue against threshold models of memory since discrete state models are believed to predict linear ROCs, which is not supported empirically (e.g., Slotnick & Dodson, 2005; Wixted, 2007; Yonelinas & Parks, 2007). However, theoretical arguments as well as the results presented here and elsewhere clearly demonstrate that rating data are not inconsistent with a discrete-state approach (Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Klauer & Kellen, 2010, 2011a; Malmberg, 2002; Province & Rouder, 2012; Schütz & Bröder, 2011). The original 2HTM was never designed to handle confidence rating data, and using those to refute it is therefore unwarranted (see Falmagne, 1985, and Klauer & Kellen, 2010, for a thorough discussion). If the model is extended with plausible mapping functions to map responses from three internal states to $k > 3$ confidence responses, the data can be accommodated including curvilinear ROCs. The results reported here supplement Province and Rouder's (2012) demonstration of conditional independence, which is implied by a discrete-state approach to recognition but incompatible with a continuous approach.

Furthermore, as we have shown here, the mapping parameters of the extended 2HTM behave as predicted: They capture the variation of response styles, leaving core parameter estimates of old/new bias and memory performance untouched. In addition, as one would expect, "detect" states were accompanied by much more confident responding than the uncertain state in all data sets analysed. The extended model provides identical core parameter estimates to comparable experimental conditions with binary responses. Therefore, we conclude that the extended model can provide a viable alternative to SDT-based measurements of memory performance and bias. This result does not reflect peculiarities of specific learning materials since it was observed for pictures of objects, faces, and word lists. We will discuss several aspects and implications of these results in turn.

Do confidence ratings capture the same processes as binary responses? The conceptualisation of SDT simply assumes that confidence judgements are nothing more than finer graded memory judgements. The criteria are placed along the same decision variable (memory strength), and hence, they are not qualitatively different from binary "old/new" responses. This assumption is rarely made explicit, and it almost never discussed as potentially problematic. However, we conjecture that binary responses and confidence judgements might very well tap different processes, namely inferential judgements and metacognitive judgements, respectively. Whereas the former consult some internal memory representation, the latter metacognitive judgement may well consult other information as well, for example hypotheses about one's own memory performance. In contrast to SDT, the extended 2HTM does not make the problematic assumption of identical decision variables underlying the finer graded response distributions. However, the extended 2HTM assumes that the coarse-grained distinction between the "old" half and the "new" half of the response scale reflects inferential processes based on the memory representation. Finer gradations of judgements within each half of the scale are not restricted to reflect the same underlying memory strength variable. Perhaps, several former results demonstrating different parameter estimates for confidence and binary data (Gardner, Macfee, & Krinsky, 1975; Grasha, 1970) within the SDT framework point to boundary conditions under which the strong equivalence assumption of SDT is not valid. Given the modest assumption in the 2HTM and our results, however, there is at the moment no indication that confidence ratings and binary judgements do differ fundamentally.

Validity of the measurement models. The labelling of confidence scales should not change the underlying memory representation, and consequently, memory measurement with a valid model should be unaffected by this nuisance variable. This lack of an effect was demonstrated for both models (except for SDT's σ parameter in Experiment 2) in the aggregated analyses. The labelling had an effect on responding, but this was adequately captured by the response mapping parameters of the 2HTM and the criterion parameters in SDT. Furthermore, the bias manipulation in Experiments 2 and 3 massively affected the guessing parameter *b* in the 2HTM as well as the criteria in SDT. Therefore, important requirements for the validity of the models are met.

There was, however, also an effect of the bias manipulation on the estimates of p_o in the 2HTM and on both μ and σ in SDT, particularly in the analyses based on individual model estimates. For measurement models with the explicit aim of disentangling bias from memory processes, this result is unfortunate. Two interpretations are possible: Either, both models are misspecified and misattribute some portion of a bias difference to differences in memory performance. In this case, it is noteworthy that both the 2HTM and SDT show the same weakness. Another interpretation is that the bias manipulation used here might not be process-pure, but also affected sensitivity. It is conceivable that participants, for example, put more effort in memory probing in the 70% than in the 30% condition. The current data do not allow for deciding between these interpretations, but validations of various bias manipulations would be a worthwhile enterprise. Note, however, that the effect on memory parameters in both models was considerably smaller than the effect on bias parameters which is important from a pragmatic point of view.

Regarding the effect of the bias manipulation on SDT parameters μ and σ, it should be noted that σ was considerably lower in the 70% old condition, and in some cases the median and aggregate estimates were even below 1. These estimates contrast with the common notion that σ is always larger than 1. Although strange, this pattern of results is not unseen; in fact, decreases in σ along with increases in the bias to respond "old" was reported by Van Zandt (2000). The experimental design used by Van Zandt is relatively uncommon in the recognition memory literature, as it consisted of multiple study–test blocks (with fast item presentation) that ran across several experimental sessions, with only 15 individual data sets collected. The present results replicate the findings of Van Zandt across two experiments using a more typical experimental design as well as a larger sample of participants. The results of Van Zandt have motivated a series of extensions of the SDT model (see Kellen, Klauer, & Singmann, 2012), which means that present results are likely to motivate and inform future developments within the SDT framework.

What are the advantages of eliciting confidence responses? The popularity of confidence ratings has increased since they allow for constructing ROCs without much effort. These have been used to dismiss threshold models. However, since these ROCs have been shown to be nondiagnostic to refute threshold models, is there still any use of confidence rating data? We think that there are still some advantages: First, the finer graded response scale produces many more degrees of freedom in the data, leading to two positive consequences at the level of a measurement model. On the one hand, the model becomes testable. Whereas standard computations of d' in SDT and *p* in the 2HTM have to assume that the underlying model is valid, extended models cannot trivially be fitted to the data. Hence, serious model violations can show up in goodness-of-fit tests and caution the researcher to view the respective measures as valid in certain situations. Furthermore, this gain in *df*s allows for a separate estimation of p_o and p_n (or the standard deviation of the old-item familiarity distribution in SDT). Setting those equal in binary response standard applications does rarely follow from theoretical considerations, but it is only a technical necessity for identifying the parameters. In most of our data sets, a detect-old state was reached with higher probability than a detect-new state, thus generating asymmetrical ROCs. This is psychologically plausible if one assumes the existence of a recollection process that can produce the former state, but not the latter. The detect-new state can only be reached by inferential processes based on indirect information as we discussed above. A ‘‘recollection’’ of a nonoccurence is not possible. An additional recollection process for old items therefore raises their probability to be detected as compared to new items that cannot profit from such a process.

Second, the enriched data base may subsequently serve to test process models which explicitly address finer graded response scales, such as the extended diffusion model (Ratcliff & Starns, 2009). Hence, to enable further analysis with a richer data base, scientists might consider confidence ratings for eliciting recognition judgements as long as the response format is expected not to interfere with the processes of interest. Therefore it seems that benefits can be gained by using confidence data without the risk of distorting the processes under scrutiny.[3]

A misconception should be clarified though: Confidence rating responses enable an unconstrained estimation of core parameters (e.g., p_o and p_n are estimated separately), and are more convenient to obtain than binary responses under different response-bias conditions. This state-of-affairs does not imply that one requires one of these two approaches to obtain unconstrained estimates of the models' core parameters. For example, Kellen and Klauer (2011) have shown that the main recognition-memory measurement models proposed in the literature can be fitted to data from a four-alternative forced-choice (4AFC) task in which first and second choices are provided.

Why should we use a discrete measurement model in the first place? The pragmatic use of measurement models strives at detecting the impact of experimental manipulations on memory performance and bias processes, respectively. This can be accomplished by continuous SDT models as well as by discrete models like the 2HTM, which show very high convergent validities in recognition and source memory applications (Bröder & Schütz, 2009; Schütz & Bröder, 2011; Snodgrass & Corwin, 1988). Given the widespread use of SDT, why should we bother using discrete models anyway? We conjecture that there is a good pragmatic reason to use discrete models in many applications: The 2HTM belongs to the class of multinomial processing tree (MPT) models which have been successfully applied to many domains in cognitive psychology (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Riefer & Batchelder, 1988). This model class is conceptually easy and mathematically tractable and comes with a fully developed statistical theory for hypothesis testing (Hu & Batchelder, 1994) as well as easy-to-use analysis software (Hu, 1991, 1999; Moshagen, 2010; Rothkegel, 1999; Singmann & Kellen, in press; Stahl & Klauer, 2007). Furthermore, parameter estimates of validated model parameters are easily interpretable as transition probabilities between latent states, and there is no ambiguity about which measure is the ‘‘best’’ to represent a cognitive process. In contrast, other measurement models like SDT may provide a multitude of potential measures for performance or bias (e.g., criterion location, β, or log(β)), which have different statistical properties and might lead to different conclusions if concurrently analysed. Last, but not least, MPT models can often easily be extended to handle new data situations, for example source memory recognition tests (Batchelder & Riefer, 1990; Bayen et al., 1996). Whereas SDT has also been extended to capture the two judgement dimensions in a standard source monitoring test (DeCarlo, 2003; Hautus, Macmillan, & Rotello, 2008), it is currently unclear how a model for more sources or source dimensions could look like. Batchelder, Riefer, and Hu (1994) discussed some of the difficulties associated to the implementation of a SDT model for three sources.

For example, Meiser and Bröder (2002) successfully extended the source memory MPT model by Bayen et al. (1996) to a situation with two independent source dimensions as well as an additional distinction to the ‘‘remember–know’’

[3] There may be exceptions, however. For example, Van Zandt (2000) used a confidence scale and simultaneously manipulated old/new bias with complex payoff schemes for different rating categories. This lead to a misfit of SDT (see Van Zandt, 2000) as well as a misfit of the 2HTM for the dichotomised data (see Bröder & Schütz, 2009) and the extended 2HTM. Hence, neither measurement model was applicable to those data. In preliminary studies, we observed that participants have difficulties understanding graded payoff instructions even for binary data. Presumably, then, too complex combinations of a graded response scale and other instructions may be a boundary condition for using confidence scales.

procedure, thus making hypotheses about specific representational assumptions testable (Meiser, 2005). The model simply conceptualises the two source judgements as additional two-high-threshold processes conditional on old detection, and it has been construct validated experimentally. More complex extensions of the source memory MPT model have also been developed in social cognition in order to account for memory for multiple source dimensions (e.g., memory for individuals, gender and/or racial groups, etc.; see Klauer & Wegener, 1998) and the associated guessing processes. Although the extension of the source-memory MPT model is relatively straightforward for such cases, it is not clear how a tractable SDT model could be implemented.

Hence, even if discrete models may turn out to be mere approximations to a continuous process, their high convergent validity with SDT as well as their flexibility in terms of adjusting them for new data situations renders them quite useful measurement tools in memory research. It should be noted that outside the memory field, models that assume discrete latent states are commonly used as useful approximations of the cognitive processes driving behavioural responses (Dutilh, Wagenmakers, Visser, & van der Maas, 2011; Ratcliff & McKoon, 2001; Schmittmann, Visser, & Raijmakers, 2006). Whether or not researchers find the extended 2HTM useful for their own analyses, our results are theoretically important by showing that there is neither theoretical nor empirical evidence favouring the continuous model over the discrete one for measurement purposes.

Is cognition discrete or continuous? Juxtaposing "continuous" versus "discrete" models of recognition is popular (see, e.g., Mickes et al., 2009), but in our view, the answer depends on the level of analysis. At the bottom level, action potentials of neurons are certainly discrete. On the other hand, firing rates of neurons or cell assemblies are approximately continuous. We do not doubt that they map onto an internal continuous psychological variable which may represent something like "familiarity" or—put into more neutral terms—a level of evidence for one of the categories "new" or "old". Such a decision variable may even be a basis for confidence judgements, although we contend that determining a mapping function remains elusive, and SDT's equating of both is at least speculative. On the other hand, on a more cognitive level, the information entering consciousness to form the basis of a categorical judgement ("old" vs. "new") may be discrete, even if it is based on a more fundamental continuous decision variable. Earlier, we conceived of this as "equivalence classes" of strength values that are accessible to consciousness. We know such threshold effects for example from the categorical perception of speech sounds (Liberman, Harris, Hoffman, & Griffith, 1957), faces (Beale & Keil, 1995), or even familiar objects (Newell & Bülthoff, 2002) where a continuous stimulus variable (e.g., the voice onset time of a consonant-vowel syllable) is mapped on perceptual categories separated by a threshold. Within these phoneme categories, stimuli cannot be readily discriminated, whereas between categories, discrimination is almost perfect. Another analogy are contrast enhancing mechanisms in perception like the well-known Mach bands which show that cognition strives at sharpening the perceived boundaries between adjacent areas. Yet another example are bistable representations like the Necker cube or Rubin's vase, which both can only be perceptually represented in two mutually exclusive ways. Nobody would deny that, at some level of description, processes leading to these perceptions are probably "continuous". However, at the phenomenological level, no intermediate representations are possible, and this is the level traditional perception research tries to describe and model. Hence, it is well conceivable that continuous and discrete models are appropriate for different levels of theorising, and although the proximal output of the memory system may be continuous, its representation in consciousness may well be "discretised".

Manuscript received 25 June 2012
Manuscript accepted 13 January 2013
First published online 8 February 2013

REFERENCES

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Learning memory and thinking* (Vol. 1, pp. 243–293). San Francisco, CA: Freeman.

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

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review*, *6*, 57–86.

Batchelder, W. H., Riefer, D. M., & Hu, X. (1994). Measuring memory factors in source monitoring: Reply to Kinchla. *Psychological Review*, *101*, 172–176.

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.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*, 217–239.

Birnbaum, M. H. (2011). Testing theories of risky decision making via critical tests. *Frontiers in Psychology*, *2*, 315.

Brandt, M. (2007). Bridging the gap between measurement models and theories of human memory. *Journal of Psychology*, *215*, 72–85.

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.

Bruce, V. (Ed.) (1991). Special Issue on face recognition. *European Journal of Cognitive Psychology*, *3*.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.

Chechile, R. A. (2009). Pooling data versus averaging model fits for some prototypical multinomial processing tree models. *Journal of Mathematical Psychology*, *53*, 562–576.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.

DeCarlo, L. T. (2003). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology*, *47*, 292–303.

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.

Dutilh, G., Wagenmakers, E. J., Visser, I., & van der Maas, H. L. J. (2011). A phase transition model for the speed-accuracy trade-off in response time experiments. *Cognitive Science*, *35*, 211–250.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Journal of Psychology*, *217*, 108–124.

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General*, *127*, 83–97.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

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

Falmagne, J.-C. (1985). *Elements of psychophysical theory*. New York, NY: Oxford University Press.

Flexser, A. J., & Tulving, E. (1978). Retrieval independence in recognition and recall. *Psychological Review*, *85*, 153–171.

Gardner, R. M., Macfee, M., & Krinsky, R. (1975). A comparison of binary rating techniques in the signal detection analysis of recognition memory. *Acta Psychologica*, *39*, 13–19.

Grasha, A. F. (1970). Detection theory and memory processes: Are they compatible? *Perceptual and Motor Skills*, *30*, 123–135.

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

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, *75*, 424–429.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, *69*, 192–203.

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 303–309.

Hautus, M. J., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin and Review*, *15*, 889–905.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*, 267–272.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers*, *16*, 96–101.

Hu, X. (1991). *Statistical inference program for multinomial binary tree models* (Version 1.0) [Computer program]. Irvine, CA: University of California.

Hu, X. (1999). Multinomial processing tree models: An implementation. *Behavior Research Methods, Instruments, and Computers*, *31*, 689–695.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47.

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 and Review*, *18*, 751–757.

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology*, *55*, 251–266.

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in Signal Detection Theory: The case of recognition memory. *Psychological Review*, *119*, 457–479.

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*, 7–31.

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

Klauer, K. C., & Kellen, D. (2011a). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (2010). *Psychological Review*, *118*, 164–173.

Klauer, K. C., & Kellen, D. (2011b). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, *55*, 430–450.

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the ‘‘Who said what?’’ paradigm. *Journal of Personality and Social Psychology*, *75*, 1155–1178.

Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, *113*, 123–127.

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368.

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.

Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General*, *141*, 233–359.

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin and Review*, *13*, 99–105.

Markowitz, J., & Swets, J. A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. *Perception and Psychophysics*, *2*, 91–97.

Meiser, T. (2005). A hierarchy of multinomial models for multidimensional source monitoring. *Methodology*, *1*, 2–17.

Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 116–137.

Mickes, L., Wais, P. E., & Wixted, J. T. (2009). Recollection is a continuous process. Implications for dual-process theories of recognition memory. *Psychological Science*, *11*, 1–7.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin and Review*, *14*, 858–865.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376–388.

Moshagen, M. (2010). MultiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*, 42–54.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*, 839–862.

Murdock, B. B. (2006a). Decision-making models of remember–know judgments: Comment on Rotello, Macmillan, and Reeder (2004). *Psychological Review*, *113*, 648–656.

Murdock, B. B. (2006b). Postscript: Reply to Macmillan and Rotello (2006). *Psychological Review*, *113*, 655–656.

Newell, F. N., & Bülthoff, H. H. (2002). Categorical perception of familiar objects. *Cognition*, *85*, 113–143.

Phillips, P. J., Moon, H., Rizvi, S., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 1090–1104.

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

Raaijmakers, J. G., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207–262). New York, NY: Academic Press.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & McKoon, G. (2001). A multinomial model for short-term priming in word identification. *Psychological Review*, *108*, 835–846.

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

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1446–1465.

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

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments, and Computers*, *31*, 696–700.

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

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

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin and Review*, *17*, 427–435.

Schmittmann, V. D., Visser, I., & Raijmakers, M. E. J. (2006). Multiple learning modes in the development of performance on a rule-based category-learning task. *Neuropsychologia*, *44*, 2079–2091.

Schulze, R. (1909). *Aus der Werkstatt der experimentellen Psychologie und Pädagogik* [*From the workshop of experimental psychology and educational science*]. Leipzig: Voigtländer.

Schütz, J., & Bröder, A. (2011). Signal detection and threshold models of source memory. *Experimental Psychology*, *58*, 293–311.

Schwarz, N., Knauper, B., Hippler, H. J., & Neumann, E. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570–582.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250.

Shiffrin, R. M., & Steyvers, M. (1997). Model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*, 145–166.

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). London: Oxford University Press.

Singmann, H., & Kellen, D. (in press). MPTinR: Analysis of Multinomial Processing Tree models with R. *Behavior Research Methods.*

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory and Cognition*, *33*, 151–170.

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

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 174–215.

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods*, *39*, 267–273.

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

Swets, J. A. (1959). Indices of signal detectability obtained with various psychophysical procedures. *Journal of the Acoustical Society of America*, *31*, 511–513.

Szekely, A., Jacobsen, T., D'Amico, Devescovi, A., Andonova, E., Herron, D., . . . Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*, 247–250.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

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

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

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

APPENDIX A

TABLE A1
Observed frequencies in the rating conditions of Experiments 1 and 2 of Bröder and Schütz (2009)

		Response						
	Item type	*"1"*	*"2"*	*"3"*	*"4"*	*"5"*	*"6"*	Σ
Experiment 1 (n =14)	Old	39	23	22	21	31	280	416
	New	187	77	41	18	31	60	414
Experiment 2 (n =9)	Old	99	58	29	49	64	421	720
	New	384	145	68	20	33	70	720

Response scale labels "1" =sure new to "6" =sure old.

TABLE A2
Observed frequencies of Experiments 1 to 3

		Response								
	Item type	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	Σ
Experiment 1										
Strong	Old	317	207	205	354	211	259	345	2062	3960
	New	1495	646	597	652	167	114	102	187	3960
Weak	Old	486	138	196	281	153	183	159	2604	4200
	New	2072	391	507	579	123	107	66	355	4200
Experiment 2										
30% strong	Old	47	44	86	153	152	32	58	328	900
	New	813	198	162	616	133	60	53	65	2100
30% weak	Old	86	34	65	159	98	17	29	448	936
	New	1101	114	125	530	139	38	40	97	2184
70% strong	Old	46	37	97	141	533	147	209	890	2100
	New	298	47	39	168	148	71	55	74	900
70% weak	Old	78	30	64	156	445	103	121	1187	2184
	New	403	27	21	132	152	44	47	110	936
Experiment 3										
30% strong	Old	15	79	101	65	55	41	132	568	1056
	New	1053	483	258	308	152	95	69	46	2464
30% weak	Old	54	91	129	16	53	24	105	680	1152
	New	1371	365	364	193	85	112	107	91	2688
70% strong	Old	12	78	78	83	228	208	329	1448	2464
	New	421	163	73	37	120	120	101	21	1056
70% weak	Old	36	97	93	40	141	225	259	1797	2688
	New	535	96	92	34	79	130	115	71	1152

APPENDIX B

For convenience in terms of parameter estimation the MPT model was reparameterised into an equivalent binary multinomial tree with the response mapping parameters r_i’ and q_i’. The original parameters for the certain state on the “new” half of the rating scale can be written as functions of these new parameters in the following way:

$$r_1 = r'_1$$
$$r_2 = (1 - r'_1) * r'_2$$
$$r_3 = (1 - r'_1) * (1 - r'_2) * r'_3$$
$$r_4 = (1 - r'_1) * (1 - r'_2) * (1 - r'_3) = 1 - r_1 - r_2 - r_3$$

The indices 1, 2, 3, and 4 correspond to 8, 7, 6, and 5 for the “old” half of the rating scale, respectively, and “q” can be substituted for “r” for the response mapping parameters in the uncertain state. This reparameterisation does not affect the number of parameters in the model, and the original parameter values can be computed from the technically reparameterised version. To render the model identifiable, at least one parameter of each mapping distribution has to be equated across the halves of the rating scale. Since we expected the most extreme response categories to be affected most by the scale manipulation, we chose the restrictions $r'_3 = r'_6$, $q'_3 = q'_6$ and $q'_2 = q'_7$ to fit the extended 2HTM to the data. Hence, there are 12 free parameters in this model (p_o, p_n, b, r'_1, r'_2, $r'_3 = r'_6$, r'_7, r'_8, q'_1, $q'_2 = q'_7$, $q'_3 = q'_6$, and q'_8), therefore 2df.