Psychon Bull Rev (2013) 20:693–719
DOI 10.3758/s13423-013-0407-2

THEORETICAL REVIEW

Recognition memory models and binary-response ROCs: A comparison by minimum description length

David Kellen · Karl Christoph Klauer · Arndt Bröder

Published online: 16 March 2013

Abstract Model comparison in recognition memory has frequently relied on receiver operating characteristics (ROC) data. We present a meta-analysis of binary-response ROC data that builds on previous such meta-analyses and extends them in several ways. Specifically, we include more data and consider a much more comprehensive set of candidate models. Moreover, we bring to bear modern developments in model selection on the current selection problem. The new methods are based on the minimum description length framework, leading to the normalized maximum likelihood (NML) index for assessing model performance, taking into account differences between the models in flexibility due to functional form. Overall, NML results for individual ROC data indicate a preference for a discrete-state model that assumes a mixture of detection and guessing states.

Keywords Model selection · Minimum description length · Normalized maximum likelihood · Recognition memory · Signal detection · Discrete-state models · Hybrid models

The ability to recognize previously encountered information is one of the most popular topics in memory research, with a substantial part of the research efforts devoted to the development of *measurement models*. These measurement models establish connections between certain cognitive processes and the observed responses (Riefer & Batchelder, 1988). The models thereby inform theories that provide more detailed accounts of memory judgments (Malmberg, 2008).

D. Kellen (✉) · K. C. Klauer
Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, 79085 Freiburg, Germany
e-mail: david.kellen@psychologie.uni-freiburg.de

A. Bröder
School of Social Sciences, University of Mannheim, Mannheim, Germany

Several distinct recognition memory measurement models have been proposed: Some assume that memory information is represented as a discrete process (e.g., Batchelder & Riefer, 1990; Klauer & Kellen, 2010), while others postulate a continuous representation of evidence (e.g., Wixted, 2007) or a combination of both (e.g., DeCarlo, 2002; Yonelinas, 1997). Many attempts to distinguish between these models have been based on the predicted receiver operating characteristics (ROC) functions. ROCs describe the recognition of studied and nonstudied items across different levels of response bias (for a review, see Yonelinas & Parks, 2007).

The purpose of this article is to bring to bear recent developments in the assessment of model flexibility on ROC data with binary responses. We present a reanalysis of published ROC data, extending previous such analyses (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube, Starns, Rotello, & Ratcliff, 2012) in several ways: We (1) consider a much wider range of models and (2) base model selection on a principled quantification of the flexibility of the different models by means of the minimum description length (MDL) principle (Grünwald, 2007), taking into account differences in the models' flexibility due to functional form. In addition, (3) we include more data sets than do these previous analyses. Beyond ROC data with binary responses, there are data from other paradigms bearing on the adequacy of the different models, such as, for example, data based on remember/know judgments (e.g., Wixted & Mickes, 2010), data from the process-dissociation procedure (e.g., Yonelinas & Jacoby, 2012), and ROC data based on confidence ratings.

This article is organized as follows. First, the different measurement models and their particular features are characterized. Second, the use of ROC functions to distinguish between the models is discussed. Third, we introduce the MDL principle and the normalized maximum likelihood

(NML) index for model selection derived from it. Finally, published ROC data are reanalyzed under the MDL principle.

The candidate models

Figure 1 presents a visualization of the models discussed here. In an item recognition memory test, previously studied items (old items) and nonstudied items (new items) are presented, and individuals indicate whether they were previously studied (responding *old* or *new*). The probabilities to respond *old* given an old item and a new item are called the *hit* and *false alarm* rates, respectively. Binary-response ROCs plot hit rates against false alarm rates across different levels of response bias. Different levels of response bias can be induced by manipulating the base rate of old items, relative to new items at test, or by payoff schedules (Macmillan & Creelman, 2005; Wickens, 2002).

Recognition memory models make different predictions for the shape of ROCs, and ROCs can therefore be used to discriminate between the different candidate models. For example, discrete-state models predict linear ROCs, continuous models predict curvilinear ROCs, and hybrid models can predict intermediate shapes, as well as more complex ones (Malmberg, 2002). Figure 2 depicts examples of differently shaped ROCs. Note that these predictions hold only if it is assumed that the ability to discriminate between old and new items remains constant across different response bias conditions (for alternative views, see Atkinson, 1963; Balakrishnan, 1999).

The two-high-threshold model (2HTM; Snodgrass & Corwin, 1988) is a discrete-state model that assumes that memory judgments are based on "detect" and "guessing" states.[1] An old item is detected with probability D_o, invariably leading to an *old* response. If the item's old/new status is not detected, with probability $(1-D_o)$, then a guessing-state is entered: The status of the item is then guessed, with the *old* response occurring with probability g and the *new* response with probability $(1-g)$. A new item is detected with probability D_n, invariably leading to a *new* response. When detection fails, with probability $(1-D_n)$, a guessing process is engaged, with the *old* response occurring with probability g and the *new* response with probability $(1-g)$. Regarding the interpretation of parameters, D_o describes correct remembering, D_n characterizes several forms of active distractor rejection processes (e.g., Strack & Bless, 1994), while response bias is captured by parameter g.

[1] The terms *threshold* and *discrete state* are used interchangeably here. In the present context, both terms are used to designate the occurrence of retrieval only, without strong assumptions on the information retrieved (see Batchelder & Riefer, 1999, p. 79; C. M. Parks & Yonelinas, 2007, p. 189).

The 2HTM is a member of the multinomial processing tree (MPT) model class (Riefer & Batchelder, 1988; for reviews, see Batchelder & Riefer, 1999; Erdfelder et al., 2009), and its proponents see it as a simple measurement model that attempts to capture the major cognitive processes involved in old/new judgments (e.g., Bröder & Schütz, 2009). The 2HTM assumes that old/new judgments reflect a mixture of responses made in "memory"/"detection" states and in a "guessing" state in which information on the status of the item is not available. Despite the likely oversimplification and misconception of the underlying cognitive processes (Kinchla, 1994), this particular model and its extensions maintain a mathematical tractability that makes them quite useful in several implementations (see Batchelder, Riefer, & Hu, 1994) and extensions (e.g., Chechile, 2004). Three distinct parameter restrictions are considered here for the 2HTM: $D_o \geq D_n$, $D_o = D_n$, and $D_n = 0$, defining submodels referred to as, in order, $2\text{HTM}_{(D_o \geq D_n)}$, $2\text{HTM}_{(D_o = D_n)}$, and 1HTM. The restriction $D_o \geq D_n$ is included given that it (1) reflects the pattern of results usually found in the literature (e.g., Bröder & Schütz, 2009; Klauer & Kellen, 2010), and (2) reflects the notion that D_n captures distractor rejection processes that are, in part, conditional on memory for studied items as captured by D_o (e.g., Rotello & Heit, 2000; Strack & Bless, 1994). The restriction $D_n = 0$ results in the one-high-threshold model (1HTM; Blackwell, 1963). The restriction $D_o = D_n$ is frequently imposed in using the model as a measurement tool (Snodgrass & Corwin, 1988).

Let $p_{o,i}$ and $p_{n,i}$ be the probabilities of hits and false alarms, respectively, in (base rate or payoff) condition i, $i = 1, \ldots, I$. The 2HTM has parameters $\theta = \{g_1, g_2, \ldots, g_I, D_o, D_n\}$, $0 \leq \theta \leq 1$, and is defined by

$$p_{o,i} = D_o + (1 - D_o)g_i,$$
$$p_{n,i} = (1 - D_n)g_i.$$

The signal detection theory (SDT) model (Banks, 1970; Green & Swets, 1966; Lockhart & Murdock, 1970; Macmillan & Creelman, 2005; T. E. Parks, 1966; Wickens, 2002) is a continuous model. It assumes a continuous memory process, often termed *familiarity*, to describe the individuals' decisions on the basis of memory information. Both old and new items evoke some degree of familiarity, with separate familiarity distributions for old and new items. The ability to discriminate between the two kinds of items is determined by the overlap between the two distributions. According to SDT, an item's familiarity is compared with an established response criterion, denoted by parameter c. If an item's familiarity is larger than the criterion, the *old* response is given; if the familiarity is lower than the criterion, the *new* response is given instead.

The familiarity distributions are assumed to be Gaussian, with parameters $\{\mu_o, \sigma_o\}$ and $\{\mu_n, \sigma_n\}$ for old and new items, respectively, with $\mu_o \geq \mu_n$, $\sigma_o > 0$, and $\sigma_n > 0$. Without loss of

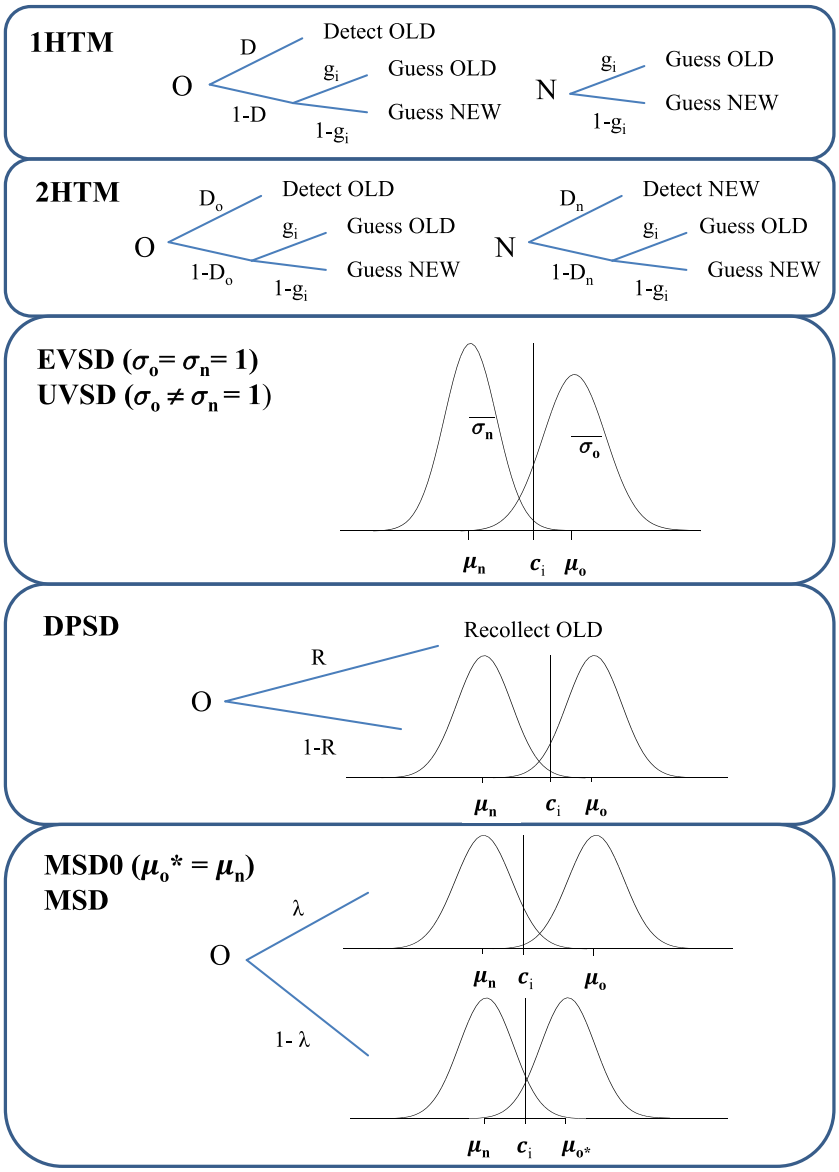

Fig. 1 The recognition memory models. See the text for the definition of parameters

generality, μ_n and σ_n are fixed to 0 and 1, respectively. The unrestricted version is referred to as the unequal-variance signal detection model (UVSD). Two parameter restrictions are considered for this model—namely, $\sigma_o \geq \sigma_n$ and $\sigma_o = \sigma_n$, defining submodels referred to as $\text{UVSD}_{(\sigma_o \geq \sigma_n)}$ and EVSD, respectively. The UVSD model with the restriction $\sigma_o \geq \sigma_n$ is included because it (1) reflects a pattern in parameter estimates that is almost invariably found in the literature (e.g., Ratcliff, McKoon, & Tindall, 1994) and (2) can be given a theoretical justification in terms of encoding variability increasing variability of the familiarity distribution of old items (e.g., DeCarlo, 2010).[2] Although several continuous distributions other than the Gaussian (e.g., Weibull, gamma, log-normal) could be used instead (Rouder, Pratte, & Morey, 2010), we restrict our analysis to the SDT model using Gaussian distributions, due to the fact that it is the most common implementation.

According to SDT, all items have a baseline level of familiarity that is determined by several characteristics (e.g., frequency and/or recency of prior occurrences; see Wixted, 2007). When a set of items (e.g., words) is studied, their average familiarity increases ($\mu_o > \mu_n$). This increase in familiarity is described by the old-item distribution being shifted to the right, relative to the new-item noise distribution (see Fig. 1).

Let F be the cumulative distribution of the standard normal distribution. The UVSD has parameters $\theta = \{c_1, c_2, \ldots, c_I, \mu_o, \sigma_o\}$ and is defined by

$$p_{o,i} = F\left(\frac{\mu_o - c_i}{\sigma_o}\right),$$
$$p_{n,i} = F(-c_i).$$

[2] Despite the fact that this interpretation of σ_o implies the inequality restriction, $\sigma_o \geq \sigma_n$, σ_o is frequently estimated without this restriction (e.g., Dube et al., 2012).

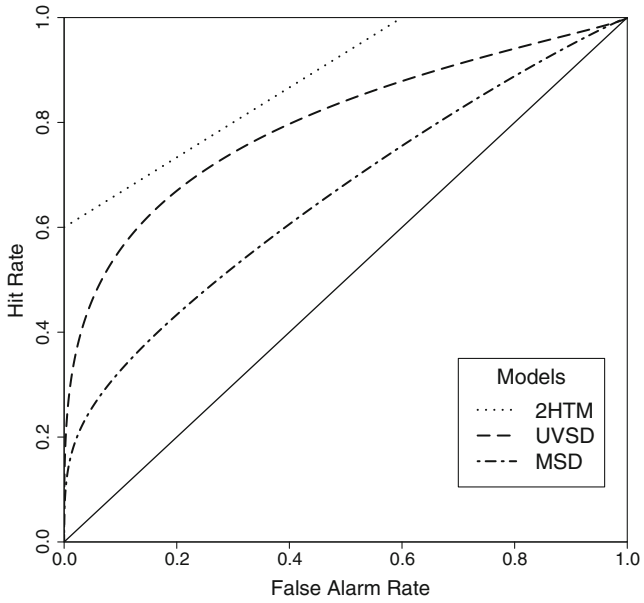

Fig. 2 Examples of receiver operating characteristics (ROC) functions for a selection of models. See the text for a description of the models. Hit and false alarm rates are the probabilities of the *old* response for old and new items, respectively. The solid diagonal represents chance-level performance

The dual-process model (DPSD; Mandler, 1980; Yonelinas, 1997) is a hybrid approach that combines a continuous familiarity process (equivalent to EVSD) and a threshold component, termed *recollection*. When judging an old item, an individual can recollect the item with probability R; when recollection fails (with probability $1-R$), the recognition judgment is based on the item's familiarity, with discriminability determined by μ_o. When judging a new item, recollection cannot occur, which means that the item is evaluated solely in terms of its familiarity. According to Yonelinas (1997), recollection and familiarity represent two independent memory systems that are associated to distinct brain regions (e.g., Yonelinas, Otten, Shaw, & Rugg, 2005).

The "butcher-on-the-bus" anecdote (Mandler, 1980, p. 252) is commonly used to distinguish the separate contributions of recollection and familiarity in the DPSD: Suppose that you are sitting in a bus and encounter a person whose face is very familiar, although you do not remember any particular detail or context about that individual. After engaging in a memory search, if you finally remember the context in which you have met that person before ("That's the butcher from the supermarket!"), then there is recollection. The recollection process thus provides concrete episodic information that is reexperienced by the individual (Yonelinas, 2001), whereas the familiarity process does not rest on the retrieval of episodic detail.

The DPSD has parameters $\theta = \{c_1, c_2, \ldots, c_I, \mu_o, R\}$ and is defined by

$$p_{o,i} = R + (1-R)F(\mu_o - c_i),$$
$$p_{n,i} = F(-c_i).$$

While the probability of the *old* response under the familiarity process depends on the response criterion c_i, recollection invariably leads to an *old* response. Note that the DPSD becomes the EVSD when $R = 0$ and the 1HTM when $\mu_o = 0$.

The mixture signal detection model (MSD; DeCarlo, 2002, 2010) builds on the EVSD and assumes that familiarity for studied items is described by two distributions, one corresponding to items that were attended during study, with mean μ_o, and a second distribution for unattended items with mean μ_o^* and $\mu_o^* \leq \mu_o$. The proportion of attended items among studied items is defined by parameter λ. The restriction $\mu_o^* = 0$ results in the MSD0 model.

The MSD model focuses on differences in encoding—in particular, differences in the level of attention to items during the study phase. It describes these differences by means of a mixture of latent classes, represented by the two distributions for old items, one for attended items, the other one for nonattended items.

The MSD has parameters $\theta = \{c_1, c_2, \ldots, c_I, \mu_o, \mu_o^*, \lambda\}$, with $0 \leq \mu_o^* \leq \mu_o$ and $0 \leq \lambda \leq 1$, and is defined by

$$p_{o,i} = \lambda F(\mu_o - c_i) + (1-\lambda)F(\mu_o^* - c_i),$$
$$p_{n,i} = F(-c_i).$$

Note that the variable-recollection dual-process model proposed by Onyper, Zhang, and Howard (2010) is mathematically equivalent to the MSD model in the present context.[3] Also note that in the present context, the MSD becomes mathematically equivalent to DPSD when $\mu_o \to +\infty$ (see DeCarlo, 2007, 2008). Decisions between mathematically equivalent models cannot be based on model selection criteria. Other criteria of the models' appropriateness can be employed, however—for example, systematic parameter validation studies reflecting the theoretical notions captured in model parameters (e.g., λ and R).

In summary, the models discussed above describe ROC data by a limited number of processes, each model focusing on a different set of processes. According to the 2HTM and restricted cases, recognition memory judgments result from a mixture of discrete memory states (retrieval and rejection) and guessing strategies. For the UVSD and restricted cases, responses are based on the familiarity of items and its

[3] Onyper et al. (2010, Footnote 4) note that the variable-recollection dual-process model could be further extended by including standard deviations of the distributions of attended and unattended old items, σ_{o^*} and σ_o, as parameters to be estimated from the data (see Sherman, Atri, Hasselmo, Stern, & Howard, 2003). In this version, the model would include UVSD as a special case as well. Following Onyper et al., we will not consider this extension.

comparison with established response criteria. For the DPSD, recognition responses can occur via a familiarity process or via episodic recollection. MSD postulates that the familiarity of studied items depends on the attention levels associated to them during study.

Despite several decades of research (for reviews, see DeCarlo, 2010; Malmberg, 2008; Yonelinas & Parks, 2007; Wixted, 2007), the issue of which model best describes ROC data is still under debate (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube, Rotelo, & Heit, 2011; Dube et al., 2012; Klauer & Kellen, 2011a, b). Several factors are likely to contribute to this state of affairs. One issue addressed below concerns the problems associated with analyses of aggregate frequencies and the complementary problems associated with analyzing individual participants' data. Another issue is what we consider to be limitations in the use of confidence-rating ROCs for discriminating between models. A third limitation is that previous analyses usually focused on comparing only two of the above models, such as 2HTM versus UVSD (Dube et al., 2012) or UVSD versus DPSD (e.g., Wixted, 2007). A final issue regards the quantification of model flexibility and how to take this into account in discriminating between the various models.

Previous meta-analyses of ROC data

ROC functions were originally conceived as functions obtained through the use of direct response bias manipulations on binary old/new responses. In most studies, they are, however, obtained by means of confidence-rating scales, which are then compiled in order to emulate changes in response bias, motivated by work in perception suggesting an equivalence of binary-response ROCs with response bias manipulations and confidence-rating-based ROCs with implied response bias variation (Green & Swets, 1966). In comparison to response bias manipulations, the use of confidence ratings is much more efficient and convenient to implement, since only a single test condition is necessary. For example, they sidestep the difficulties of manipulating response bias in an efficient manner (e.g., Cox & Dobbins, 2011). In addition, the use of confidence ratings does not affect any of the predictions made by models such as the UVSD or MSD: According to these models, responses on an $(I + 1)$-point confidence scale are given by establishing a set of ordered parameters $c_1 \leq c_2 \leq \ldots \leq c_I$. Whenever the familiarity of an item falls between c_i and c_{i+1}, for $i = 1, \ldots, I$, the rating response i is given. When the familiarity of an item is lower than c_1, the rating response 1 (corresponding to maximum confidence that the item is new) is given, and when the familiarity of an item is larger than c_I, the rating response $I + 1$ (maximum confidence that the item is old) is given instead.

Although the use of confidence ratings is innocuous for models like the UVSD or MSD, for models like the 2HTM or the DPSD with detect/recollection states, it raises the question of how responses from such states should be mapped onto confidence ratings. One reasonable ancillary assumption is to map them on highest-confidence ratings, leading to the same predicted shapes for the ROCs based on confidence ratings as for ROCs based on binary responses. For example, 2HTM with this ancillary assumption predicts linear ROCs. On the other hand, the existence of individual differences in response styles in the use of extreme response categories and response strategies (e.g., Hamilton, 1968; Tourangeau, Rips, & Rasinski, 2000), intraindividual variations and sequential dependencies in scale usage (e.g., Haubensak, 1992; Malmberg & Annis, 2012) and the possibility of random errors (e.g., Rieskamp, 2008) suggest that a certain proportion of responses generated from detect/recollection states might be mapped on less than highest confidence ratings. But as soon as this possibility is admitted, models with detection/recollection states can predict ROCs shaped like those predicted by UVSD (e.g., Klauer & Kellen, 2010), diminishing the potential of ROC data to discriminate between these models.

Nevertheless, even allowing for nontrivial response mapping, confidence-rating data retain some diagnosticity for discriminating between these models that can be exploited for critical tests (Province & Rouder, 2012), and the added flexibility implied by assuming a nontrivial response mapping can, in principle, be taken into account using the MDL methods detailed below. In the present article, we focus on ROCs based on binary data as a first step, because such data do not require ancillary assumptions regarding the state–response mapping and for the practical reason that we have developed tractable methods for computing the modern MDL-based selection indices only for binary-response ROCs so far.

The fact that binary-response ROCs sidestep the issue of specifying state–response mapping functions makes them particularly attractive for the comparison of models. This advantage was exploited by Bröder and Schütz (2009), who conducted a meta-analysis on ROCs based on binary responses and response bias manipulations. Aggregate data from 59 experiments were fitted with the UVSD and 2HTM and submodels thereof. Goodness-of-fit results, as indexed by the G^2 statistic, favored the UVSD model, especially when focusing on ROCs constructed with a larger number of response bias conditions (a corrected analysis was reported by Bröder & Schütz, 2011). Still, the 2HTM did not reliably produce statistically significant deviations from the data when considering power-adjusted statistics (see Faul, Erdfelder, Lang, & Buchner, 2007). Bröder and Schütz (2009) argued that the diagnosticity of the analyzed ROCs is somewhat questionable, since many were obtained from experiments that included additional manipulations and/or large numbers of study–test blocks across multiple

sessions, which might affect memory discriminability or induce response strategies not accounted for by the models. These possible confounds led Bröder and Schütz (2009) to implement new studies focused on determining the shape of the ROCs. In three experiments, they collected ROC data by varying test item base rates across five levels (10 %, 30 %, 50 %, 70 %, and 90 % of old items). In the three experiments conducted, linear ROCs were obtained, as predicted by the 2HTM, but inconsistent with the curvilinear ROCs based on confidence ratings that have been reported almost ubiquitously in the literature.

Bröder and Schütz's (2009) claims were challenged by Dube and Rotello (2012), who pointed out that most of the ROCs considered by Bröder and Schütz (2009) in their meta-analysis have only two points (i.e., two response bias conditions) and that these cannot be used to assess the shape of ROCs. When focusing on ROCs with at least three points, the goodness-of-fit results indicate a strong preference for the UVSD. Like Bröder and Schütz (2009), Dube and Rotello also reported new experimental data, but with response bias manipulated by (mis)informing participants about the proportions of old and new items. Across the two experiments reported by Dube and Rotello, goodness-of-fit results in general favored the UVSD over the 2HTM, which were again the only models considered. A larger number of test trials was collected in the second experiment (77 old and 77 new items per implied base rate condition), which encouraged Dube and Rotello to compare the two models with both individual and aggregated data. In both cases, goodness-of-fit results supported the UVSD. Similar results were reported by Dube et al. (2012), who used two types of studied items (weak and strong) and a base rate manipulation.

One limitation of the analyses by Bröder and Schütz (2009, 2011), Dube and Rotello (2012), and Dube et al. (2012) that we intend to overcome is that they focus on comparing only 2HTM and UVSD. Moreover, the previous meta-analyses are based on goodness-of-fit results, ignoring possible differences in the models' flexibility, that is in their ability to fit data in general (Pitt & Myung, 2002). Kellen and Klauer (2011) and Klauer and Kellen (2011b) have shown that these models differ pronouncedly in terms of their flexibility, differences which need to be taken into account when comparing the goodness of fit of the different models (Roberts & Pashler, 2000).

Model selection and the minimum description length principle

Traditional model selection approaches

The question of how to compare and choose from competing models is a core aspect of cognitive modeling endeavors (see Myung, Forster, & Brown, 2000; Wagenmakers & Waldorf, 2006). One of the main concerns in model selection is the weighting of goodness of fit and model flexibility (or complexity; both terms will be used interchangeably). The problem is that models can differ in their ability to produce good fits to data in general. An overly flexible model will fit many data sets well, including some that can be seen as inconsistent with its core assumptions (e.g., Roberts & Pashler, 2000). This bears on the support that a model can gather from observed data: As was thoroughly discussed by Roberts and Pashler, strong support for a model requires that it provides a good fit of the data, but also that the a priori probability that the model will provide a good fit is low. In particular, if the model fits well almost any data that could possibly be observed, the support coming from good fits is rather low. Model selection should take this into account by weighing goodness of fit against how flexible each model is in terms of fitting data in general. As is elaborated on below, this notion is straightforwardly implemented in the NML index that flows from the MDL principle.

Another way to characterize this notion of model flexibility is to say that flexible models tend to capitalize on random error in the data to a greater extent than do simpler models, compromising their ability to make accurate predictions for new data. For this reason, more flexible models tend to produce less stable parameter estimates and a greater number of prediction errors when attempting to generalize to new observations. This problem of *overfitting* and *generalization error* encourages the search for a model that strikes the best trade-off between model fit and model parsimony (Hastie, Tibshirani, & Friedman, 2008). Several methods based on different philosophies have been proposed for this purpose (e.g., Myung, Navarro, & Pitt, 2006; Myung & Pitt, 2004).

The most frequently used method is model selection by the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). These indices basically quantify model flexibility in terms of the number of parameters:

$$\mathrm{AIC} = -2\log\, f\left(\mathrm{x}\middle|\widehat{\theta}(\mathrm{x})\right) + 2p,$$
$$\mathrm{BIC} = -2\log\, f\left(\mathrm{x}\middle|\widehat{\theta}(\mathrm{x})\right) + \log(N)p,$$

with p denoting the number of parameters and N the sample size (total number of trials). For both equations, the first term corresponds to (minus) two times the maximum log-likelihood of observed data vector x, where the maximum-likelihood estimates of the p model parameters are denoted by $\widehat{\theta}(x), \theta = \left(\theta_1, \ldots, \theta_p\right)$. Note that the first term in the AIC and BIC formulae thereby quantifies the model's goodness of fit, whereas the second term quantifies the penalty factor quantifying model complexity.

Despite their similarities and joint use in the literature, AIC and BIC are based on rather distinct principles: AIC is based on the expected bias in the Kullback–Leibler divergence of a model (see Burnham & Anderson, 2002), and BIC is an approximation of the *Bayes factor* (Jeffreys, 1961; Kass & Rafferty, 1995). In the derivation of AIC, no assumption is made regarding whether or not the true data-generating process is included in the set of models being evaluated, nor does the sample size play any role. The latter feature is visible in the formulation of AIC, since the penalty factor is not dependent on the sample size N. This characteristic causes AIC to be biased toward more flexible models as sample size increases, in contrast with BIC, which selects the true model (if it is included in the set of candidate models) with probability approaching 1 as sample size increases.

In the present case, both AIC and BIC are of limited value, because many of the models considered here (e.g., 2HTM, UVSD, DPSD, MSD0) use the same number of parameters, so that they are treated as equally flexible by AIC and BIC. But it is easy to see that models can differ considerably in flexibility despite having the same numbers of parameters. For example, AIC and BIC are unable to take into account inequality restrictions that might be imposed on the models. Take the case of UVSD and $\mathrm{UVSD}_{(\sigma_o \geq \sigma_n)}$: Although it is obvious that the latter model is less flexible than the first, due to the inequality restriction imposed, according to AIC and BIC they are equally flexible. More generally, AIC and BIC cannot account for differences in flexibility due to functional form (e.g., Myung, 2000). For example, an additional parameter can have anything between a negligible and a dramatic impact on a model's ability to account for data in general, depending on how it is entered into the model equations, and AIC and BIC gloss over such differences in flexibility. We elaborate on this issue below.

Another approach to the problem of model flexibility is based on simulations (Bröder & Schütz, 2009; Cohen, Sanborn, & Shiffrin, 2008; Dube et al., 2012; Jang, Wixted, & Huber, 2011; Wixted, 2007), the most prominent simulation method being the *parametric bootstrap cross-fitting method* (PBCM) introduced by Wagenmakers, Ratcliff, Gomez, and Iverson (2004). PBCM is also of limited value for the current selection problem, since it assesses *model mimicry* (see Jang et al., 2011) for a given pair of models conditional on a particular data set. Model mimicry is concerned with the distinguishability of a specific pair of models. In contrast, model flexibility is concerned with the ability of a model to fit data in general and can be calculated in the absence of other candidate models.

Importantly, the assessment of model mimicry with the PBCM is inherently limited to pairwise comparisons of models, and it is not clear whether and how it can be extended to selecting from larger sets of models as considered here (Wagenmakers et al., 2004, p. 47). In addition, by comparing two models in terms of their ability to fit data generated from either of the two models on the basis of parameters estimated from the observed data, its relevance for the case that neither model generated the observed data is unclear. For example, for many of the data sets analyzed in previous meta-analyses, neither UVSD nor 2HTM provided a good fit of the data, so that the simulated data sets in a PBCM comparison of these two models necessarily bear little resemblance to the actual data. Given the differences between model mimicry and model flexibility, Wagenmakers et al. advised against the use of the PBCM for the purpose of assessing model flexibility, explaining that the method is unsuited for that purpose and, in addition, is biased in favor of the more flexible model (pp. 40–42) when used to quantify flexibility. To assess model flexibility, Wagenmakers et al. proposed the *data-uninformed* PBCM, which is also known as a *landscaping analysis* (Navarro, Pitt, & Myung 2004) and yields results closely comparable to those obtained with the MDL approach (Cohen et al., 2008; Wagenmakers et al., 2004) considered next.

The minimum description length approach

An approach that overcomes several limitations of traditional model selection methods is the MDL principle (Rissanen, 1984), a framework stemming from information theory (Cover & Thomas, 1991) that is widely used in statistics and machine learning (for a comprehensive introduction, see Grünwald, 2007). According to MDL, data can be seen as a code whose length can be compressed by a model (itself a code with a particular length) according to the regularities present in the data. The more regularities are present in the data, the more the data can be compressed into a smaller description and the more is learned from it, since these described regularities can be used to predict future observations (Grünwald, 2007). MDL has been successfully applied in diverse areas of psychological research, ranging from the class of MPT models (Wu, Myung, & Batchelder, 2010a, b), to human categorization learning (Myung, Pitt, & Navarro, 2007), strategy identification (Davis-Stober & Brown, 2011), clustering (Lee & Navarro, 2005), working memory (Rouder et al., 2008), hypothesis testing (Lee & Pope, 2006), and structural equation (Preacher, 2006) modeling.

Model selection according to the MDL principle can be stated in very general terms: Let $M_1, \ldots, M_J$ be a set of models, and let $L(\cdot)$ be a function that indicates code length. The best model to describe data D is the model that minimizes the sum $L(D|M) + L(M)$, where $L(D|M)$ is the length of the description of the data provided by the model and $L(M)$ is the length of the description of the model itself. The

first term of this sum corresponds to the goodness of fit of a model, and the second term to a quantification of the model's complexity serving as a penalty factor. MDL thereby provides a general theory of inductive inference that inherently implements a form of Occam's razor, taking into account not only the ability of a model (hypothesis) to describe data, but also the flexibility of the model itself.

MDL does not presuppose the existence of an "underlying truth" and, instead, focuses on finding the model that most concisely describes the regularities present in the data (Grünwald, 2007). Furthermore, simpler models tend to be preferred by MDL not because it assumes that "simple models are more likely to be true" but because it is frequently the case that the data available are not sufficient to identify a complex model and all its predicted regularities with any reliability. Note that we think that this "agnosticism" regarding an underlying truth in the MDL framework is appropriate given that none of the simple measurement models considered here is likely to provide more than a first approximation of the underlying processes generating the data. Nevertheless, this agnosticism does not compromise the consistency of the indices that emerge from it; as in the case of one the models being true, they will select it as sample size increases (see Grünwald, 2007, Chap. 7).

The Fisher information approximation and normalized maximum likelihood

Two indices derived from the MDL principle are the Fisher information approximation (FIA; Rissanen, 1996) and NML (Myung, Navarro, & Pitt, 2006), with FIA being an asymptotic approximation of NML. It is instructive to consider FIA first:

$$\mathrm{FIA} = -\log f\left(\mathrm{x}\middle|\widehat{\theta}(\mathrm{x})\right) + \frac{p}{2}\log\frac{N}{2\pi} + \mathrm{FIA}_f,$$

where

$$\mathrm{FIA}_f = \log\int \sqrt{\det \mathrm{I}(\theta)}\; d\theta.$$

The first term of FIA corresponds to (minus) the maximum log-likelihood of observed data x in a particular experiment, quantifying model fit, and the second and third terms correspond to the model penalties.[4] The second term takes the number of parameters p and sample size N into account, similarly to BIC. The third term, FIA_f, accounts for the flexibility of the model due to its functional form by integrating over the determinant of the Fisher information matrix ($I(\theta)$) of the model for a sample of size 1 (Schervish, 1995).

Importantly, FIA behaves as one would intuitively expect of an index that takes functional form into account, as many examples attest (see, e.g., Kellen & Klauer, 2011; Su, Myung, & Pitt, 2005; Wu et al., 2010a, b). For example, if inequality restrictions are imposed on a model's parameter, as in the case of $\mathrm{UVSD}_{(\sigma_o \geq \sigma_n)}$, the restricted model still has the same number of parameters as the original model, UVSD, but it is obviously less flexible than the original model. This is reflected in FIA via FIA_f. The penalty FIA_f is decreased for $\mathrm{UVSD}_{(\sigma_o \geq \sigma_n)}$ relative to UVSD, because the integral of the determinant of the Fisher information matrix is now computed over only a subset of the parameter space—namely, over those parameters that satisfy the inequality restrictions, implying a smaller value for the integral. Neither AIC nor BIC would correct for such a change in model flexibility, because both models employ the same number of parameters. In simulation studies, use of the model selection index based on FIA led to more valid results than the use of only goodness-of-fit values, AIC (Klauer & Kellen, 2011b; Myung et al., 2007), or BIC (Klauer & Kellen, 2011b; Su et al., 2005), as is further illustrated below. Still, the penalty factor due to functional form included in FIA is asymptotic, which can sometimes lead to illogical results when sample size is small (see Navarro, 2004).

FIA is linked to the MDL principle because it provides an asymptotic approximation to the NML index, which can be derived as the "optimal" implementation of the MDL principle (Myung et al., 2006; Rissanen, 2001). The (logarithm of the) NML index is given by

$$\mathrm{NML} = -\log f\left(\mathrm{x}\middle|\widehat{\theta}(\mathrm{x})\right) + \log\sum_{y} f\left(\mathrm{y}\middle|\widehat{\theta}(\mathrm{y})\right).$$

The NML formula has two terms. The first term is identical to the first term in FIA. The second term is a penalty factor that is the sum of the maximum log-likelihoods of all possible data patterns y that could in principle be observed in such experiment. In other words, the first term quantifies a model's goodness of fit (in terms of the maximum likelihood), and the second term penalizes the model for its ability to account for any data that might be observed (again in terms of maximum likelihoods). The flexibility measure derived from the MDL principle thus penalizes a given model to the extent that it provides good fits in general, capturing Roberts and Pashler's (2000) notion of model support. The NML penalty term is defined in terms of the model's fit to all possible data sets (i.e., sets of response frequencies) that can, in principle, arise in a specific experimental design. By considering all possible data patterns that

[4] Note that in FIA (and other MDL indices), goodness of fit is represented by $-\log f\left(x\middle|\widehat{\theta}(x)\right)$, while in AIC and BIC, it is represented by $-2\log f\left(x\middle|\widehat{\theta}(x)\right)$. This means that in order to compare MDL indices with AIC and BIC, one needs to adjust them. This adjustment can be done by either multiplying the MDL indices by 2 or dividing AIC and BIC by 2. We will use the latter option in our comparisons.

could be observed in one experiment, NML takes into account the experiment's design, which means that the NML penalties depend not only on the total number of items, but also on the numbers of items of each type (e.g., old and new items) in every condition of the experiment. NML penalties are thus tailored to each experimental design, in contrast to asymptotic indices such as AIC, BIC, and FIA, whose penalties constitute large-sample approximations.

An illustration

To further illustrate the relationship between model flexibility and NML, and the advantages of the latter in comparison with AIC and BIC, let us consider the present recognition memory models for the case of the 4AFC-2R task (Kellen & Klauer, 2011; C. M. Parks & Yonelinas, 2009). In this task, individuals are presented with a set of four items, one previously studied and the other three not studied. The individuals' task is to choose the item of which they most strongly believe that it was previously studied and then choose a second item among the remaining alternatives as the second most likely item to be the old item. Let π_1, π_2, and π_3 denote the (unconditional) probabilities that the studied item is chosen as the first choice, is chosen as the second choice, or is not chosen, respectively.

Figure 3 depicts the model spaces—that is, the sets of joint values of π_1 and π_2 that each model can describe perfectly, for some of the recognition memory models considered here (π_3 is redundant, because the three predicted probabilities have to sum to one). The range of predictions of both EVSD and 1HTM is limited to a single curve each, while the more flexible models are able to account for defined regions (see Kellen & Klauer, 2011; Theorems 1–4). Still, the size of these regions varies greatly between models. For example, MSD0, MSD, and DPSD make the same range of predictions that is enclosed by the curves for 1HTM and EVSD. On the other hand, 2HTM and UVSD are able to account for larger regions, with UVSD being by far the most flexible model. According to model selection indices like AIC and BIC, the models 2HTM, UVSD, DPSD, and MSD0 are equally flexible, despite the fact that their ability to make predictions varies considerably. In contrast, by fitting the models to each possible data pattern, NML visits all possible $\{\pi_1, \pi_2\}$ values in determining the flexibility penalty and, thereby, captures the differences in model flexibility that are visible in Fig. 3 as demonstrated by Kellen and Klauer (2011).

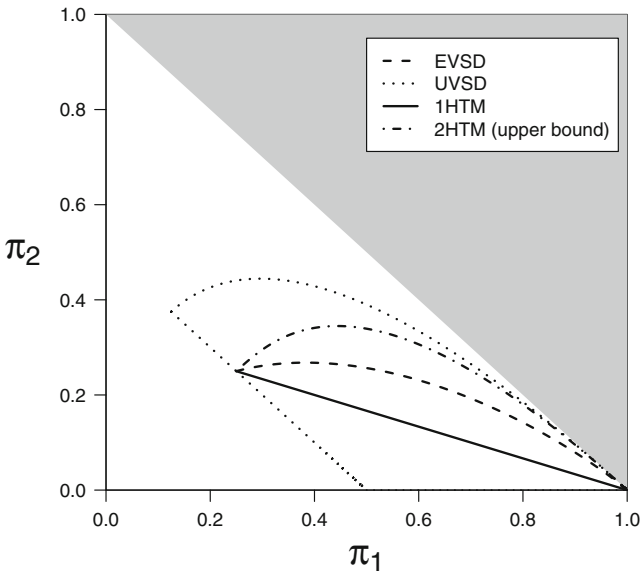

Fig. 3 Range of $\{\pi_1, \pi_2\}$ predictions for a selection of models in the 4AFC-2R task (for details, see Kellen & Klauer, 2011). For EVSD and 1HTM, the predictions are limited to their respective curves (no areas are defined). The range of predictions of both DPSD and MSD correspond to the area enclosed by the curves for EVSD and 1HTM. The predictions of 2HTM correspond to the area enclosed by the curves labeled 2HTM (upper bound) and 1HTM. The predictions of the UVSD model correspond to the area enclosed by the curve labeled UVSD

Computing and interpreting NML

The requirement in computing NML to fit all possible data sets that could, in principle, arise in a given experiment quickly becomes unfeasible as the number of trials increases (e.g., Kellen & Klauer, 2011), a difficulty that contributed to a more common use of FIA (e.g., Pitt, Myung, & Zhang, 2002). As was shown by Klauer and Kellen (2011b), NML can, however, be computed even for relatively large sample sizes via Monte Carlo integration (Robert & Casella, 2004), sidestepping the need to fit all possible data sets that could arise in principle. The estimates of FIA and NML converge as the sample size increases, given that FIA is an asymptotic approximation to NML. More specifically, $\mathrm{NML}_f = \log \sum_y f\left(y \middle| \widehat{\theta}(y)\right) - \frac{p}{2} \log \frac{N}{2\pi}$ converges to FIA_f as $N \to \infty$. Note that both FIA_f and NML_f quantify the flexibility of models due to functional form. Due to its asymptotic nature, FIA can, however, sometimes lead to erroneous and even illogical results when sample sizes are small (e.g., Navarro, 2004), making NML a more suitable measure despite the computational difficulties associated with it.

NML can also be seen as a complementary (and sometimes more convenient) method for Bayesian model selection: As was shown by Balasubramanian (1997), model selection by Bayes factors (Jeffreys, 1961; Kass & Rafferty, 1995) is equivalent to NML under certain conditions (see also Grünwald & Navarro, 2009; Karabatsos & Walker, 2006; Lee & Pope, 2006). The connection between NML and the Bayes factor can be used to obtain rough guidelines on how to interpret the size of NML differences between models (Lee, 2004). The Bayes factor is defined as

$\frac{P(D|M_i)}{P(D|M_j)}$, with $i \neq j$, and quantifies the evidence obtained from the data (D) for one model (M_i) relative to another model (M_j). Following Jeffreys (1961), the value of the Bayes factor on a natural-logarithmic scale (corresponding to the NML scale here used) can be roughly interpreted as follows: Values between 0 and 1.1 provide "anecdotal" evidence for M_i, values between 1.1 and 2.3 represent substantial relative evidence for M_i, values between 2.3 and 3.4 represent strong relative evidence for M_i, and values larger than 3.4 represent very strong relative evidence for M_i. Summing NML values across independent data sets is also legitimate. The summed NML value is simply the NML index of the model fitted to the joint data, with different parameter values permitted for each independent data set.

The flexibility of the models due to functional form

Despite their advantages and successful track record in psychological research, the use of MDL measures such as FIA and NML has been hindered by the difficulty of their computation. Klauer and Kellen (2011b) recently developed tractable methods for computing FIA and NML for the present models for the case of binary-response ROCs and found considerable differences in flexibility. Figure 4 shows model flexibility values that arise from functional form (i.e., FIA_f and NML_f) across different sample sizes for 5-point ROCs for each of the recognition memory models. Note that values can be compared directly for models with the same number of parameters. These are linked by lines in the figure. First, note that the NML_f values converge to FIA_f, as sample size increases, as expected. Of special interest in Fig. 4 is the considerable flexibility of the UVSD, which is the highest among the candidate models, a result that is likely to lead to significant changes in the conclusions of the previous meta-analyses that did not take flexibility due to functional form into account. These FIA_f and NML_f results per se already run counter to the notion that the UVSD and DPSD are approximately equal in flexibility and corroborate previous analyses that indicated a greater propensity for the UVSD model to overfit data (Bröder & Schütz, 2009; Jang et al., 2011), although it should be emphasized that differences in flexibility do not necessarily generalize across tasks or variations of a particular task (see Kellen & Klauer, 2011; Klauer & Kellen, 2011b). For example, the 2HTM is more flexible than the DPSD and MSD0 in the context of the 4AFC-2R task, but not for the case of binary-response ROCs.[5] Model complexity or flexibility is a property of the model *and* of the experimental design.

[5] There are tasks in which the UVSD is less flexible than many of the candidate models: In a two-alternative forced choice task, the UVSD reduces to the EVSD (see Wickens, 2002, Chap. 6), which is a special case of DPSD, MSD0, and MSD.

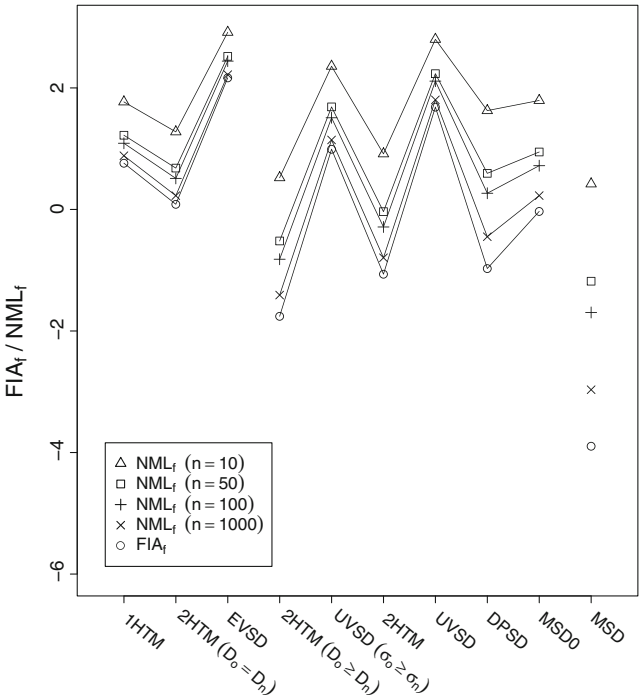

Fig. 4 FIA_f and NML_f for five-point ROCs, and n = 10, 50, 100, and 1,000 trials per hit and false alarm rates. Note that FIA_f and NML_f values can be directly compared only across models with the same number of parameters (models connected by lines)

Additionally, the differences in NML_f presented in Fig. 4 show that the number of parameters is not a good proxy for model flexibility—that is, for the model's ability to fit data in general—given that the models connected by a line in the figure have the same number of parameters and still show considerable differences in the ability to fit data in general (i.e., in NML). More strongly, the UVSD and the MSD provide an example in which a model with more parameters (MSD) is even less flexible than a model with fewer parameters (UVSD), at least for data sets of realistic sizes: For 5-point ROCs with 10, 50, 100, and 1000 old and new items per response bias condition, the differences in NML penalties between MSD and UVSD are, in order, −1.00, −1.23, −1.27, and −1.08. These penalty differences indicate that UVSD has a greater flexibility despite the fact that MSD has one more parameter than UVSD and would accordingly be punished as more flexible by AIC and BIC. For example, for BIC, the penalty differences in quantifying goodness of fit would be 2.30, 3.11, 3.45, and 4.61, respectively, favoring UVSD (see Footnote 4). Given the common use of AIC and BIC in the literature, it becomes clear that models such as MSD have been unfairly penalized so far. Differences in flexibility can also be found among the models with fewer numbers of parameters, like $2HTM_{(D_o = D_n)}$ and EVSD: According to AIC and BIC, these models are equally flexible. Also, these models are highly constrained in terms of the ROCs that they can account

for: $2HTM_{(D_o = D_n)}$ can only predict linear ROCs with slope 1, and EVSD is restricted to symmetrical curvilinear ROCs. Still, NML results indicate that EVSD is able to fit data in general much better than $2HTM_{(D_o = D_n)}$: For 5-point ROCs with 10, 50, 100, and 1,000 old and new items per response bias condition, the differences in NML penalties between $2HTM_{(D_o = D_n)}$ and EVSD are, in order, −1.64, −1.84, −1.93, and −1.99.

Model recovery by AIC, BIC, and NML

We want to apply NML to selecting from the set of models shown in Fig. 4. As a preparatory step, we conducted a model recovery study known as landscaping analysis (Myung et al., 2007; Navarro et al., 2004) or *data-uninformed* PBCM (Wagenmakers et al., 2004) across the ten candidate models considered here (as listed in Fig. 4), complementing similar simulations previously reported by Klauer and Kellen (2011b) for selecting from subsets of these models. We compared model recovery performance for AIC, BIC, and NML. Ten thousand 5-point binary-response ROC data sets were generated from each model for each of three sample sizes: n = 50, n = 100, and n = 1,000 per item type and response bias condition. Sample sizes n = 50 and n = 100 are representative of analyses at the level of individual participants, and n = 1,000 is representative of the analyses with aggregated data. Base rates were balanced in each response bias condition, but different response bias parameters governed each response bias condition, as might occur in manipulating response biases via payoff or implied base rate manipulations (e.g., Dube & Rotello, 2012).

Following Navarro et al. (2004) and Myung et al. (2007), parameter values were sampled from Jeffreys's (1961) noninformative distribution for each model. An important property of this sampling scheme is that it is the only one that makes sampling independent of the particular way in which the model is parameterized. In other words, we thereby ensured that the model recovery contrasts the different models independently of the particular way in which they are mathematically stated (see Navarro et al., 2004; for details on how to sample from Jeffreys's distribution for the recognition memory models, see Klauer & Kellen, 2011b).

Use of NML implies that everything else being equal, simple models (in terms of NML) will be selected more frequently and complex models less frequently than under AIC and BIC. The question is whether the implied reduction in model recovery for complex models is outweighed by the implied increase in model recovery for simple models. In the language of recognition memory models, relative to the use of AIC and BIC, not only should there be shifts in response bias toward or against certain models, but also overall discrimination performance should improve as well.

The simulation results presented in Table 1 show that NML outperforms AIC and BIC in terms of overall recovery accuracy, irrespective of sample size (see the "Overall" column). This difference is proportionally larger for the smaller sample sizes reported (e.g., n = 50), in which NML provides a 16 % and 26 % increase in overall accuracy in comparison with AIC and BIC, respectively. Also, note that AIC and BIC tend to overpenalize certain models so that they are recovered with chance or below-chance accuracy. For example, when considering realistic sample sizes for individual data (n = 50 or n = 100), MSD is virtually never correctly recovered when using AIC or BIC. In contrast, model recovery exceeds the 10 % chance baseline for each model under NML. As was expected, the differences between the model selection indices are proportionally smaller for larger sample sizes, reflecting the reduction in sampling variability, as well as the asymptotic properties of these indices.[6]

These results are thus consistent with the simulations previously reported by Klauer and Kellen (2011b), in which NML fared consistently better than AIC and BIC across binary-response ROC data with different numbers of points and different sample sizes in selecting from subsets of the above models. In the particular case of pairwise model recovery simulations, Klauer and Kellen (2011b) showed that model recovery rates obtained with NML are very close to the optimal rates that can be achieved, a finding that has previously been reported in the literature (Cohen et al., 2008, p. 698; Wagenmakers et al., 2004, p. 43). Similar results showing the advantages of model selection indices coming from the MDL principle over AIC and BIC have been reported in several studies that have focused on different types of models (e.g., Bozdogan, 2000; Cohen et al., 2008; Myung, 2000; Myung, Balasubramanian, & Pitt, 2000; Myung et al., 2007; Pitt & Myung, 2002; Pitt et al., 2002; Su et al., 2005). The BIC results are not surprising given the connections between BIC and the MDL framework. Like NML and FIA, BIC can be seen as an approximation to the Bayes factor, in which the term quantifying functional flexibility (i.e., FIA_f) is simply neglected (see Myung et al., 2006, p. 173), making BIC a less accurate approximation of the Bayes factor than FIA or NML. Although NML shows superior performance in terms of model recovery, it is important to note that NML (like AIC or BIC) was designed not with the purpose of maximizing model recovery accuracy, but of minimizing overfitting and generalization error.

In the simulation discussed above, a set of ten models was considered, although much of the literature is, in fact, focused on a subset of these. For example, more complex models such

[6] NML performance was also superior to AIC and BIC for ROCs obtained with sample sizes consistent with base rate manipulations (e.g., Bröder & Schütz, 2009).

Table 1 Model-recovery simulation results

Sample size	Method	Data-generating model										Overall
		1HTM	$2HTM_{(D_o = D_n)}$	$2HTM_{(D_o \geq D_n)}$	2HTM	EVSD	$UVSD_{(\sigma_o \geq \sigma_n)}$	UVSD	DPSD	MSD0	MSD	
50	AIC	.61	.50	.20	.22	.83	.64	.38	.09	.20	.01	.37
	BIC	.73	.58	.03	.13	.93	.54	.32	.00	.10	.00	.34
	NML	.67	.85	.36	.31	.67	.49	.32	.16	.27	.19	.43
100	AIC	.66	.57	.33	.30	.83	.71	.41	.20	.28	.03	.43
	BIC	.80	.67	.11	.20	.95	.63	.36	.05	.18	.00	.40
	NML	.72	.88	.44	.34	.73	.55	.36	.25	.35	.23	.49
1,000	AIC	.76	.73	.70	.43	.83	.85	.47	.56	.55	.26	.61
	BIC	.93	.87	.52	.39	.99	.82	.45	.41	.49	.05	.59
	NML	.87	.95	.70	.43	.90	.70	.45	.60	.63	.44	.67

Note. The values correspond to proportion of cases (out of 10,000) in which the data-generating model had the smallest value with a particular method (AIC, BIC, or NML). Column "Sample Size" refers to the number of items of each type (old and new), for each response bias condition. Column "Overall" indicates the overall proportion of data sets for which its data-generating model had the smallest value with a particular method. Chance-level recovery is .10.

as UVSD, DPSD, MSD0, and MSD were developed in part to account for discrepancies between the EVSD and ROC data, as well as to account for similar discrepancies found with other tasks (for a review, see Yonelinas & Parks, 2007). The same is true for 2HTM relative to 1HTM (e.g., Bayen, Murnane, & Erdfelder, 1996). When excluding simpler models from the selection set, an important concern is whether or not the data sets being used are *diagnostic*. For example, it could be the case that the ROC points are too close to each other to provide reliable information on the function's shape, or it could be that the data sets are well accounted for by a common restricted model such as EVSD. In both cases, the data can be seen as nondiagnostic. They would be well fit by all complex models that include the restricted model as a special case, and in consequence, the most simple of these complex models would invariably be selected as an application of the parsimony principle.

This issue was recently discussed by Jang et al. (2011), who focused on the comparison between the DPSD and the UVSD. Jang et al. noted that when comparing DPSD and UVSD using confidence-rating ROCs, many of the cases for which DPSD was chosen as the best model were cases in which the data were actually consistent with the EVSD. Given that both UVSD and DPSD have EVSD as a common restricted case, their predictions converge for these data sets, which severely reduces the ability to discriminate between the two models. Jang et al. pointed out that for these data sets, DPSD will tend to be selected simply because it is less complex than UVSD. In contrast, for the data sets that were not consistent with EVSD, UVSD was consistently selected over DPSD as the best model. Given that the role of both UVSD and DPSD is to account for data sets that are not well accounted for by EVSD, it is important to focus the comparisons on data sets that actually *diverge* from the latter (Jang et al., 2011, p. 756). The findings of Jang et al. indicate that model selection results from among more complex models (e.g., DPSD and UVSD) can be distorted when data sets that are well accounted by common restricted models (e.g., EVSD) are not screened out.

Because we screen out such data sets in some of the analyses reported below, let us consider how this affects model recovery from among the more complex models $2HTM_{(D_o \geq D_n)}$, 2HTM, $UVSD_{(\sigma_o \geq \sigma_n)}$, UVSD, DPSD, MSD0, and MSD. For these simulations, we excluded simulated data for which NML favored one of the simple models 1HTM, $2HTM_{(D_o = D_n)}$, and EVSD. Note that 1HTM is a restricted submodel of $2HTM_{(D_o \geq D_n)}$, 2HTM, DPSD, MSD0, and MSD; $2HTM_{(D_o = D_n)}$ of $2HTM_{(D_o \geq D_n)}$ and 2HTM; EVSD of $UVSD_{(\sigma_o \geq \sigma_n)}$, UVSD, DPSD, MSD0, and MSD. This excludes ROCs that can be accounted for by common restricted models, as well as ROCs whose points are too close to each other to be diagnostic. Note that in addition to the rationale by Jang et al. (2011) based on diagnosticity, this exclusion of nondiagnostic data sets gains support from similar recent MDL-based methods developed for optimizing model discriminability in experimental designs (see Myung & Pitt, 2009). The results from this simulation are presented in Table 2, and show that overall model recovery increases when excluding nondiagnostic data sets. The proportion of excluded data sets decreases as sample sizes increases, as was expected. As in the previous simulation, NML performs better overall than AIC and BIC and, unlike AIC and BIC, guarantees recovery rates above the chance baseline of 14.3 % for each model.

To summarize, MDL indices such as NML provide a principled quantification of flexibility in recognition memory models tailored to the parameters of a given study (in

Table 2 Model-recovery simulation results excluding non-diagnostic datasets

Sample size	Diagnostic data sets	Data-generating model							Overall
		Method	$2HTM_{(D_o \geq D_n)}$	2HTM	$UVSD_{(\sigma_o \geq \sigma_n)}$	UVSD	DPSD	MSD0	MSD	
50	.56	AIC	.74	.44 (.72)	.84	.49 (.90)	.43	.40	.01 (.39)	.48 (.63)
		BIC	.74	.44 (.72)	.84	.49 (.90)	.43	.40	.00 (.38)	.48 (.63)
		NML	.93	.59 (.94)	.68	.44 (.77)	.40	.51	.29 (.68)	.55 (.70)
100	.64	AIC	.79	.48 (.81)	.85	.49 (.91)	.52	.45	.04 (.40)	.52 (.68)
		BIC	.79	.48 (.81)	.86	.49 (.91)	.52	.45	.00 (.38)	.51 (.67)
		NML	.94	.58 (.97)	.69	.45 (.79)	.53	.58	.32 (.70)	.58 (.74)
1000	.83	AIC	.92	.53 (.95)	.89	.49 (.95)	.76	.64	.27 (.56)	.64 (.81)
		BIC	.92	.53 (.95)	.90	.49 (.95)	.77	.66	.05 (.42)	.62 (.80)
		NML	.97	.54 (.98)	.75	.48 (.87)	.84	.76	.47 (.79)	.69 (.85)

Note. The values correspond to proportion of cases (out of diagnostic data sets) in which the data-generating model had the smallest value with a particular method (AIC, BIC, or NML). Column "Sample Size" refers to the number of items of each type (old and new), for each response bias condition. Column "Overall" indicates the overall proportion of data sets for which its data-generating model had the smallest value with a particular method (AIC, BIC, or NML). Column "Diagnostic data sets" gives the proportion of data sets (out of 10,000) not screened-out as nondiagnostic. Values in parentheses are the proportions of data sets that were correctly recovered by the data-generating model (2HTM, UVSD, and MSD) or by a restricted version of that model ($2HTM_{(D_o \geq D_n)}$, $UVSD_{(\sigma_o \geq \sigma_n)}$, and MSD0, respectively)

terms of numbers of items and implemented base rates). The results reported in this section suggest that NML outperforms AIC and BIC in model recovery.

MDL analysis of ROC data for item recognition

The data sets

In this section, we reanalyze binary-response ROC data published in the literature. Following the meta-analysis by Dube and Rotello (2012), we consider only ROCs with at least 3 points, stemming from Curran et al. (2007), Henriques, Glowacki, and Davidson (1994), T. E. Parks (1966), Ratcliff, Sheu, and Gronlund (1992), Snodgrass and Corwin (1988), Starns, Ratcliff, and McKoon (2012), and Van Zandt (2000). We additionally include the data from the original experiments reported by Bröder and Schütz (2009), by Dube and Rotello (2012), by Dube et al. (2012), and by Starns et al., for a total of 41 data sets.

For the sake of comparison with the previous meta-analyses, we begin by analyzing aggregate frequencies, with aggregate data set as the unit. The analyses at this level and the data sets have several limitations that we then overcome by an analysis of data at the level of individual participants, where available.

The limitations comprise theoretical and practical issues. First, as was already noted by Bröder and Schütz (2009), many of the studies analyzed employed nonstandard designs with unusually high numbers of study–test blocks administered across several experimental sessions (Ratcliff et al., 1994; Starns et al., 2012; Van Zandt, 2000) and unusually short presentation times at the study phase (e.g., 50 and 200 ms per item; Ratcliff et al., 1994). In addition, these studies have small numbers of participants (e.g., 4 participants in a single experiment; Starns et al., 2012), and they comprise 20 out of the 41 data sets, giving them a high weight when the data set is the unit of analysis.

A practical limitation stems from the fact that the studies by Dube et al. (2012), Ratcliff et al. (1994), and Starns et al. (2012) implemented designs with more than one class of old and/or new items, calling for a joint analysis with different parameters permitted per item class. Unfortunately, the computational complexities of computing the NML index allow us to analyze only standard designs for aggregate data with one class of old and new items. For this reason, we followed Bröder and Schütz (2009, 2011) and Dube and Rotello (2012) in splitting up such data sets into several data sets that share the same distractor data and, therefore, are nonindependent for the analyses at the level of aggregate data. Nonindependence is also an issue for two data sets gleaned from Curran, DeBuse, and Leynes (2007; see Bröder & Schütz, 2011).

A final issue is that there is growing awareness of the problems that arise in aggregating data across individuals that differ systematically in performance for analyses of nonlinear models (Estes & Maddox, 2005; Klauer, 2010; Rouder & Lu, 2005). For example, even if all individual ROCs are generated from one of the models considered here, the aggregate ROC need not be well described by that model. Problems of this kind are compounded to the extent to which many observations are nested within comparatively few individuals (e.g., Riefer & Batchelder, 1988, 1991),

as was the case especially for the nonstandard designs just discussed.

For such reasons, Bröder and Schütz (2009), Dube and Rotello (2012), and Dube et al. (2012) conducted new experiments using more standard designs and used individual-participant ROCs in their analyses. Our main analysis focuses on these data comprising 186 participants, with participant as the level of analysis. As was already mentioned, one issue in selecting from among the more complex models is diagnosticity of the data, and given that smaller numbers of trials are available per participant than for the analysis of aggregate data, the diagnosticity issue is more pressing for the individual-level analyses than for the aggregate analyses.[7] For this reason, we also present an individual-level analysis in which nondiagnostic data are screened out as exemplified in the above model recovery study.

NML results for aggregated ROCs

Table 3 shows the NML analyses of the aggregate data for the independent data sets (upper half) and the nonindependent data sets (lower half). We conducted two analyses, a vote-counting analysis and a quantitative analysis. Vote counting is based on the frequencies with which each model is selected as best by NML across data sets. For the quantitative analysis, NML values are summed across data sets for each model.

The vote-counting analysis is in line with the model selection rationale according to which the model with the smallest NML value is the one that strikes the best balance between fit and flexibility in describing a given data set. Moreover, it is likely to be robust against a few outlying data sets, producing unusually large NML values and differences therein. The quantitative analysis, on the other hand, compares the models in terms of their ability to describe the joint data comprising all data sets, with different parameters per data set. It is probably more sensitive to outlying data sets, as well as to the problem of nonindependence of data sets that arises for many of the aggregate data, as was discussed above. Note that both levels of analysis need not converge on the same conclusion, given that it is possible that a model describes many data sets best but fails spectacularly (in terms of large NML values) for others.

Table 4 presents estimates for the parameters (other than response bias and guessing parameters) of the major models for the aggregate analyses. As can be seen, the parameter values are consistent with the values typically reported in the literature. Figures 5 and 6 show the ROCs for the aggregate data sets. Visual inspection of the ROCs shows that some of them appear to have a curvilinear shape. In other cases, the shape of the ROCs seems better described as linear or does not follow any of the forms that the models discussed here can account for. In any case, the shapes of many of the ROCs seem to differ from the smoother and more curvilinear shape that is almost invariably found in confidence-rating ROCs (Wixted, 2007; Yonelinas & Parks, 2007), although we acknowledge that an eyeball analysis of this kind has a subjective element.

Vote counting

Consider the frequencies of model selections given in brackets in the last row of Table 3 labeled "Total." A χ^2 test for equality of these counts across models reveals that the models differ significantly in their likelihood of being selected, $\chi^2(9) = 51.44$, $p < .001$.[8] It can be seen that models from the 2HTM family are chosen most frequently as the best model. In terms of individual models, there is a group of three models that fare best, $2HTM_{(D_o = D_n)}$, DPSD, and $2HTM_{(D_o \geq D_n)}$, with, in order, 13, 10, and 9 selections. These are followed by a second group of models, comprising EVSD, MSD0, and 1HTM, with, in order, 4, 3, and 2 selections, which are selected significantly less often than the first group ($p < .001$ by an exact binomial test on the cases that one of these six models was selected). A final group of models, 2HTM, $UVSD_{(\sigma_o \geq \sigma_n)}$, UVSD, and MSD, comprises models that are never selected and that are selected significantly less often as a group than the second group ($p = .004$). The differences between models within these three groups are not significant.

The quantitative analysis

The row labeled "Total" in Table 3 also shows the summed NML values. Interestingly, these present a

[7] For reference, Bröder and Schütz (2009) implemented a base rate manipulation with a total of 60 trials per response bias condition. The number of new items was 6, 15, 30, 45, and 54 across five response bias conditions. Dube et al. (2012) collected 96 trials per response bias condition, and the number of new items was 24, 32, 48, 64, and 72 across five response bias conditions. Dube and Rotello (2012) collected 5-point ROCs with balanced base rates in each response bias condition. In Experiment 1, 40 old and 40 new items were collected per response bias condition. In Experiment 2, 77 old and 77 new items were collected per response bias condition instead. In Van Zandt (2000), both base rate (Experiment 1) and payoff (Experiment 2) manipulations were used: There are slight variations in the number of trials per participant (at least in the raw data made available by the author), with the number of trials for any item type ranging between 137 and 640.

[8] A bootstrap analysis that does not rely on asymptotic approximations confirms that the differences between models are significant.

Table 3 NML results for aggregated datasets

Data	1HTM	$2HTM_{(D_o = D_n)}$	$2HTM_{(D_o \geq D_n)}$	2HTM	EVSD	$UVSD_{(\sigma_o \geq \sigma_n)}$	UVSD	DPSD	MSD0	MSD
Data sets with a single class of studied items (signal trials)										
Henriques et al. (1994), Exp. 1, control*	**15.91**	18.97	17.00	17.58	18.42	18.37	18.97	17.45	17.66	17.82
Henriques et al. (1994), Exp. 1, dysphorics*	**17.39**	20.34	18.55	19.14	19.99	20.09	20.70	18.98	19.19	19.35
Snodgrass & Corwin (1988), Exp. 1, high imagery*	19.89	**12.84**	14.19	14.47	13.43	15.04	15.62	14.15	14.34	14.45
Snodgrass & Corwin (1988), Exp. 1, low imagery*	18.11	**13.07**	14.42	14.33	13.60	15.21	15.59	14.32	14.50	14.61
Bröder & Schütz (2009), Exp.1	45.63	**21.24**	22.64	23.20	25.35	26.88	27.50	24.91	26.30	25.74
Bröder & Schütz (2009), Exp.2	28.06	33.51	**25.23**	25.82	42.74	29.56	30.20	26.87	28.52	27.77
Bröder & Schütz (2009), Exp.3	51.59	26.11	**25.75**	26.61	34.34	30.78	31.42	28.14	30.58	29.38
Dube & Rotello (2012), Exp. 1, pictures	60.07	**25.92**	28.10	28.73	36.47	33.81	34.47	31.35	33.75	32.71
Dube & Rotello (2012), Exp. 1, words	75.16	31.59	32.88	33.51	30.80	31.30	31.95	**30.01**	30.17	30.78
Dube & Rotello (2012), Exp. 2	68.64	45.27	34.46	35.11	83.18	29.39	30.05	29.48	**28.64**	29.25
T. E. Parks (1966), Exp. 1, fixed format	25.53	**21.98**	23.16	22.15	23.50	25.19	24.11	23.90	24.40	24.52
T. E. Parks (1966), Exp. 1, free format	23.58	**17.28**	18.46	19.01	19.00	20.65	21.25	19.36	19.87	19.98
Ratcliff et al. (1992), Exp.1, pure, weak	71.33	33.36	34.24	34.89	31.50	30.90	31.54	**29.25**	30.02	30.61
Ratcliff et al. (1992), Exp.1, pure, strong	93.01	35.46	31.50	32.15	33.66	30.98	31.62	**28.41**	31.06	29.76
Ratcliff et al. (1992), Exp.2, pure, weak	257.55	50.28	**35.69**	36.34	110.57	60.89	61.58	41.62	79.01	43.23
Ratcliff et al. (1992), Exp.2, pure, strong	383.00	99.10	96.38	97.03	51.52	52.38	53.06	**49.85**	52.88	51.46
Van Zandt (2000), Exp. 1, slow	62.34	**26.22**	28.44	28.68	28.99	29.59	30.26	27.92	28.72	29.23
Van Zandt (2000), Exp. 1, fast	47.43	30.09	32.28	31.69	**29.09**	31.85	32.52	30.23	30.83	31.49
Van Zandt (2000), Exp. 2	140.65	95.10	83.02	83.62	52.29	39.43	40.10	44.94	**37.05**	37.72
Subtotal 3-point ROCs (Single Class)	71.30 [2]	65.22 [2]	**64.16** [0]	65.52 [0]	65.44 [0]	68.72 [0]	70.88 [0]	64.89 [0]	65.70 [0]	66.23 [0]
Subtotal 5-point ROCs (single class)	1433.58 [0]	592.50 [5]	552.22 [3]	558.54 [0]	633.00 [1]	503.59 [0]	511.63 [0]	**466.24** [4]	511.79 [2]	473.64 [0]
Subtotal (single class)	1504.88 [2]	657.72 [7]	616.38 [3]	624.06 [0]	698.44 [1]	572.31 [0]	582.51 [0]	**531.13** [4]	577.48 [2]	539.87 [0]
Data sets with multiple classes of studied items (signal trials), or with inconsistent reconstruction										
Curran et al. (2007), Exp. 3, collapsed ratings*	34.71	32.41	30.47	31.09	**28.56**	30.09	30.72	29.31	29.10	29.36
Curran et al. (2007), Exp. 3, reconstructed binary*	49.54	25.33	27.40	27.93	**24.91**	27.20	25.85	26.13	26.40	26.70
Dube et al. (2012), Exp.1, study × 1	33.65	**22.50**	24.45	24.75	24.87	25.93	26.59	24.60	25.00	25.40
Dube et al. (2012), Exp.1, study × 5	33.97	30.64	**25.07**	25.55	28.57	28.79	29.47	27.10	28.29	28.06
Dube et al. (2012), Exp.2, study × 1	39.36	**29.44**	31.47	31.94	32.61	31.08	31.78	30.57	29.98	30.42
Dube et al. (2012), Exp.2, study × 10	51.46	30.30	27.76	28.21	**26.53**	27.86	28.55	26.54	26.91	27.34
Ratcliff et al. (1992), Exp.1, mixed, weak	70.39	**33.21**	35.15	35.79	40.16	38.52	39.16	35.68	38.00	37.04
Ratcliff et al. (1992), Exp.1, mixed, strong	110.77	40.67	**35.60**	36.24	39.82	40.66	41.30	37.69	40.73	39.04
Ratcliff et al. (1992), Exp.2, mixed, weak	155.99	48.99	50.03	50.66	37.15	30.46	31.11	**28.79**	30.46	30.05
Ratcliff et al. (1992), Exp.2, mixed, strong	158.16	104.46	74.61	75.24	40.57	40.72	41.37	39.13	**38.96**	39.58

Table 3 (continued)

Data	1HTM	$2HTM_{(D_o = D_n)}$	$2HTM_{(D_o \geq D_n)}$	2HTM	EVSD	$UVSD_{(\sigma_o \geq \sigma_n)}$	UVSD	DPSD	MSD0	MSD
Starns et al. (2012), HF, speed, study 1×	25.98	**24.19**	25.99	26.35	27.48	28.73	29.45	27.05	27.74	27.94
Starns et al. (2012), HF, speed, study 2×	28.84	**26.89**	27.39	27.75	28.18	28.49	29.19	27.55	27.66	28.09
Starns et al. (2012), HF, speed, study 4×	30.59	27.61	**27.59**	27.96	28.69	29.15	29.87	27.93	28.30	28.72
Starns et al. (2012), HF, accuracy, study 1×	40.75	28.93	28.48	28.83	27.00	27.16	27.86	**25.94**	26.31	26.73
Starns et al. (2012), HF, accuracy, study 2×	51.30	34.36	30.20	30.56	26.21	26.96	27.68	**25.59**	26.20	26.48
Starns et al. (2012), HF, accuracy, study 4×	54.14	49.64	29.82	30.18	29.28	28.60	29.33	**26.70**	28.23	27.59
Starns et al. (2012), LF, speed, study 1×	36.04	**26.81**	29.01	29.37	30.56	32.09	32.81	30.48	31.33	31.36
Starns et al. (2012), LF, speed, study 2×	30.62	27.42	**26.61**	26.98	28.36	27.72	28.43	26.67	26.80	27.25
Starns et al. (2012), LF, speed, study 4×	32.62	30.52	**26.51**	26.88	29.00	28.73	29.45	27.26	27.97	28.14
Starns et al. (2012), LF, accuracy, study 1×	59.55	27.12	29.20	29.56	27.69	28.33	29.05	**27.11**	27.60	27.90
Starns et al. (2012), LF, accuracy, study 2×	60.67	34.74	31.69	32.04	27.05	27.28	27.98	**26.30**	26.31	26.73
Starns et al. (2012), LF, accuracy, study 4×	63.74	43.38	**27.77**	28.13	28.43	29.93	30.64	28.32	29.59	29.17
Subtotal 3-point ROCs (multiple classes)	84.25 [0]	57.74 [0]	57.87 [0]	59.03 [0]	**53.47** [2]	57.29 [0]	56.57 [0]	55.44 [0]	55.50 [0]	56.06 [0]
Subtotal 5-point ROCs (multiple classes)	1168.60 [0]	721.82 [6]	644.38 [6]	652.99 [0]	608.24 [1]	607.20 [0]	621.07 [0]	**576.99** [6]	592.38 [1]	593.05 [0]
Subtotal (multiple classes)	1252.85 [0]	779.56 [6]	702.25 [6]	712.01 [0]	661.71 [3]	664.49 [0]	677.64 [0]	**632.43** [6]	647.88 [1]	649.11 [0]
Total 3-point ROCs	155.55 [2]	122.96 [2]	122.04 [0]	124.55 [0]	**118.91** [2]	126.01 [0]	127.46 [0]	120.33 [0]	121.19 [0]	122.29 [0]
Total 5-point ROCs	2602.18 [0]	1314.32 [11]	1196.60 [9]	1211.53 [0]	1241.24 [2]	1110.79 [0]	1132.69 [0]	**1043.23** [10]	1104.17 [3]	1066.69 [0]
Total	2757.73 [2]	1437.28 [13]	1318.64 [9]	1336.07 [0]	1360.14 [4]	1236.80 [0]	1260.15 [0]	**1163.56** [10]	1225.37 [3]	1188.98 [0]

Note. Bold values correspond to model with lowest NML. Values inside squared-brackets indicate the number of aggregate data sets for which a given model provided the lowest individual NML value. Data sets with an asterisk (*) correspond to 3-point ROCs. HF = high-frequency words, and LF = low-frequency words

Table 4 Parameter estimates for a selection of models for the aggregated datasets

Data	2HTM		UVSD		DPSD		MSD0		MSD		
	D_o	D_n	μ_o	σ_o	μ_o	R	μ_o	λ	μ_o	μ_o^*	λ
Curran et al. (2007), Exp. 3, collapsed ratings*	.58	.32	1.40	1.21	0.98	.28	1.54	.86	1.54	0.00	.86
Curran et al. (2007), Exp. 3, reconstructed binary*	.20	.23	0.53	0.85	0.61	.00	0.61	1	0.61	0.00	1
Henriques et al. (1994), Exp. 1, control*	.78	.00	4.57	3.86	0.00	.78	7.19	.78	7.19	0.00	.78
Henriques et al. (1994), Exp. 1, dysphorics*	.77	.03	3.75	3.23	0.00	.77	7.27	.77	7.29	0.00	.77
Snodgrass & Corwin (1988), Exp. 1, high imagery*	.60	.69	1.87	0.97	1.90	.00	1.90	1	1.90	0.00	1
Snodgrass & Corwin (1988), Exp. 1, low imagery*	.21	.40	0.78	0.86	0.83	.00	0.83	1	0.83	0.00	1
Bröder & Schütz (2009), Exp.1	.54	.50	1.61	1.13	1.24	.27	1.62	.95	6.56	1.24	.27
Bröder & Schütz (2009), Exp.2	.70	.31	2.69	2.20	0.52	.67	2.83	.77	4.25	0.51	.67
Bröder & Schütz (2009), Exp.3	.56	.45	1.71	1.38	0.95	.41	1.88	.83	6.34	0.95	.41
Dube & Rotello (2012), Exp. 1, pictures	.70	.69	2.58	1.52	1.43	.55	2.50	.90	6.61	1.43	.55
Dube & Rotello (2012), Exp. 1, words	.43	.38	1.21	1.16	0.93	.19	1.40	.83	1.40	0.00	.83
Dube & Rotello (2012), Exp. 2	.46	.31	1.52	1.74	0.56	.40	2.18	.64	2.32	0.15	.59
Dube et al. (2012), Exp.1, study 1×	.40	.45	1.27	1.25	0.90	.23	1.57	.77	1.57	0.00	.77
Dube et al. (2012), Exp.1, study 5×	.76	.54	2.45	1.40	1.36	.58	2.29	.93	6.41	1.36	.58
Dube et al. (2012), Exp.2, study 1×	.36	.34	1.07	1.35	0.66	.23	1.60	.64	1.60	0.00	.64
Dube et al. (2012), Exp.2, study 10×	.67	.52	1.95	1.17	1.52	.31	1.96	.94	1.97	0.10	.94
T. E. Parks (1966), Exp. 1, fixed format	.00	.76	1.06	0.19	1.54	.00	1.54	1	1.54	1.54	.63
T. E. Parks (1966), Exp. 1, free format	.49	.52	1.46	1.07	1.30	.12	1.50	.95	3.97	1.30	.12
Ratcliff et al. (1992), Exp.1, mixed, weak	.34	.30	0.95	1.19	0.65	.19	1.28	.72	5.94	0.65	.19
Ratcliff et al. (1992), Exp.1, mixed, strong	.55	.40	1.46	1.12	1.13	.25	1.48	.94	5.41	1.13	.25
Ratcliff et al. (1992), Exp.1, pure, weak	.37	.30	1.02	1.16	0.74	.18	1.26	.78	1.68	0.52	.44
Ratcliff et al. (1992), Exp.1, pure, strong	.54	.40	1.51	1.21	1.04	.30	1.62	.88	4.90	1.04	.30
Ratcliff et al. (1992), Exp.2, mixed, weak	.60	.55	1.96	1.26	1.40	.34	2.00	.91	2.89	1.25	.47
Ratcliff et al. (1992), Exp.2, mixed, strong	.86	.63	2.82	1.16	2.27	.41	2.68	.99	2.68	0.00	.99
Ratcliff et al. (1992), Exp.2, pure, weak	.64	.51	2.17	1.45	1.17	.49	2.16	.89	7.20	1.17	.49
Ratcliff et al. (1992), Exp.2, pure, strong	.77	.71	2.56	1.10	2.24	.28	2.48	.99	6.11	2.24	.28
Starns et al. (2012), HF, speed, study 1×	.16	.09	0.35	1.15	0.15	.13	1.38	.28	6.32	0.15	.13
Starns et al. (2012), HF, speed, study 2×	.29	.15	0.67	1.23	0.36	.19	1.29	.52	1.29	0.00	.52
Starns et al. (2012), HF, speed, study 4×	.34	.18	0.80	1.22	0.45	.22	1.28	.61	1.28	0.00	.61
Starns et al. (2012), HF, accuracy, study 1×	.25	.16	0.67	1.13	0.48	.11	1.02	.65	1.48	0.33	.31
Starns et al. (2012), HF, accuracy, study 2×	.34	.20	0.87	1.11	0.68	.13	1.07	.80	4.59	0.68	.13
Starns et al. (2012), HF, accuracy, study 4×	.46	.22	1.16	1.16	0.81	.23	1.32	.85	5.93	0.81	.23
Starns et al. (2012), LF, speed, study 1×	.32	.30	0.90	1.16	0.62	.18	1.20	.73	5.98	0.62	.18
Starns et al. (2012), LF, speed, study 2×	.50	.30	1.34	1.37	0.70	.35	1.68	.74	1.68	0.00	.74
Starns et al. (2012), LF, speed, study 4×	.60	.36	1.64	1.35	0.88	.43	1.79	.83	5.83	0.88	.43
Starns et al. (2012), LF, accuracy, study 1×	.44	.41	1.32	1.15	1.04	.18	1.47	.86	2.09	0.88	.38
Starns et al. (2012), LF, accuracy, study 2×	.59	.45	1.71	1.18	1.32	.26	1.79	.91	1.79	0.00	.91
Starns et al. (2012), LF, accuracy, study 4×	.70	.50	2.01	1.12	1.62	.30	1.95	.98	6.04	1.62	.30
Van Zandt (2000), Exp. 1, slow	.56	.60	1.83	1.26	1.32	.31	1.94	.88	2.99	1.24	.39
Van Zandt (2000), Exp. 1, fast	.33	.46	1.06	1.04	1.03	.02	1.15	.91	1.15	0.00	.91
Van Zandt (2000), Exp. 2	.32	.20	0.94	1.21	0.67	.15	1.30	.70	1.30	0.00	.70

Note. Study 1 ×, 2 ×, 4 ×, 5 ×, and 10 × refer to the amount of times that items were presented in the study phase. Data sets with an asterisk (*) correspond to 3-point ROCs. HF = high-frequency words, and LF = low-frequency words

different ordering of the models. From best to worst, the models are ordered as DPSD, MSD, MSD0, $\text{UVSD}_{(\sigma_o \geq \sigma_n)}$, UVSD, $\text{2HTM}_{(D_o \geq D_n)}$, 2HTM, EVSD, $\text{2HTM}_{(D_o = D_n)}$, 1HTM. The quantitative differences between models

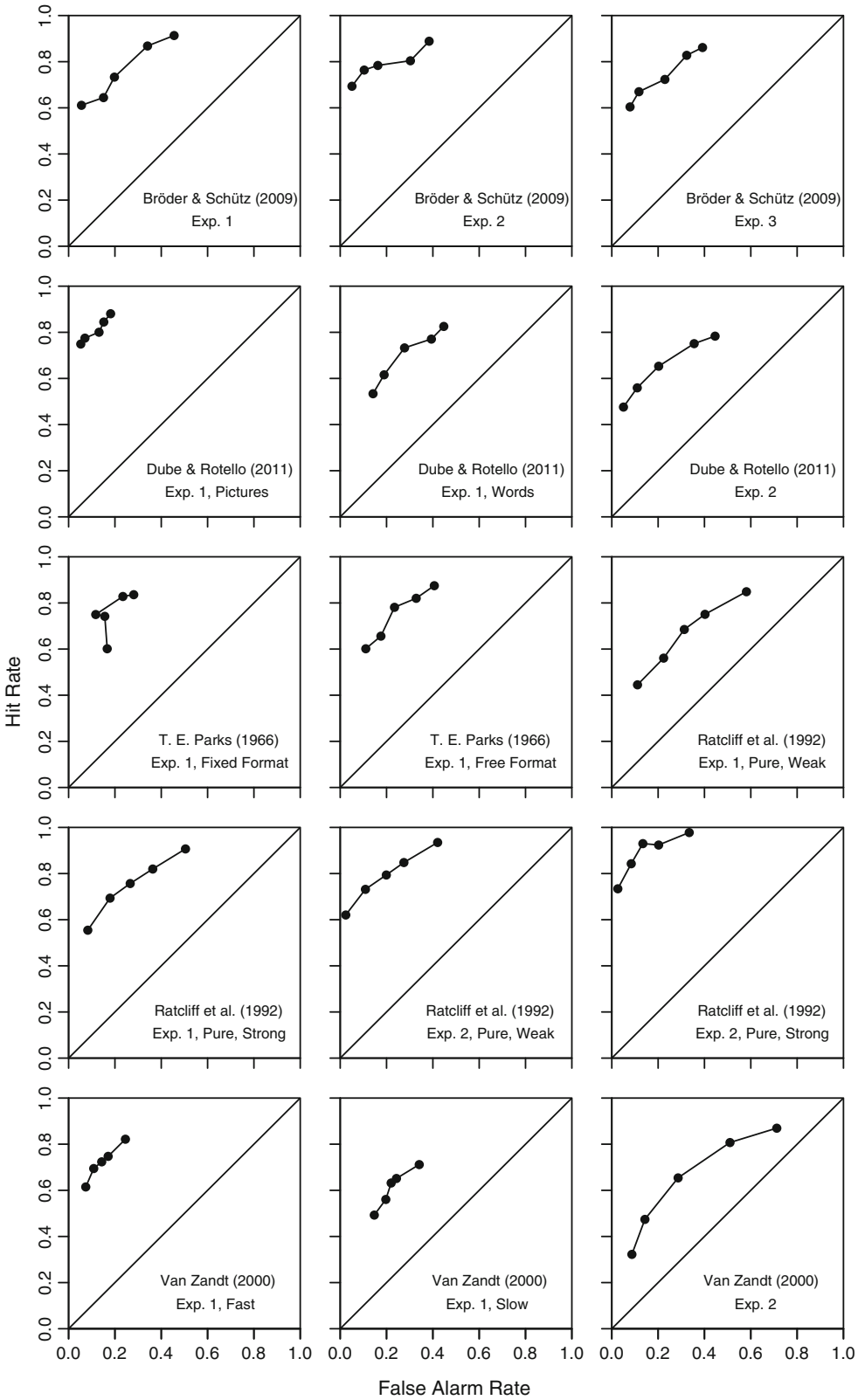

Fig. 5 Observed five-point receiver operating characteristics (ROCs) for aggregated data sets that included a single class of studied items (signal trials)

in NML values are significant across data sets according to the Friedman rank-sum test, $\chi^2(9) = 132.53$, $p < .001$. DPSD significantly outperforms all models, as attested to by Wilcoxon tests (largest $p < .014$), except

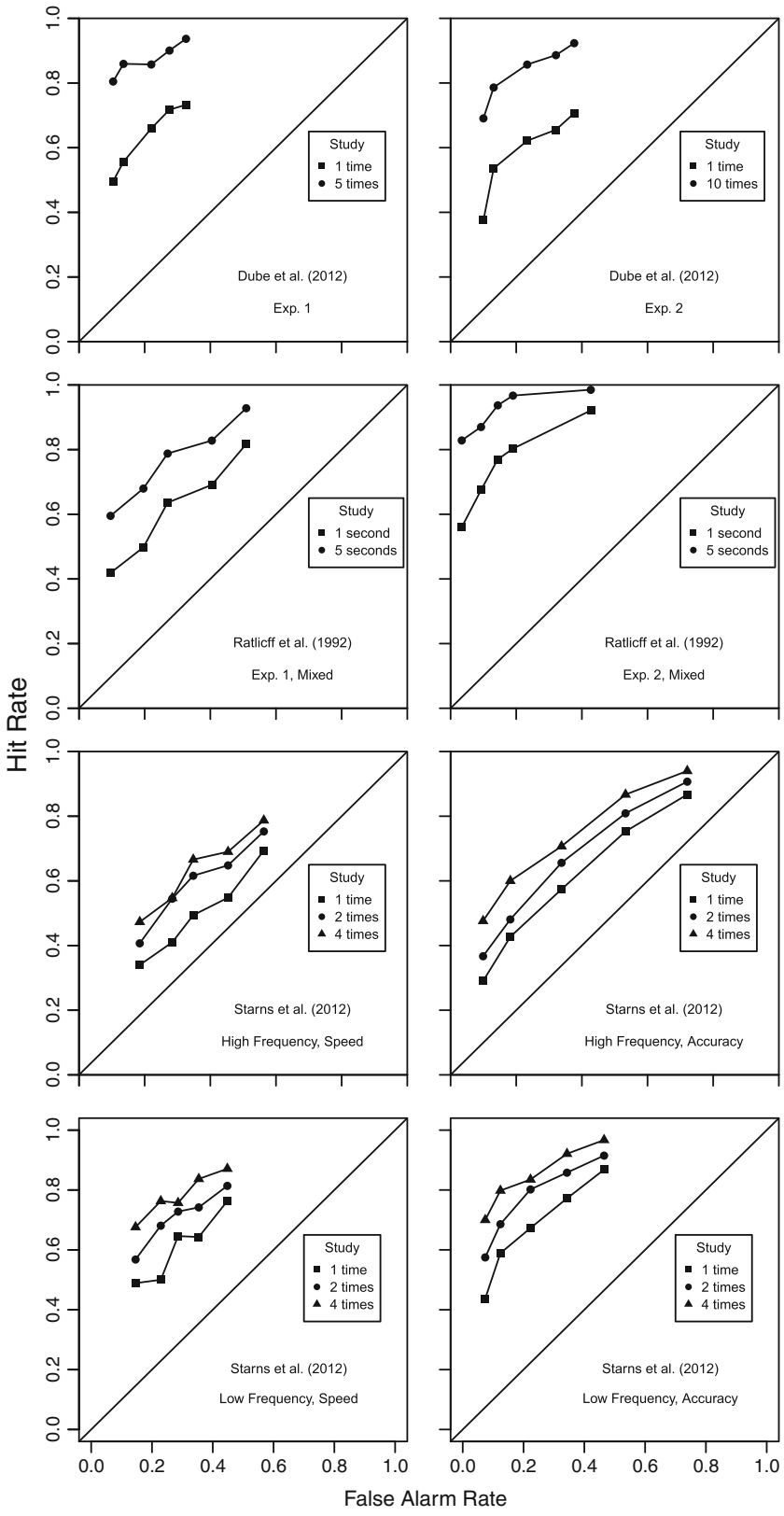

Fig. 6 Observed five-point receiver operating characteristics (ROCs) for aggregated data sets that included more than one class of studied items (signal trials)

$2HTM_{(D_o \geq D_n)}$, for which the difference is not significant (p=.15). This indicates that the numerically sizable difference between DPSD and $2HTM_{(D_o \geq D_n)}$ in summed NML values is driven by only a few data sets and

does not generalize across data sets. Because the quantitative analysis is analogous to the quantitative comparisons in terms of G^2 from previous meta-analyses contrasting UVSD and 2HTM, we also tested UVSD against 2HTM using a Wilcoxon test and found that they do not differ significantly ($p = .22$); the same is true for comparing $UVSD_{(\sigma_o \geq \sigma_n)}$ and $2HTM_{(D_o \geq D_n)}$ ($p = .32$).

Summary of the aggregate analyses

The aggregate analyses do not present a clear winner. Nevertheless, they converge on a couple of conclusions. In terms of vote counting, as well as in the quantitative analyses, DPSD performed quite well. Likewise, models from the 2HTM family (i.e., $2HTM_{(D_o = D_n)}$ and $2HTM_{(D_o \geq D_n)}$) performed well in vote counting, and in the quantitative analysis, the $2HTM_{(D_o \geq D_n)}$ is the only model that is not significantly outperformed by DPSD. Nor is $2HTM_{(D_o \geq D_n)}$ significantly outperformed by any other model in this analysis.

Previous meta-analyses focused on comparing UVSD and 2HTM. Regarding this comparison, neither in the vote-counting analysis nor in the quantitative analysis was UVSD ($UVSD_{(\sigma_o \geq \sigma_n)}$) found to perform better than 2HTM ($2HTM_{(D_o \geq D_n)}$), which is also true if the vote-counting analysis is restricted to just these pairs of models. In other words, once the ability of these models to fit data in general is factored in, there is no decisive evidence in these analyses for or against one of these models.

As was already mentioned, the analyses at the level of aggregate data are limited in several ways concerning a large proportion of data sets with nonstandard designs, the issue of nonindependence of some of the data sets, and the problems related to aggregating data across heterogeneous participants. For these reasons, our main analysis focuses on individual data to see whether these present a more consistent picture across vote-counting analysis and quantitative analysis and confirm the suggestive evidence for DPSD and the 2HTM family that is nevertheless found in these aggregate analyses.

NML results for individual-participant ROCs

Individual data are available from experiments by Bröder and Schütz (2009), Dube and Rotello (2012), and Dube et al. (2012) using relatively standard recognition designs, comprising 186 participants. Following Bröder and Schütz (2009) and Dube and Rotello, we focus our analysis on individual data sets from recent studies specifically designed to compare these models, therefore not including the 15 individual data sets from Van Zandt (2000), who implemented experimental designs that are nonstandard in several respects.[9]

The data from Dube et al. (2012) comprised two types of old items, weak and strong items. At the level of the individual analyses, we are able to compute NML values for the model that analyses new items and weak and strong old items jointly, with different parameters for the two kinds of old items. In consequence, the issue of nonindependence does not arise in the analyses by participant.[10]

Vote counting

Table 5 presents the NML analyses. Consider the upper half of the table, labeled "All Data sets," first. The row labeled "Total" provides the frequencies of model selections in brackets for each model. A χ^2 test for equality of these counts across models reveals that the models differ significantly in their likelihood of being selected, $\chi^2(9) = 407.12$, $p < .001$ (see also Footnote 8). As can be seen, $2HTM_{(D_o = D_n)}$ emerges as the clear winner, with 48 % of individuals best described by it, followed at a significant distance ($p < .001$) by 1HTM (24 %), which does not differ significantly from 2HTM (18 %; $p = .25$), and fewer than 4 % per model for the remaining models, each of which was selected significantly less often than 2HTM (largest $p < .001$).

This suggests that many of the individual-level data sets are not diagnostic for discriminating between the more complex models, because they are already well described by simpler models such as 1HTM and $2HTM_{(D_n = D_o)}$ (Jang et al., 2011). Another possibility is, of course, that the simple models and,

[9] For the Van Zandt (2000) data, models from the 2HTM family of models provide the best NML account for 10 of the 15 individual data sets. There is one outlier (participant 1, Experiment 2) for which the models from the 2HTM family perform extremely poorly, determining the summed NML for these data sets so that the 2HTM family does not fare well in the quantitative analysis. Including the outlier, the ranking of the models in terms of summed NML is, from best to worst, DPSD, MSD0, MSD, $UVSD_{(\sigma_o \geq \sigma_n)}$, UVSD, EVSD, 2HTM, $2HTM_{(D_o \geq D_n)}$, $2HTM_{(D_o = D_n)}$, and 1HTM, with summed NML values of 362.06, 368.02, 371.38, 378.98, 382.52, 385.05, 452.54, 452.67, 469.62, and 546.34, respectively. Removing the outlier considerably changes the ranking of the models in terms of summed NML, from best to worst being DPSD, $2HTM_{(D_o = D_n)}$, MSD0, MSD, 2HTM, $2HTM_{(D_o \geq D_n)}$, $UVSD_{(\sigma_o \geq \sigma_n)}$, EVSD, UVSD, and 1HTM, with summed NML values of 338.01, 342.05, 343.25, 346.64, 346.71, 347.40, 353.79, 356.57, 356.67, and 427.97, respectively.

[10] We followed Dube et al.'s (2012) specification of UVSD for this analysis in permitting different μ_o parameters for weak and strong items and keeping σ_o (and response bias parameters) equal for both kinds of items. For the other models, different parameters are permitted for weak and strong items for all parameters other than the response bias and guessing parameters. For the case of $2HTM_{(D_o = D_n)}$, we somewhat arbitrarily let $D_o = D_n$ for the weak items and permitted a new D_o parameter for the strong items.

Table 5 NML results for individual data

All data sets

Data	Participants	1HTM	$2HTM_{(D_o = D_n)}$	$2HTM_{(D_o \geq D_n)}$	2HTM	EVSD	$UVSD_{(\sigma_o \geq \sigma_n)}$	UVSD	DPSD	MSD0	MSD
Bröder & Schütz (2009), Exp. 3	40	608.92 [8]	**569.29** [25]	581.06 [3]	587.38 [4]	635.55 [0]	652.46 [0]	668.65 [0]	620.62 [0]	629.61 [0]	629.22 [0]
Dube & Rotello (2012), Exp. 1, pictures	39	817.37 [11]	783.04 [9]	779.43 [2]	**709.77** [14]	788.76 [1]	795.98 [0]	779.79 [1]	764.15 [0]	773.06 [1]	773.74 [0]
Dube & Rotello (2012), Exp. 1, words	36	696.73 [9]	630.70 [20]	637.21 [0]	**626.43** [5]	687.94 [2]	696.96 [0]	693.81 [0]	662.59 [0]	674.06 [0]	675.14 [0]
Dube & Rotello (2012), Exp. 2	24	536.70 [2]	500.92 [10]	485.60 [2]	**484.06** [3]	551.94 [0]	510.15 [1]	512.20 [0]	488.17 [3]	493.12 [2]	492.87 [1]
Dube et al. (2012), Exp. 1	21	473.74 [7]	**457.07** [11]	465.61 [0]	466.04 [3]	515.38 [0]	513.45 [0]	522.38 [0]	497.10 [0]	504.38 [0]	515.43 [0]
Dube et al.(2012), Exp. 2	26	613.58 [7]	**584.88** [14]	597.54 [0]	585.58 [4]	638.13 [1]	642.54 [0]	647.09 [0]	625.86 [0]	634.33 [0]	648.65 [0]
Total	186	3747.04 [44]	3525.90 [89]	3546.46 [7]	**3459.27** [33]	3817.69 [4]	3811.55 [1]	3823.92 [1]	3658.50 [3]	3708.57 [3]	3735.05 [1]

Diagnostic data sets

Data	Participants	$2HTM_{(D_o \geq D_n)}$	2HTM	$UVSD_{(\sigma_o \geq \sigma_n)}$	UVSD	DPSD	MSD0	MSD
Bröder & Schütz (2009), Exp. 3	7	103.34 [3]	**100.60** [4]	113.98 [0]	115.65 [0]	106.91 [0]	108.82 [0]	108.48 [0]
Dube & Rotello (2012), Exp. 1, pictures	18	399.37 [2]	**322.02** [14]	378.91 [0]	354.82 [1]	363.43 [0]	366.67 [1]	367.59 [0]
Dube & Rotello (2012), Exp. 1, words	5	107.64 [0]	**89.87** [5]	111.17 [0]	98.64 [0]	106.00 [0]	107.64 [0]	107.90 [0]
Dube & Rotello (2012), Exp. 2	12	256.81 [2]	**250.30** [3]	258.31 [1]	253.65 [0]	251.92 [3]	252.40 [2]	252.55 [1]
Dube et al. (2012), Exp. 1	3	71.21 [0]	**66.40** [3]	73.59 [0]	74.24 [0]	71.77 [0]	73.08 [0]	74.53 [0]
Dube et al. (2012), Exp. 2	4	97.73 [0]	**83.75** [4]	96.59 [0]	92.34 [0]	93.14 [0]	94.89 [0]	96.85 [0]
Total	49	1036.09 [7]	**912.96** [33]	1032.54 [1]	989.35 [1]	993.17 [3]	1003.50 [3]	1007.90 [1]

Note. Bold values correspond to model with lowest sum of NML values. Values inside squared-brackets indicate the number of participants for which a given model provided the lowest individual NML value. The "Diagnostic data sets" are data sets that were not better accounted for in terms of NML by EVSD, $2HTM_{(D_o=D_n)}$, or EVSD

in particular, $2HTM_{(D_o = D_n)}$ are truly the most appropriate models. If so, the quantitative analysis should mirror their superiority. In contrast, if the good performance of these simple models reflects a diagnosticity issue, we can expect them to perform quite poorly on the still substantial proportion of cases in which none of them provided the best description (i.e., in the diagnostic data sets), implying that their performance in the quantitative analysis may not mirror their vote-counting superiority.

Quantitative analysis

Table 5 also presents the summed NML values. They agree better with the vote-counting results than in the aggregate-level analyses, but there are a couple of suggestive discrepancies. From best to worse, the models are ordered as 2HTM, $2HTM_{(D_o = D_n)}$, $2HTM_{(D_o \geq D_n)}$, DPSD, MSD0, MSD, 1HTM, $UVSD_{(\sigma_o \geq \sigma_n)}$, EVSD, and UVSD. These differences generalize across participants, as attested to by a Friedman rank-sum test, $\chi^2(9) = 647.01$, $p < .001$. 2HTM significantly outperformed all models according to Wilcoxon tests (largest $p < .001$) other than $2HTM_{(D_o = D_n)}$ ($p = .06$) and $2HTM_{(D_o \geq D_n)}$ ($p = .07$). 1HTM fared second-best in vote counting but dropped to the seventh place in the quantitative analysis, suggesting that it performs poorly on more diagnostic data sets.

To summarize, the individual-level analyses suggest that models from the 2HTM family strike the best balance between fit and flexibility both in terms of vote counting and in terms of the quantitative analysis. On the other hand, the good performance of DPSD seen for the aggregate data is not replicated at the level of the individuals' data. One possibility is that this reflects a diagnosticity issue, as discussed by Jang et al. (2011); another possibility is that the relatively good performance of DPSD in the aggregate analysis is an aggregation artifact;.These two possibilities that are not mutually exclusive. Our final set of analyses excluded nondiagnostic data sets, as described in the above model recovery study.

Analyses restricted to diagnostic data sets

The final set of analyses is restricted to the individual-level data sets for which none of the simpler models (1HTM, $2HTM_{(D_o = D_n)}$, EVSD) was selected by NML, for a total of 49 data sets. The model selection results are shown in the lower half of Table 5. In terms of vote counting, the selection frequencies for the seven more complex models differed significantly from each other, $\chi^2(6) = 116.57$, $p < .001$ (see Footnote 8). The clear winner is 2HTM, with 67 % of data sets favoring it, with a vote count that is significantly larger than the vote count for any of the other models (largest $p < .001$).

The same is true for the quantitative analysis. The differences between models in NML values generalize across data sets, as attested to by a significant Friedman rank-sum test, $\chi^2(6) = 94.16$, $p < .001$. As in vote counting, 2HTM performed best: It was associated with the smallest summed NML, and in Wilcoxon tests, the differences between 2HTM and any other model were significant (largest $p < .001$).

Relative to the analysis on all individual data sets, an important change in the results is that the performance of UVSD becomes slightly better than DPSD (although not significant with a Wilcoxon test, $p = .52$), which suggests that when excluding nondiagnostic data sets the flexibility of UVSD is better justified by the remaining data. This outcome mimics, to a certain extent, Jang et al.'s (2011) above-described results and reinforces the notion that model selection efforts need to consider potential biases produced by nondiagnostic data.

Note that the percentage of data sets excluded as nondiagnostic is substantial, but this was to be expected on the basis of the rates of exclusion reported for the simulation study on model recovery summarized in Table 2, which represent a best-case scenario in the sense that one of the more complex models truly generated the data. Nevertheless, it is noteworthy that the analyses of all individual data, as well as the analyses of the diagnostic subset of them, converge without exception in that they favor 2HTM significantly relative to each of the other six complex models (significance being only marginal for the comparison of 2HTM and $2HTM_{(D_o \geq D_n)}$ in the quantitative analysis of all individual data sets as detailed above).[11]

General discussion

The purpose of this article is to bring to bear modern developments in model selection based on the MDL principle on the debate as to which of several prominent recognition memory models provides the best measurement model for ROC data. The NML index derived from the MDL principle overcomes several limitations of previous model selection methods, such as those based on AIC, BIC, and the data-informed PBCM; in particular, it provides a principled and intuitively plausible quantification of model flexibility due to functional form. The penalty for flexibility built into

[11] One possibility is that the criterion used to screen out data sets excluded curvilinear ROCs (consistent with EVSD) that would lead to a rejection of both 2HTM and $2HTM_{(D_o \geq D_n)}$. We checked this by redoing the analysis, this time excluding only data sets for which 1HTM or $2HTM_{(D_o = D_n)}$ had the smallest NML value. Out of the remaining 53 individual data sets (only 4 additional data sets), 2HTM provided the best summed NML and was the most frequently selected model. Significance tests in the vote counting and quantitative analysis did not differ from the ones reported in body of text.

NML can be given a straightforward interpretation: It quantifies the ability of the model to fit data in general. The NML index thereby addresses the concerns raised by Roberts and Pashler (2000) and others (e.g., Chechile, 1998; Myung, 2000) in a head-on fashion: A model' s ability to fit the observed data should be put in relation to its ability to provide good fits in general. A model recovery study confirmed that model selection based on NML outperforms model selection based on AIC, as well as BIC, for the models and kind of data sets considered here. Previously, Klauer and Kellen (2011a, b) showed that selection by NML closely approaches the optimal recovery rates that can be attained in pairwise comparisons of recognition memory models (see also Cohen et al., 2008; Wagenmakers et al., 2004). Beyond model recovery, NML will also perform its objective of identifying the model that provides the best compression of the data, minimizing overfitting and generalization error, when none of the candidate models truly generated the data—a fact that we consider reassuring given that we feel it unlikely that any of the models considered here provides more than a rough first approximation of the actual data-generating process.

We focus here on binary-response ROC data with experimental manipulations of response bias that have been the subject of recent debates assessing the relative abilities of UVSD and 2HTM to fit such data (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube et al., 2011; Dube et al., 2012; Klauer & Kellen, 2011a, b). We extended this work in several ways: By bringing modern selection methods in terms of the MDL principle to bear on the issue, by considering a much wider range of candidate models, and by including additional data. We analyzed the data both at the aggregate level and at the level of individual participants, where individual-level data were available. At each level, we conducted a vote-counting analysis of the frequencies with which the different models were selected as best by the MDL index NML and a quantitative analysis focusing on summed NML values. We pointed out several theoretical and practical limitations of the analyses of the aggregate data that have the potential to compromise conclusions based on them. We nevertheless present the aggregate analyses for comparison with previous meta-analyses that focused on aggregate data.

In fact, the NML analyses of the aggregate data did not agree well between vote-counting and quantitative analysis. Nevertheless, for both vote-counting and quantitative analysis, they suggested a preference for DPSD and models from the 2HTM family, although the preference was much more clearly expressed for DPSD than for the 2HTM family in the quantitative analysis. This also provides a correction of previous meta-analyses that compared UVSD and 2HTM on a subset of the present data and concluded that 2HTM performed significantly worse than UVSD, on the basis of quantitative (G^2-based) analyses that did not take differences in model flexibility due to functional form into account.

Because of the problems discussed for the aggregate analysis, our main interest was on individual-participant data sets for which many of the problems associated with the aggregate data sets do not arise. This level of analysis is, however, more vulnerable to the issue of possible nondiagnosticity of data sets given the comparatively small numbers of trials typically administered per participant. To assess the impact of nondiagnosticity, we conducted the NML analyses on all the 186 individual data sets available, as well as after excluding all individuals whose data were best accounted for by one of the simpler models, thereby excluding nondiagnostic data sets as per Jang et al. (2011). A preparatory model recovery study implementing this exclusion scheme showed that overall selection accuracy is thereby increased for the complex models, although substantial proportions of data sets have to be excluded as nondiagnostic. Again, NML performed better than AIC and BIC in terms of overall recovery performance.

The individual-level data permit a simple summary. Models from the 2HTM family emerged as the clear winner. For the analyses including diagnostic and nondiagnostic data sets, $2HTM_{(D_o = D_n)}$ and 2HTM were preferred over all other models in terms of vote-counting and quantitative analyses, respectively; for the analyses restricted to diagnostic data sets and the more complex models, 2HTM was preferred over all other models in terms of both vote-counting and quantitative analysis.

A number of conclusions can be drawn. First, there is little evidence at any level of analysis that the mathematical complexity implied by UVSD is supported by the present data. Second, the goal of measuring individuals' recognition memory performance from binary old/new recognition judgments parsimoniously (Snodgrass & Corwin, 1988) is best fulfilled by members of the 2HTM family. This result is especially relevant for cases in which the 2HTM is used as a building block in measurement models for extended recognition memory designs such as source monitoring (Bayen et al., 1996; Klauer & Kellen, 2010; Klauer & Wegener, 1998; Meiser & Bröder, 2002), since some criticisms of these models focused on binary-response ROC data like the ones analyzed here (e.g., Kinchla, 1994).

Note that the adequacy of measurement is a distinct goal from one's attempts to characterize in a more fine-grained manner the processes that actually generated the data (see Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002). In addition, it is a common finding in several fields of research that models based on discrete states provide suitable descriptions of what are believed to be continuous processes (e.g., Dutilh, Wagenmakers, Visser, & van der Maas, 2011; Ratcliff & McKoon, 2001; Schmittmann, Visser, &

Raijmakers, 2006). For such reasons, we agree with George Box's famous words: "All models are wrong, but some are useful" (Box, 1979, p. 202).

The present results show that the 2HTM is a viable model for the present data, followed by the DPSD, on the basis of its good performance for the aggregate data. This conclusion is tempered by the observation that these models (i.e., 2HTM and DPSD) are also among the least flexible among the more complex models (such as UVSD, MSD0, and MSD) in the principled flexibility metric provided by NML. In consequence, it may be the case that the ROC data, even after excluding obviously nondiagnostic data sets, are still not strong enough to provide reliable support for the mathematical complexity of more complex models, such as UVSD or MSD. The present results thereby provide formal and documented encouragement for recent endeavors to develop alternative tasks and tests that have the potential of providing even more diagnostic data. Promising avenues comprise administering different recognition memory tasks in one session and modeling them jointly (e.g., Jang, Wixted, & Huber, 2009; Kellen, Klauer, & Singmann, 2012), the development of new paradigms focusing on aspects in which different models diverge (e.g., O'Connor, Guhl, Cox, & Dobbins, 2011), and the use of response latency data (e.g., Dube et al., 2012; Province & Rouder, 2012; Starns et al., 2012). The analysis of response latency data is becoming increasingly popular, but so far the results have been mixed, since some are consistent with continuous models (Dube et al., 2012; Starns et al., 2012) and others with discrete-state models (Province & Rouder, 2012). Accounting for response latencies represents an important effort in recognition memory modeling, but this effort still needs to be complemented with an assessment of model flexibility (see Luce, 1986, p. 344) and an assessment of the ability of response latency data to distinguish between different processing accounts (e.g., Ratcliff, 1988). The use of alternative tasks and response latency data might lead to data sets that justify the complexity of some models, with the potential to change and correct the present results.

Having said this, it should be noted that recent such work with confidence-rating ROCs obtained results favoring the 2HTM. Province and Rouder (2012) tested a critical invariance property of state–response mapping functions that are required by 2HTM to account for confidence-rating data as discussed above: State–response mappings are not a function of the probability of the different discrete memory states being reached. This property leads to a signature prediction for memory-strength manipulations. Memory-strength manipulations should *only* affect the detection of studied items, leaving the state–response mapping functions unaffected. In terms of the distribution of responses across the confidence-rating scale, this means that the component distributions (defined by the state–response mappings) should remain invariant under memory-strength manipulations, with only the mixture weights (defined by the probabilities of entering the different detection states) being affected (see also Falmagne, 1985, p. 255). In contrast, continuous models like UVSD assume that memory-strength manipulations should lead to shifts in the response distributions. The results reported by Province and Rouder are consistent with 2HTM's predictions, but did not support continuous models. Moreover, recent work by Bröder, Kellen, Schütz, and Rohrmeier (in press) focusing on the modeling of confidence-rating ROCs with the 2HTM provided an experimental validation of state–response mapping functions, showing that the latter can be selectively manipulated without affecting the detection parameters. The results of Province and Rouder and of Bröder et al. (in press), together with the ones reported here, suggest not only that 2HTM provides a parsimonious account of the data, but also that this account can be corroborated by focused validation tests.

The usefulness of measurement models stems from the fact that they allow us to go beyond the raw data and make theoretically significant statements about the underlying cognitive processes. The search for the most appropriate measurement model in recognition memory has a long history but has been riddled with limitations in terms of the methods used and the scope of the comparisons made. The use of MDL-based measures such as NML overcomes some of those limitations, as is shown here and in other work (e.g., Myung et al., 2006). Ongoing research in MDL is quickly making these measures available for an increasing number of models (Wu et al., 2010a, b). For example, current research is addressing new methods for computing NML indices for categorical data with more than two response categories, which comprises confidence-rating ROC data (e.g., Kontkanen & Myllymäki, 2007). The use of MDL-based measures is not limited to ROC data and can potentially be applied to any experimental paradigm.

An important caveat should be highlighted, though: It is important to emphasize that model selection in general is *not* to be reduced to an automated comparison of indices that take into account model fit and model flexibility. Other important criteria for evaluating models include (but are not limited to) (1) the explanatory adequacy of the accounts, (2) the validity of parameter interpretations, (3) the testability of models, and (4) the heuristic value of the models in generating predictions for new contexts. Different weights can be given to these criteria depending on the particular goals of the researcher (Cohen et al., 2008). Model selection indices such as NML are useful statistical tools that contribute to scientific development and should not be seen as the sole arbiters of truth or adequacy, regardless of their sophistication.

Author Note We thank Chad Dube and Jeff Starns for providing their original data sets. The research reported in this article was supported by grant Kl 614/32-1 from the Deutsche Forschungsgemeinschaft to Karl Christoph Klauer. Model-fitting routines used in this article can be obtained upon request

References

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

Atkinson, R. C. (1963). A variable sensitivity theory of signal detection. *Psychological Review, 70,* 91–106.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology. Human Perception and Performance, 25,* 1189–1206.

Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation, 9,* 349–368.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99.

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 & 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.

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

Box, G. E. P. (1979). Robustness in scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York: Academic Press.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44,* 62–91.

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (in press). Validating a two-high threshold model for confidence rating data in recognition memory. *Memory.*

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.

Bröder, A., & Schütz, J. (2011). Correction to Bröder and Schütz (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1301.

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

Chechile, R. A. (1998). A new method for estimating model parameters for multinomial data. *Journal of Mathematical Psychology, 42,* 432–471.

Chechile, R. A. (2004). New multinomial models for the Chechile-Meyer task. *Journal of Mathematical Psychology, 48,* 364–384.

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

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley Interscience.

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities standard, distractor-free, and target-free recognition. *Memory & Cognition, 39,* 925–940.

Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 2–17.

Davis-Stober, C. P., & Brown, N. (2011). A shift in strategy or "error"? strategy classification over multiple stochastic specifications. *Judgment and Decision Making, 6,* 800–813.

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

DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 18–33.

DeCarlo, L. T. (2008). Process dissociation and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1565–1572.

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

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.

Dube, C., Rotello, C. M., & Heit, E. (2011). The belief bias effect is aptly named: A reply to Klauer and Kellen (2011). *Psychological Review, 118,* 155–163.

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

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. *Zeitschrift für Psychologie / Journal of Psychology, 217,* 108–144.

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

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

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

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

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

Grünwald, P., & Navarro, D. J. (2009). NML, Bayes and true distributions: A comment on Karabatsos and Walker (2006). *Journal of Mathematical Psychology, 53,* 43–51.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.

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

Henriques, J. B., Glowacki, J. M., & Davidson, R. J. (1994). Reward fails to alter response bias in depression. *Journal of Abnormal Psychology, 103,* 460–466.

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

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

Jeffreys, H. (1961). *The theory of probability.* Oxford: Oxford University Press.

Karabatsos, G., & Walker, S. G. (2006). On the normalized maximum likelihood and Bayesian decision theory. *Journal of Mathematical Psychology, 50,* 517–520.

Kass, R., & Rafferty, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

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). One the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review, 119,* 457–479.

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101,* 166–171.

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

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

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 exibility 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.

Kontkanen, P., & Myllymäki, P. (2007). A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters, 103,* 227–233.

Lee, M. D. (2004). An efficient method for the minimum description length evaluation of cognitive models. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 807–812). Mahwah: Erlbaum.

Lee, M. D., & Navarro, D. J. (2005). Minimum description length and psychological clustering models. In P. Grünwald, J. I. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 355–384). Cambridge: MIT Press.

Lee, M. D., & Pope, K. J. (2006). Model selection for the rate problem: A comparison of significance testing, Bayesian, and minimum description length statistical inference. *Journal of Mathematical Psychology, 50,* 193–202.

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

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A users guide* (2nd ed.). Mahwah: Erlbaum.

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

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

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.

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

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

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000a). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences of the United States of America, 97,* 11170–11175.

Myung, J. I., Forster, M., & Brown, M. W. (2000b). A special issue on model selection. *Journal of Mathematical Psychology, 44,* 1–2.

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

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

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116,* 499–518.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review, 14,* 1043–1050.

Navarro, D. J. (2004). A note on the applied use of MDL approximations. *Neural Computation, 16,* 1763–1768.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49,* 47–84.

O'Connor, A. R., Guhl, E. N., Cox, J. C., & Dobbins, I. G. (2011). Some memories are odder than others: Judgments of episodic oddity violate known decision rules. *Journal of Memory and Language, 64,* 299–315.

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

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review, 114,* 188–201.

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

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review, 73,* 44–58.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6,* 421–425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41,* 227–259.

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

Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling the accumulation of partial information. *Psychological Review, 95,* 238–255.

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

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

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

Riefer, D. M., & Batchelder, W. H. (1991). Statistical inference for multinomial processing tree models. In J. P. Doignon & J. C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 313–335). New York: Springer.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment, 14,* 184–201.

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

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory, 29,* 629–636.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42,* 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47,* 1712–1717.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). New York: Springer.

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

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

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

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

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 5976–5979.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

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.

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

Sherman, S. J., Atri, A., Hasselmo, M. E., Stern, C. E., & Howard, M. W. (2003). Scopolamine impairs human recognition memory: Data and modeling. *Behavioral Neuroscience, 117,* 526–539.

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

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64,* 1–34.

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

Su, Y., Myung, J. I., & Pitt, M. A. (2005). Minimum description length and cognitive modeling. In P. Grünwald, J. I. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 411–433). Cambridge: MIT Press.

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.

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

Wagenmakers, E. J., & Waldorf, L. (2006). Editors' introduction. *Journal of Mathematical Psychology, 50,* 99–100.

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

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

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054.

Wu, H., Myung, J. I., & Batchelder, W. H. (2010a). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review, 17,* 275–286.

Wu, H., Myung, J. I., & Batchelder, W. H. (2010b). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology, 54,* 291–303.

Yonelinas, A. P., & Jacoby, L. J. (2012). The process-dissociation approach two decades later: Convergence, boundary conditions, and new directions. *Memory and Cognition, 40,* 663–680.

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

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology. General, 130,* 361–379.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience, 25,* 3002–3008.

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