Journal of Mathematical Psychology 67 (2015) 8–25

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

The flexibility of models of recognition memory: The case of confidence ratings

Karl Christoph Klauer[a,*,1], David Kellen[b,1]

[a] Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany
[b] Center for Cognitive and Decision Sciences, Faculty of Psychology, University of Basel, Switzerland

HIGHLIGHTS

- A method for computing NML for models with categorical data is presented.
- NML penalties are given for confidence-rating based models of recognition memory.
- In simulation studies, NML outperforms AIC and BIC in model recovery.
- A meta-analysis based on NML supports dual-process signal-detection models.

ARTICLE INFO

Article history:
Received 17 December 2014
Received in revised form
7 May 2015

Keywords:
Model selection
Minimum-description length
Normalized maximum likelihood
Recognition memory
Confidence ratings

ABSTRACT

The normalized maximum likelihood (NML) index is a model-selection index derived from the minimum-description length principle. In contrast to traditional model-selection indices, it also quantifies differences in flexibility between models related to their functional form. We present a new method for computing the NML index for models of categorical data that parameterize multinomial or product-multinomial distributions and apply it to comparing the flexibility of major models of recognition memory for confidence-rating based receiver-operating-characteristic (ROC) data. NML penalties are tabulated for datasets of typical sizes and interpolation functions are fitted that allow one to interpolate NML penalties for datasets with sizes between the tabulated ones. Recovery studies suggest that the NML index performs better than traditional model-selection indices in model selection from ROC data. In an NML-based meta-analysis of 850 ROC datasets, versions of the dual-process signal detection models received most support followed by the finite mixture signal detection model and constrained versions of two-high threshold models.

Recognition memory has frequently been studied by means of mathematical models (for a review, see Malmberg, 2008 and Yonelinas & Parks, 2007). A range of models has been proposed. In some models, information from memory is represented in terms of discrete states; in others, a continuous representation of evidence is postulated. Discrete-state models are variants of the so-called threshold models (e.g., Blackwell, 1963 and Snodgrass & Corwin, 1988); the continuous models are variants of the so-called signal-detection models (Macmillan & Creelman, 2005). Finally, hybrid models implement combinations of both ideas. Model fits and comparisons are frequently based on the shape of the observed receiver operating characteristic (ROC) functions.

1. Receiver operating characteristics

In the most basic recognition experiment, participants study items to be remembered later. In a subsequent test phase, they are shown the studied items mixed with new items, so-called distractors, and their task is to discriminate studied items from new items. These data are typically modeled in terms of two probabilities, the probability to respond OLD given an old item and the probability to respond OLD given a new item, also called the hit and false-alarm rates, respectively. An important line of research is to obtain hit and false-alarm rates at different levels of response bias and to plot hit rates against false alarm rates across levels of response bias resulting in a so-called ROC function. Fig. 1 shows examples of typical ROCs: From top to bottom, they are typical of an ROC generated by a threshold model, a signal-detection model with higher variability of the memory response for old items than for new items, and a simpler signal detection model with equal

[*] Corresponding author.
E-mail address: christoph.klauer@psychologie.uni-freiburg.de (K.C. Klauer).
[1] Both authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.jmp.2015.05.002
0022-2496/© 2015 Elsevier Inc. All rights reserved.

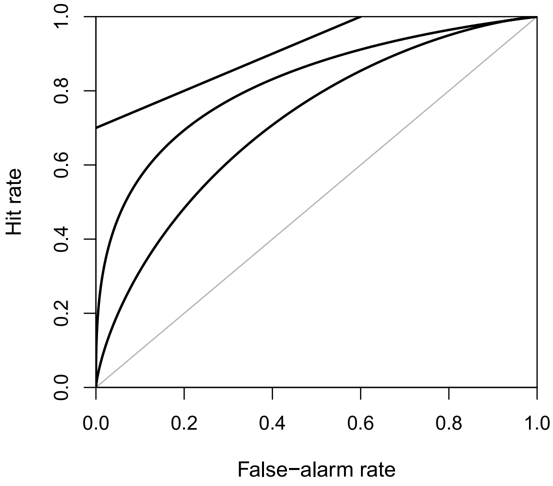

Fig. 1. Example ROC functions.

variability for old and new items. All of these fall above the chance-level diagonal that is also shown.

Different levels of response bias are traditionally produced via manipulations of base rates of old relative to new items in the test phase or by payoff manipulations (Bröder & Schütz, 2009). The vast majority of studies in the field has, however, relied on a less expensive method of generating ROC data via confidence ratings (see Wixted, 2007 and Yonelinas & Parks, 2007, for reviews). That is, ratings of confidence in the OLD or NEW response, as the case may be, are collected and the different levels of confidence interpreted as expressing different levels of response bias. Although most widely used in research on recognition memory, confidence-rating data and (some of) the above models also play an important role in perception (e.g., Swets, Tanner, & Birdsall, 1961) and reasoning (e.g., Dube, Rotello, & Heit, 2010).

Despite decades of research, the question which of the above models provide the best description of the data in recognition memory is still under debate (e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012; Kellen & Klauer, 2014; Kellen, Klauer, & Bröder, 2013; Kellen, Singmann, Vogt, & Klauer, 2015; Onyper, Zhang, & Howard, 2010; Province & Rouder, 2012; Wixted, 2007 and Yonelinas & Parks, 2007). To our minds, several factors have prevented this very productive field from reaching a clear and non-contested decision on the most adequate model. Among these are the distorting, but often ignored influences of individual differences in memory performance and response-bias settings and of analogous differences between items (Klauer & Kellen, 2010; Rouder & Lu, 2005), an over-reliance on one method, the confidence-rating paradigm, and the absence of model-selection measures that take into account differences between models in flexibility related to functional form. The purpose of the present manuscript is to address this last problem for the important case of confidence-rating data in a similar fashion as described by Kellen et al. (2013) and Klauer and Kellen (2011a) for binary OLD/NEW ROC data (see also Kellen & Klauer, 2011). It turns out that some of the solutions developed here are even more widely applicable than defined by this original purpose as elaborated on below.

2. Model selection

Given data and a range of models, the task to select the model that most parsimoniously accounts for the data is discussed under the heading "model selection". There is a growing awareness in psychology that goodness of fit and model flexibility should both be weighed when evaluating mathematical models. For example, two special issues of the Journal of Mathematical Psychology (Myung, Forster, & Brown, 2000; Wagenmakers & Waldorf, 2006) were recently devoted to this topic.

In recognition memory, model selection has usually relied on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC; see, e.g., Burnham & Anderson, 2005). Let f be the model's probability function, $\boldsymbol{x}$ the observed data, and $\hat{\theta}(\boldsymbol{x})$ the maximum-likelihood estimate of the p model parameters, $\theta = (\theta_1, \ldots, \theta_p)$. AIC and BIC are given by

$$\mathrm{AIC} = -2 \log f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x})) + 2p, \quad \text{and}$$

$$\mathrm{BIC} = -2 \log f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x})) + p * \log(N),$$

where N is the number of data points. In both indices, the first term quantifies the model's goodness of fit (minus twice the maximum log-likelihood), and the second term is the flexibility penalty. AIC and BIC thus gauge model flexibility basically in terms of the number of parameters. This is both crude and not very helpful in the present context. It is crude because an extra parameter can have anything between very little effect on the model's flexibility (i.e., on its capability to fit diverse datasets) and a tremendous effect depending on the functional form via which it enters the model equations. It is also not very helpful given that many of the major models employ the same number of parameters in fitting ROC data (plus or minus one or two). This problem has repeatedly been noted in the literature, without definitive solution so far (Macho, 2004; Onyper et al., 2010 and Wixted, 2007; see also Cohen, Rotello, & Macmillan, 2008).

The purpose of the present project is to bring functional form into consideration using recent developments in the model-selection field based on the minimum-description-length (MDL) principle (Myung, Navarro, & Pitt, 2006). Roughly, a model reduces the complexity of a code (e.g., a string of binary values) needed to describe the data, because only the model, the estimated parameter values as well as the residuals have to be encoded, once the model has been fit. Inasmuch as the residuals show less variability than the original data, a good model thereby reduces the code needed to describe the dataset. The code length is a function of both the model's complexity and its ability to account for the data (Grünwald, 2007).

Model selection based on the minimum-description-length principle has a strong track record in psychology. These modern methods have to date been applied to the class of multinomial processing tree models (Wu, Myung, & Batchelder, 2010a,b), to models of human categorization (Myung, Pitt, & Navarro, 2007), to clustering models (Navarro & Lee, 2005), to models of recognition memory (Kellen & Klauer, 2011; Kellen et al., 2013; Klauer & Kellen, 2011a), to decision making (Davis-Stober & Brown, 2011; Moshagen & Hilbig, 2014), and to structural equation models (Preacher, 2006), among others.

In the present context, the principle leads to the normalized maximum likelihood index (NML) for model selection. Much like AIC and BIC, the index adds minus the maximum log-likelihood of the data given the model and a penalty term for the model's flexibility. The penalty is given by the logarithm of the sum of maximum likelihood values summed over the entire set of possible data patterns $\boldsymbol{y}$ that might in principle occur in the experimental setting. NML[2] is given by

$$\mathrm{NML} = -\log f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x})) + \log \sum_{\boldsymbol{y}} f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y})).$$

[2] Strictly speaking, this expression gives the logarithm of NML. NML cannot be computed for the case of continuous data as the penalty term is not finite for continuous data (e.g., Karabatsos & Walker, 2006). This problem can be sidestepped through the introduction of a (informative) prior distribution for the data in the computation of the penalty (see Zhang, 2011).

The penalty term, $\log \sum_{\boldsymbol{y}} f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y}))$, is a measure of the model's ability to fit data in general, whatever the data. The resulting model-selection index is principled and has a couple of desirable properties (Grünwald, 2007, Chapter 7; Myung, Balasubramanian, & Pitt, 2000 and Rissanen, 1996, 2001) such as, in the present context, consistency: Roughly, if one of the models is the correct one, use of NML will select it as sample size increases. Importantly, NML behaves as one would intuitively expect of an index that takes functional form into account (see e.g., Kellen & Klauer, 2011; Klauer & Kellen, 2011a; Su, Myung, & Pitt, 2005 and Wu et al., 2010a,b). In simulation studies, use of MDL-based model-selection indices such as NML led to more valid results than the use of only goodness-of-fit values, or AIC (Klauer & Kellen, 2011a; Myung et al., 2007), or BIC (Klauer & Kellen, 2011a; Su et al., 2005).

NML has a simple interpretation: The model's fit to the data is put in relation to the model's ability to fit data in general. This addresses in a head-on fashion concerns raised by Roberts and Pashler (2000) and others (e.g., Chechile, 1998 and Myung, 2000) that a given level of model fit should be related to the a priori probability that the model provides a given level of fit.

Although more sophisticated than AIC and BIC, a more widespread use of NML has been hampered by the difficulty of computing the index. Computing NML involves processing all data patterns that can in principle occur in an experiment, which implies prohibitively many computations even for modern computers. Klauer and Kellen (2011a) developed a new method to compute NML as well as an asymptotic approximation thereof, the Fisher information approximation (FIA), for data from base rate and payoff experiments with binary (OLD/NEW) responses for the major recognition models.

3. Overview

A first question pursued here is whether similar methods can be developed for the case of confidence-rating data. It turns out that it is possible to use a similar approach to compute the NML indices for confidence-rating data. In fact, the method that we propose is sufficiently general to be applicable for the purpose of computing NML for any model based on categorical data parameterizing multinomial or product-multinomial distributions (e.g., Kellen, Singmann, & Klauer, 2014). The new method should thereby be helpful for the entire huge field of categorical data analysis (e.g., Agresti, 1990 and Bishop, Fienberg, & Holland, 1975) beyond the narrower fields of recognition memory, perception, and reasoning on which we focus here.

Having described the new method, we go on to apply it to the computation of NML indices for the major models in use in recognition memory in order to (a) assess and compare their flexibility and (b) provide researchers with a range of precomputed and tabulated indices as well as interpolation functions covering the most frequent sizes of datasets encountered in the literature.

In a recovery study, we then compare the performance of the NML, AIC, and BIC indices in selecting the underlying models from sets of candidate models. Finally, we apply the new method in a meta-analytic study to a large set of existing confidence-rating datasets. We begin by a brief description of the range of models considered here.

4. Models

Confidence ratings in recognition memory are typically given on (or recoded to) a $2k$-point rating scale ranging from 1 (highest-confidence NEW judgment) to $2k$ (highest-confidence OLD judgment) allowing one to discriminate between k levels of confidence in the OLD or NEW judgment as the case may be. The data are then modeled in terms of the probabilities of producing response category r, $P(R = r \mid \text{old})$ and $P(R = r \mid \text{new})$, given an old item and a new item, respectively, for $r = 1, \ldots, 2k$. Most often, the number of response categories r is even, but depending upon the precise response format, odd numbers can occur. Let M be the number of response categories minus one.

In terms of models, the major players are the One-High Threshold Model (1HTM), the Two-High Threshold Model (2HTM), the Equal-Variance Signal Detection Model (EVSDT), the Unequal-Variance Signal Detection Model (UVSDT), the Dual-Process Signal Detection Model (DPSDT), the Finite Mixture Signal Detection Model (MSDT), and the Variable-Recollection Dual-Process model (VRDP). Let us consider these in turn, beginning with the continuous models. Fig. 2 gives a graphical depiction of core assumptions of the discrete, continuous, and hybrid models.

In the signal detection models (Macmillan & Creelman, 2005), it is assumed that each item evokes a value on a familiarity dimension. Furthermore, decision makers are assumed to place M criteria, $c_1 < c_2 < \cdots < c_M$, on that dimension partitioning it into $M + 1$ intervals that are mapped onto the $M + 1$ response categories. Old and new items generate distributions on the familiarity dimension assumed to be normal with means separated by μ. In the EVSDT, both distributions are assumed to have equal variances, which, without loss of generality, can be set equal to one, whereas the mean of the distribution of new items can be set to zero. The model equations are:

$$P(R = r \mid \text{old}) = F(c_r - \mu) - F(c_{r-1} - \mu),$$
$$P(R = r \mid \text{new}) = F(c_r) - F(c_{r-1}),$$

where F is the cumulative distribution function of the standard normal distribution, $r = 1, \ldots, M + 1$, $c_0 = -\infty$, $c_{M+1} = \infty$, and $F(-\infty) = 0, F(\infty) = 1$.

In the UVSDT, the distribution of old items is assumed to have a variance σ^2 that may differ from the variance of the distribution of new items. Hence,

$$P(R = r \mid \text{old}) = F\left(\frac{c_r - \mu}{\sigma}\right) - F\left(\frac{c_{r-1} - \mu}{\sigma}\right),$$
$$P(R = r \mid \text{new}) = F(c_r) - F(c_{r-1}),$$

for $r = 1, \ldots, M + 1$.

In the MSDT (DeCarlo, 2002), the distribution of old items is assumed to be a mixture of two equal-variance normal distributions—one corresponding to items that were attended to during study with mean μ, the other one to unattended items with mean μ^*, with $\mu \geq \mu^*$ and with μ^* often set equal to zero. The proportion of attended items is described by the mixture coefficient λ. Hence,

$$\begin{aligned} P(R = r \mid \text{old}) &= \lambda(F(c_r - \mu) - F(c_{r-1} - \mu)) \\ &\quad + (1 - \lambda)(F(c_r - \mu^*) - F(c_{r-1} - \mu^*)) \end{aligned}$$
$$P(R = r \mid \text{new}) = F(c_r) - F(c_{r-1}),$$

for $r = 1, \ldots, M + 1$. The MSDT is equivalent to the VRDP model by Onyper et al. (2010) in the present context. We will refer to the submodel with $\mu^* = 0$ as MSDT$_0$.

The DPSDT (Yonelinas, 1997) combines high-threshold and signal-detection assumptions. It is assumed that a certain proportion R of old items is recollected, leading to a highest-confidence OLD response. If recollection fails, responses are governed by the EVSDT. Hence,

$$P(R = M + 1 \mid \text{old}) = R + (1 - R)(F(c_{M+1} - \mu) - F(c_M - \mu)),$$
$$\begin{aligned} P(R = r \mid \text{old}) &= (1 - R)(F(c_r - \mu) - F(c_{r-1} - \mu)) \\ &\quad \text{for } r < M + 1, \end{aligned}$$
$$P(R = r \mid \text{new}) = F(c_r) - F(c_{r-1}) \quad \text{for } r = 1, \ldots, M + 1.$$

In the 1HTM (Blackwell, 1963), it is assumed that memory can be described by two discrete states, termed (1) "detection"

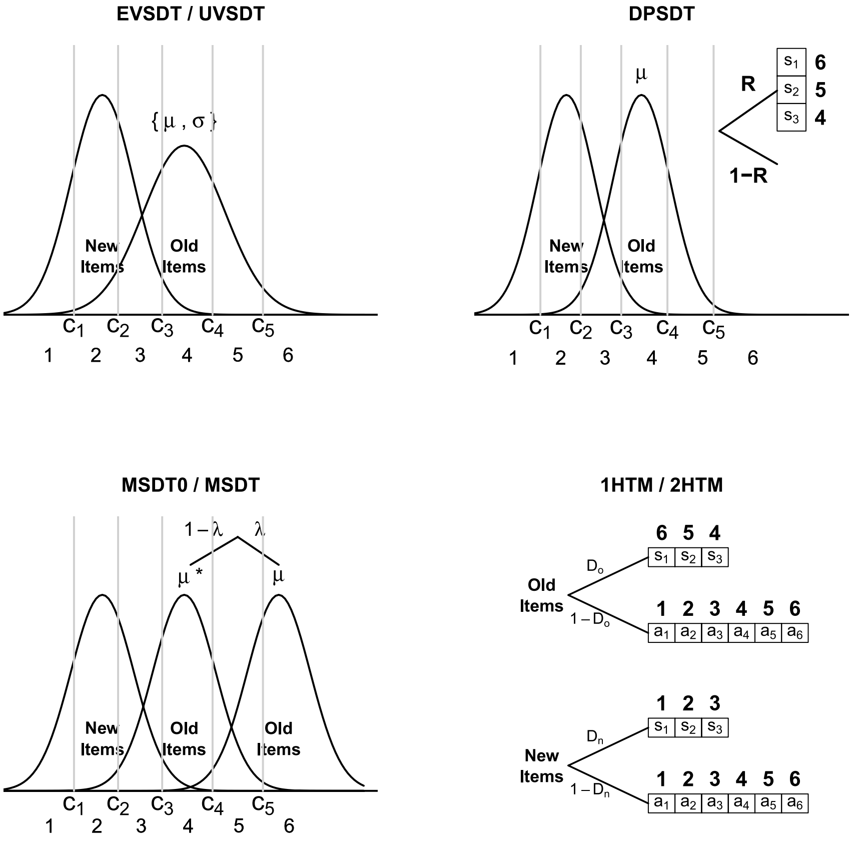

Fig. 2. Graphical depictions of major models of recognition memory for confidence-rating data.

or "remember" and (2) "guess". Given an old item, the item can either be remembered with probability D_o, leading to a highest-confidence OLD response, or not, in which case a response category r is guessed with probability $a_r,\ r = 1, \ldots, M+1$. Responses for new items are based only on guessing. This leads to the following model equations:

$$P(R = M+1 \mid \text{old}) = D_o + (1 - D_o)a_{M+1},$$
$$P(R = r \mid \text{old}) = (1 - D_o)a_r \quad \text{for } r < M+1,$$
$$P(R = r \mid \text{new}) = a_r \quad \text{for } r = 1, \ldots, M+1.$$

Note that one of the guessing parameters a_r is redundant, because the a parameters have to sum to one.

In the 2HTM (Snodgrass & Corwin, 1988), there is an additional state of distractor detection that is reached with probability D_n given a new item and leads to the highest-confidence NEW response. Hence, the model equations are:

$$P(R = M+1 \mid \text{old}) = D_o + (1 - D_o)a_{M+1},$$
$$P(R = r \mid \text{old}) = (1 - D_o)a_r \quad \text{for } r < M+1,$$
$$P(R = r \mid \text{new}) = (1 - D_n)a_r \quad \text{for } r > 1,$$
$$P(R = 1 \mid \text{new}) = D_n + (1 - D_n)a_1.$$

A model with $D_o = D_n$ is also often considered (e.g., Bröder & Schütz, 2009) which we will refer to as 2HTM$_{D_o=D_n}$.

1HTM, 2HTM$_{D_o=D_n}$, and EVSDT employ $M + 1$ parameters, 2HTM, UVSDT, DPSDT, and MSDT$_0$ $M + 2$ parameters, and MSDT $M+3$ parameters, whereas the confidence-rating data provide $2M$ non-redundant observed response frequencies.

4.1. Order restrictions and response mappings

One of the advantages of the minimum-description-length approach to model selection is that it is able to capture reductions in model flexibility produced by a priori order restrictions on parameters, despite the fact that these restrictions do not reduce the number of parameters.[3] In the case of the continuous and hybrid models (EVSDT, UVSDT, DPSDT, MSDT$_0$, MSDT), we constrain parameter μ to be larger or at best equal to zero a priori; for MSDT we impose $\mu \geq \mu^* \geq 0$. The constraints imply that performance in discriminating between old and new items should not be reliably lower than chance (with the exception of UVSDT that can accommodate below-chance performance even under the restriction $\mu \geq 0$). These constraints are accepted by most researchers working in the field on theoretical grounds, and they lead to a fairer comparison with the discrete-state models that rule out below-chance performance via the functional form of their model equations without additional parametric constraints.

Furthermore, a common finding in recognition memory and perception is that of asymmetric ROCs (e.g. Glanzer, Kim, Adams, & Hilford, 1999; Ratcliff, Sheu, & Gronlund, 1992 and Swets et al., 1961). The asymmetry can be accommodated by constraining UVSDT parameter σ to be larger or at best equal to one. Like Klauer and Kellen (2011a), we therefore consider two variants of the UVSDT model, one without constraint on σ and one with $\sigma \geq 1$ to which we will refer as UVSDT$_{\sigma\geq 1}$. Imposing the constraint implies a benefit in computing the NML penalty, and leads to a fairer comparison with models such as DPSDT and MSDT that accommodate the asymmetry via the functional form of the model equations without additional parametric constraint (for $R > 0$ in

[3] We already noted in the definition of the continuous and hybrid models that the response criteria c_i are ordered so that $c_1 \leq c_2 \leq \cdots \leq c_M$. This constraint ensures that the probabilities defined by the model equations are nonnegative. It is not a constraint on the flexibility of the model in the sense that relaxing the constraint would lead to a model that can accommodate additional datasets. Instead relaxing the constraint would lead to a ill-defined model predicting negative probabilities.

DPSDT and $\lambda < 1$ in MSDT). Considering the 2HTM, the asymmetry is accommodated by the constraint $D_o \geq D_n$ (Klauer & Kellen, 2011a) and like for the UVSDT, we consider two versions of the 2HTM, one without and one with the constraint, the latter termed $2\text{HTM}_{D_o \geq D_n}$.

The 1HTM and 2HTM often perform poorly in fitting confidence-rating based ROC data, whereas the other models perform better. The discrete-state models as stated above predict linear ROCs, whereas ROCs obtained with confidence-rating data are frequently curved (Yonelinas & Parks, 2007). However, as already pointed out by Erdfelder and Buchner (1998) and Malmberg (2002; see also Bröder, Kellen, Schütz, & Rohrmeier, 2013; Bröder & Schütz, 2009; Klauer & Kellen, 2010, 2011b; Krantz, 1969; Luce, 1963 and Schütz & Bröder, 2011), the threshold models' predictions of linear ROCs depend on assumptions about how detect states are mapped on rating categories. The prediction of linear ROCs is predicated on the assumption that in a "detect" state, the highest confidence level is *invariably* chosen on the correct side of the rating scale (i.e., on the OLD side in a "detect old" or "recollection" state, and on the NEW side in a "detect new" or distractor-detection state). But well-documented individual differences in response styles regarding the use of extreme response categories and response strategies (Böckenholt, 2012), intra-individual variations and sequential dependencies in scale usage, and simply, the possibility of random errors (Rieskamp, 2008) all suggest that a certain proportion of responses generated from detect/recollection states might be mapped on less than highest confidence ratings. As soon as this possibility is admitted, models with detection/recollection states can predict curved ROCs similar to those predicted by continuous models.

For these reasons, we will also consider versions of the discrete-state and hybrid models with (non-deterministic) response mapping. For example, the 1HTM, the 2HTM, and the DPSDT have a detect state for old items that is reached with probability D_o (1HTM, 2HTM) or R (DPSDT). In the version with response mapping, it is assumed that aside from highest-confidence OLD responses, lower-confidence OLD responses also occur with certain probabilities s. Let k_{OLD} be the lowest-confidence OLD response. Then, in the detect state for old items, response r with $r \geq k_{\text{OLD}}$ is chosen with probability s_{M+2-r}, where $s_1, s_2, \ldots, s_{M+2-k_{\text{OLD}}}$ are parameters to be estimated from the data with the constraint that they have to sum to one. Thus, s_1 is the probability that the highest-confidence OLD response is chosen in the OLD detect state, s_2 the probability of choosing the next lower confidence OLD response, and so forth. For example, the DPSDT with response mapping is defined by the following equations:

$$\begin{aligned} P(R = r \mid \text{old}) &= R \times s_{M+2-r} + (1-R)(F(c_{M+1} - \mu) \\ &\quad - F(c_M - \mu)) \quad \text{for } r \geq k_{\text{OLD}} \\ P(R = r \mid \text{old}) &= (1-R)(F(c_r - \mu) - F(c_{r-1} - \mu)) \\ &\quad \text{for } r < k_{\text{OLD}}, \end{aligned}$$

$$P(R = r \mid \text{new}) = F(c_r) - F(c_{r-1}) \quad \text{for } r = 1, \ldots, M+1.$$

The 2HTM also comprises a detect state for new items, for which an analogous response mapping can be defined with new parameters. In many situations, it makes sense to assume that the same s parameters describe the response mapping of OLD and NEW detect states on confidence levels (so that $s_{M+1} = s_1$, $s_M = s_2$, and so forth), and we employ this simplifying assumption for the present analyses (but see Kellen et al., 2015 and Klauer & Kellen, 2010). Results for models from the 2HTM family with different response mappings for new and old items can be found in the online supplement (see Appendix B). Discrete-state models with response mappings have been successfully fitted by Erdfelder and Buchner (1998), Klauer and Kellen (2010, 2011b), and Schütz and Bröder (2011), among others.

In what follows, let us refer to the 1HTM, $2\text{HTM}_{D_o=D_n}$, 2HTM, and DPSDT without (non-deterministic) response mapping as 1HTM_{wo}, $2\text{HTM}_{D_o=D_n,\text{wo}}2\text{HTM}_{\text{wo}}$, and DPSDT_{wo} (wo for "without"), and let us use 1HTM, $2\text{HTM}_{D_o=D_n}$, 2HTM, and DPSPDT for the models with response mapping. For the models with response mapping, it is possible, for example, that lowest confidence ratings would be strongly preferred even in detect/recollection states (i.e., these confidence levels can have the largest s parameters). This has prompted criticisms to the effect that the models with response mapping deal with rating scales in an arbitrary and post-hoc manner (e.g., Dubé, Rotello, & Pazzaglia, 2013, Pazzaglia, Dubé, & Rotello, 2013; see also Batchelder & Alexander, 2013). According to the critics, these model variants are overly complex (Dube, Rotello, & Heit, 2011).

The problem critiqued is somewhat mitigated if models are evaluated using NML which penalizes the models for increases in flexibility. Nevertheless, as we have argued elsewhere (Klauer, Singmann, & Kellen, 2015), it is possible to define psychologically more meaningful response mappings by imposing restrictions on the mapping parameters s. Specifically, for model variants with subscript "r", 1HTM_{r}, $2\text{HTM}_{D_o=D_n,\text{r}}$, 2HTM_{r}, and DPSDT_{r}, we impose the restriction that parameters s have to increase monotonically as confidence level increases. Even more restrictively, for model variants with subscript "r2", 1HTM_{r2}, $2\text{HTM}_{D_o=D_n,\text{r2}}$, 2HTM_{r2}, and DPSDT_{r2}, this same order restriction is in force and in addition, only the s parameters for the two highest confidence levels are allowed to differ from zero. In other words, under these models detect/recollection states are mapped onto the correct response with either highest confidence level, or with next-to-highest confidence (with smaller probability). Note that the different response mappings define a nested series of models for each model with discrete-state elements. For example, DPSDT_{wo} is a submodel of DPSDT_{r2}, which is a submodel of DPSDT_{r}, which is a submodel of DPSDT. Remember also that NML is particularly well suited to quantify the increase in flexibility within these series of nested models even where number of parameters and hence AIC and BIC do not distinguish between the nested models (e.g., between DPSDT_{r} and DPSDT). Table 1 provides an overview of the resulting set of 25 models considered.

5. Computing NML for categorical data

In computing NML, the challenging part is to compute the sum over all data patterns $\boldsymbol{y}$ of the maximum likelihood under a given model, $\sum_{\boldsymbol{y}} f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y}))$. One idea in Klauer and Kellen (2011a) was to state that sum as the integral over $f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y}))$ with respect to the counting measure ν that assigns measure 1 to each data pattern $\boldsymbol{y}$ that can occur:

$$T = \sum_{\boldsymbol{y}} f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y})) = \int f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y})) \mathrm{d}\nu(\boldsymbol{y}).$$

This makes it possible to apply methods of Monte Carlo integration to approximate the sum via the independent importance sampling algorithm (Evans & Swartz, 2000, Chapter 6; see also Roos, 2008, for an independent application of Monte Carlo integration to the computation of NML). For this purpose, data patterns $\boldsymbol{y}_i$, $i = 1, \ldots, m$, are sampled from a density g defined on the data space with $g > 0$ for all data patterns that can occur, and the integral $T = \int f$ is approximated by $T_m = \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{y}_i \mid \hat{\theta}(\boldsymbol{y}_i))/g(\boldsymbol{y}_i)$.

One condition that is sufficient for T_m to converge to the integral in question as m becomes large is that g dominates f, that is that there is a value $M > 0$ such that $f(\boldsymbol{y}) \leq Mg(\boldsymbol{y})$ for all $\boldsymbol{y}$ in the data space. In addition, the rate of convergence will largely depend upon the similarity of f and g. Ideally, $\frac{f}{g}$ should be a constant.

Table 1
The set of recognition memory models considered.

Acronym	Description	p
1HTM_{wo}	One-high threshold model without response mapping	$M+1$
1HTM_{r2}	· · · with order-constrained response mapping restricted to the two highest confidence levels	$M+2$
1HTM_{r}	· · · with order-constrained response mapping	$M+(M+1)/2$
1HTM	· · · with unconstrained response mapping	$M+(M+1)/2$
$2\text{HTM}_{D_o=D_n,\text{wo}}$	Two-high threshold model with $D_o = D_n$ and without response mapping	$M+1$
$2\text{HTM}_{D_o=D_n,\text{r2}}$	· · · with order-constrained response mapping restricted to the two highest confidence levels	$M+2$
$2\text{HTM}_{D_o=D_n,\text{r}}$	· · · with order-constrained response mapping	$M+(M+1)/2$
$2\text{HTM}_{D_o=D_n}$	· · · with unconstrained response mapping	$M+(M+1)/2$
EVSDT	Equal-variance-signal detection model	$M+1$
$\text{UVSDT}_{\sigma\geq 1}$	Unequal-variance signal detection model with the constraint $\sigma \geq 1$	$M+2$
UVSDT	Unequal-variance signal detection model	$M+2$
$2\text{HTM}_{D_o\geq D_n,\text{wo}}$	Two-high threshold model with $D_o \geq D_n$ and without response mapping	$M+2$
$2\text{HTM}_{D_o\geq D_n,\text{r2}}$	· · · with order-constrained response mapping restricted to the two highest confidence levels	$M+3$
$2\text{HTM}_{D_o\geq D_n,\text{r}}$	· · · with order-constrained response mapping	$M+1+(M+1)/2$
$2\text{HTM}_{D_o\geq D_n}$	· · · with unconstrained response mapping	$M+1+(M+1)/2$
2HTM_{wo}	Two-high threshold model without response mapping	$M+2$
2HTM_{r2}	· · · with order-constrained response mapping restricted to the two highest confidence levels	$M+3$
2HTM_{r}	· · · with order-constrained response mapping	$M+1+(M+1)/2$
2HTM	· · · with unconstrained response mapping	$M+1+(M+1)/2$
DPSDT_{wo}	Dual-process signal detection model without response mapping	$M+2$
DPSDT_{r2}	· · · with order-constrained response mapping restricted to the two highest confidence levels	$M+3$
DPSDT_{r}	· · · with order-constrained response mapping	$M+1+(M+1)/2$
DPSDT	· · · with unconstrained response mapping	$M+1+(M+1)/2$
MSDT_0	Finite mixture signal detection model with $\mu^* = 0$	$M+2$
MSDT	Finite mixture signal detection model	$M+3$

Note. p = number of parameters; M = number of response categories minus one; "/" refers to integer division. For odd $M+1$ such as for seven-point scales, the scale has a neutral midpoint given by $(M+2)/2$.

A second idea in Klauer and Kellen (2011a) was to use a density g that is proportional to the maximum likelihood of the saturated model. The saturated model is a model that can fit each data pattern $\boldsymbol{y}$ perfectly and the maximum likelihood of which can be stated explicitly. Obviously, g dominates any model-based maximum-likelihood function $f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y}))$, because the maximum likelihood of the saturated model cannot be smaller than that of a restricted model. In addition, g is similar to the maximum-likelihood of recognition-memory models to the extent to which these are similar to the saturated model, that is impose few restrictions on the data. For binary OLD/NEW ROCs with base rate or payoff manipulation of response bias, it was possible to sample data patterns from the distribution defined by density g relatively efficiently. The resulting algorithm was sufficiently efficient to allow one to compute the NML indices for datasets of realistic sizes.

Computing a model's NML index in this way involves three steps:

1. Sampling datasets from a density proportional to the maximum likelihood of the saturated model;
2. fitting the model to each dataset; and
3. computing the average ratio of the maximum likelihood of the model and that of the saturated model across sampled datasets.

In Klauer and Kellen's (2011a) situation, sampling (Step 1) can be performed for each response-bias condition and for trials with old and new items separately due to an independence property of the saturated model for binary OLD/NEW data. The problem here is that a similar independence property of the saturated model across response-bias conditions does not exist for confidence-rating data. This entails that the sampling method used by Klauer and Kellen (2011a) for the independent importance sampling algorithm, namely rejection sampling (Gelman, Carlin, Stern, & Rubin, 2004, Chapter 11), cannot be employed with any efficiency.

The major new idea pursued here is to propose a highly efficient Gibbs sampler (Gelman et al., 2004, Chapter 11) for sampling from the distribution defined by the density g that is proportional to the maximum-likelihood function of the saturated model. We specify the Gibbs sampler for a single multinomial distribution; the case of product-multinomial distributions comprising several category systems (here, the two rating scales for old and new items) follows directly, because the distributions over separate category systems are stochastically independent under the distribution defined by the maximum-likelihood function of the saturated model.

Consider a set of frequency counts $\boldsymbol{y} = (y_1, \ldots, y_{M+1})$. The density g is proportional to

$$g \propto \binom{n}{y_1, \ldots, y_{M+1}} \prod_{i=1}^{M+1} \left(\frac{y_i}{n}\right)^{y_i},$$

where $n = \sum_{i=1}^{M+1} y_i$ is a given constant. Because of this, one of the y_i is redundant. Let us therefore consider only y_i for $i \leq M$, noting that $y_{M+1} = n - \sum_{j=1}^{M} y_j$.

For the Gibbs sampler, we need to specify, and be able to sample from, each of the conditional distributions of y_i given all other non-redundant frequency counts $\boldsymbol{y}^{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_M)$, for $i = 1, \ldots, M$. It turns out that

$$P(y_i \mid \boldsymbol{y}^{(i)}) \propto \frac{n!}{y_i! y_{M+1}!} \left(\frac{y_i}{n}\right)^{y_i} \left(\frac{y_{M+1}}{n}\right)^{y_{M+1}}.$$

Note further $y_{M+1} = n - y_i - \sum_{j\neq i, 1\leq j\leq M} y_j$ and let $m = n - \sum_{j\neq i, 1\leq j\leq M} y_j$, which is a constant given $\boldsymbol{y}^{(i)}$ (and n). This implies

$$\begin{aligned} P(y_i \mid \boldsymbol{y}^{(i)}) &\propto \frac{n!}{y_i!(m-y_i)!} \left(\frac{y_i}{n}\right)^{y_i} \left(\frac{m-y_i}{n}\right)^{m-y_i} \\ &= \frac{m!}{y_i!(m-y_i)!} \frac{n!}{m!} \left(\frac{y_i}{m}\right)^{y_i} \left(\frac{m-y_i}{m}\right)^{m-y_i} \left(\frac{m}{n}\right)^{m} \\ &\propto \binom{m}{y_i} \left(\frac{y_i}{m}\right)^{y_i} \left(\frac{m-y_i}{m}\right)^{m-y_i}, \end{aligned}$$

because the other terms in the products are given and can hence be absorbed in the proportionality constant. In words, the required conditional distribution function is proportional to the maximum-likelihood function of the saturated binomial model given a total of m trials. It is not difficult to sample from this distribution efficiently (Klauer & Kellen, 2011a).

Table 2
NML penalties (NML_P) for a six-point rating scale.

Model	p	Number of trials with old items (same as for new items)								
		30	50	100	300	500	1000	3000	5000	10 000
1HTM_{wo}	6	9.98	11.35	13.21	16.29	17.76	19.75	22.98	24.49	26.58
1HTM_{r2}	7	10.60	12.13	14.22	17.73	19.40	21.71	25.45	27.20	29.59
1HTM_{r}	8	11.04	12.68	14.96	18.83	20.70	23.29	27.51	29.50	32.18
1HTM	8	12.09	13.84	16.26	20.28	22.18	24.85	29.16	31.17	33.86
$2\text{HTM}_{D_o=D_n,\text{wo}}$	6	9.65	10.96	12.85	15.88	17.36	19.34	22.57	24.08	26.15
$2\text{HTM}_{D_o=D_n,\text{r2}}$	7	9.96	11.41	13.46	16.87	18.52	20.79	24.49	26.23	28.60
$2\text{HTM}_{D_o=D_n,\text{r}}$	8	10.13	11.67	13.85	17.58	19.36	21.86	25.96	27.93	30.62
$2\text{HTM}_{D_o=D_n}$	8	10.94	12.64	14.96	18.89	20.78	23.36	27.59	29.55	32.31
EVSDT	6	10.85	12.26	14.17	17.27	18.75	20.76	24.03	25.54	27.61
$\text{UVSDT}_{\sigma\geq 1}$	7	12.10	13.73	15.92	19.57	21.26	23.61	27.41	29.19	31.57
UVSDT	7	12.64	14.27	16.52	20.19	21.87	24.26	28.03	29.81	32.23
$2\text{HTM}_{D_o\geq D_n,\text{wo}}$	7	10.32	11.80	13.82	17.25	18.89	21.18	24.85	26.60	28.96
$2\text{HTM}_{D_o\geq D_n,\text{r2}}$	8	10.91	12.53	14.80	18.61	20.46	23.00	27.19	29.18	31.88
$2\text{HTM}_{D_o\geq D_n,\text{r}}$	9	11.27	13.00	15.42	19.52	21.54	24.36	28.95	31.14	34.10
$2\text{HTM}_{D_o\geq D_n}$	9	12.32	14.16	16.70	20.97	23.01	25.91	30.59	32.81	35.78
2HTM_{wo}	7	10.73	12.25	14.30	17.78	19.46	21.77	25.48	27.26	29.60
2HTM_{r2}	8	11.40	13.02	15.33	19.24	21.07	23.65	27.87	29.84	32.54
2HTM_{r}	9	11.77	13.52	16.02	20.15	22.16	24.98	29.59	31.80	34.78
2HTM	9	12.77	14.64	17.26	21.54	23.58	26.52	31.23	33.45	36.46
DPSDT_{wo}	7	11.08	12.59	14.61	18.02	19.60	21.86	25.52	27.27	29.61
DPSDT_{r2}	8	11.31	12.87	15.05	18.77	20.52	23.07	27.17	29.13	31.81
DPSDT_{r}	9	11.51	13.17	15.47	19.45	21.39	24.11	28.66	30.82	33.75
DPSDT	9	12.28	14.08	16.56	20.78	22.81	25.61	30.24	32.46	35.42
MSDT_0	7	11.31	12.87	14.99	18.49	20.13	22.43	26.17	27.96	30.33
MSDT	8	11.35	12.95	15.12	18.78	20.51	23.00	27.08	28.97	31.63

Note. p = number of parameters.

One Gibbs step cycles through the M conditional distributions of y_i given $\boldsymbol{y}^{(i)}$. Specifically, starting with $\boldsymbol{y}^{\langle 0\rangle} = \boldsymbol{y}$, do for $i = 1, \ldots, M$, in order:

- Set $m = y_i^{\langle i-1\rangle} + y_{M+1}^{\langle i-1\rangle}$.
- If $m = 0$, set $\boldsymbol{y}^{\langle i\rangle} = \boldsymbol{y}^{\langle i-1\rangle}$,
- else:
  - sample a number x from the distribution with density function proportional to the maximum-likelihood function of the saturated binomial model with a total of m trials, and
  - set $y_i^{\langle i\rangle} = x$, $y_{M+1}^{\langle i\rangle} = m - x$, and $y_j^{\langle i\rangle} = y_j^{\langle i-1\rangle}$, for $1 \leq j \leq M$, $j \neq i$.

The outcome of one such Gibbs step is given by $\boldsymbol{y}^{\langle M\rangle}$. Only $\boldsymbol{y}^{\langle M\rangle}$ is used for Steps 2 and 3 of the NML algorithm, whereas the intermediate $\boldsymbol{y}^{\langle i\rangle}$ are discarded. Furthermore, the outcome of one Gibbs step, $\boldsymbol{y}^{\langle M\rangle}$, is the starting point, $\boldsymbol{y}^{\langle 0\rangle}$, for the next Gibbs step.

Gibbs sampling is employed for Step 1 of the above NML algorithm. For Step 2 (fitting the models), it is essential that maximum likelihood estimation is implemented efficiently for each model and data pattern. We used a fast modified Newton algorithm (procedure E04LYF from the NAG FORTRAN library) for the purpose. The outcome of Step 3 of the above NML algorithm approximates the desired integral up to a multiplicative constant. In Appendix A, we describe how we computed the constant.

For the current computations we implemented an additional refinement of the algorithm that further increases its speed considerably. More details on the algorithm (convergence monitoring for the Gibbs sampler, computation of the multiplicative constant, assessment of approximation error, the additional refinement) are described in Appendix A. The FORTRAN code of the program used for NML computations can be obtained from the first author. It calls routines from the NAG and IMSL libraries for numerical computation.

The algorithm computes the set of NML indices for the set of 25 models considered here within hours for six-point rating scales and dataset sizes below $n = 100$ old and $n = 100$ new items on a fast personal computer (2 Intel® Xeon® X5675 processors with clock speed 3.07 GHz enabling 32 parallel threads on a 64 bit operating system). It takes days to compute the entire set of 25 NML penalties for larger values of M and n.

Importantly, the current algorithm is not restricted to models of recognition memory nor confidence-rating paradigms, but can be applied to any model of categorical data specifying multinomial or product-multinomial distributions, opening up a huge range of potential applications (e.g., Agresti, 1990; Bishop et al., 1975; for binomial distributions, see Klauer & Kellen, 2011a). In consequence, it is also likely to be useful for many researchers outside the fields of recognition memory, perception, and reasoning.

The speed of convergence of the algorithm depends strongly on the flexibility of the model in question: More flexible models are more similar to the proposal distribution based on the saturated model and can therefore be approximated faster. Thus, the algorithm is likely to be most useful where the models under scrutiny are not too restrictive relative to the saturated model as roughly quantified by the degrees of freedom left for the model's goodness-of-fit test. In any event, the present algorithm is of course much faster than computing NML by enumerating all data patterns, which is completely infeasible even for small datasets due to a combinatorial explosion.

6. Flexibility of the models

Table 2 shows the models' NML penalty terms for six-point rating scales and trial numbers ranging from $n = 30$ and $n = 30$ trials with old and new items, respectively, to $n = 10{,}000$ trials of each kind as might arise in analyzing aggregate data. Analogous tables for seven-point and eight-point rating scales are provided in the online supplement (see Appendix B).

Several trends can be seen in Table 2. First, the penalties of submodels are smaller than those of their superordinate models, even if the same number of parameters is employed (compare, e.g., $2\text{HTM}_{D_o\geq D_n}$ and 2HTM). Second, the flexibility of the discrete models and the DPSDT models depend strongly on the constraints imposed on the response mappings. Third, consider datasets of typical size ($n \leq 300$), order-constrained response mappings, and the models that are tailored to accommodate asymmetric

Table 3
Interpolation function details and selected FIA_f penalties.

Model	*RMSE*		*Parameters*			FIA_f
	Fit	CV	*a*	*b*	*c*	
1HTM_{wo}	0.01	0.03	4.55	0.47	2.30	2.32
1HTM_{r2}	0.01	0.02	6.59	0.44	1.24	1.27
1HTM_{r}	0.01	0.01	9.54	0.42	−0.28	−0.23
1HTM	0.01	0.03	7.30	0.45	1.50	1.56
$2\text{HTM}_{D_o=D_n,\text{wo}}$	0.01	0.01	5.01	0.47	1.87	1.87
$2\text{HTM}_{D_o=D_n,\text{r2}}$	0.01	0.02	7.54	0.41	0.19	0.22
$2\text{HTM}_{D_o=D_n,\text{r}}$	0.02	0.03	10.83	0.36	−2.07	−1.99
$2\text{HTM}_{D_o=D_n}$	0.01	0.04	8.49	0.41	−0.16	−0.11
EVSDT	0.01	0.03	3.67	0.48	3.36	
$\text{UVSDT}_{\sigma\geq 1}$	0.01	0.02	4.22	0.44	3.26	
UVSDT	0.01	0.01	3.73	0.45	3.93	
$2\text{HTM}_{D_o\geq D_n,\text{wo}}$	0.00	0.02	7.30	0.40	0.54	0.61
$2\text{HTM}_{D_o\geq D_n,\text{r2}}$	0.02	0.01	9.67	0.39	−0.67	−0.58
$2\text{HTM}_{D_o\geq D_n,\text{r}}$	0.01	0.01	12.31	0.34	−2.73	−2.49
$2\text{HTM}_{D_o\geq D_n}$	0.01	0.03	10.12	0.35	−0.92	−0.70
2HTM_{wo}	0.00	0.04	6.86	0.43	1.24	1.30
2HTM_{r2}	0.01	0.01	8.89	0.39	0.03	0.11
2HTM_{r}	0.02	0.03	11.55	0.33	−2.07	−1.80
2HTM	0.02	0.02	9.53	0.36	−0.20	−0.01
DPSDT_{wo}	0.01	0.01	6.53	0.33	1.04	
DPSDT_{r2}	0.02	0.02	10.50	0.35	−0.88	
DPSDT_{r}	0.02	0.02	14.00	0.33	−3.25	
DPSDT	0.02	0.03	10.78	0.33	−1.42	
MSDT_0	0.02	0.01	5.80	0.40	1.93	
MSDT	0.02	0.01	10.01	0.30	−1.29	

Note. CV = cross validation. The interpolation function for the NML penalty term is $\text{NML}_\text{P} = \frac{p}{2}\log\frac{2n}{2\pi} + an^{-b} + c$, where n is the number of signal trials and the number of noise trials so that the total N is $2n$, and p is the number of model parameters.

ROCs, that is $\text{UVSDT}_{\sigma\geq 1}$, $2\text{HTM}_{D_o\geq D_n}$, DPSDT, MSDT_0, and MSDT. In this competition, the discrete models with order-constrained response mappings are seen to be least flexible, followed by the hybrid models MSDT and then DPSDT with order-constrained response mapping, followed by the continuous $\text{UVSDT}_{\sigma\geq 1}$. Note, however, that $\text{UVSDT}_{\sigma\geq 1}$ employs fewest parameters in this set of models. Traditional measures such as AIC and BIC therefore attribute a greater penalty to the MSDT and to the DPSDT with order-constrained response mapping than to the UVSDT although they are seen to be less flexible than the UVSDT in terms of their ability to fit data in general (see also Kellen & Klauer, 2011).

Given that the penalties are scaled on a log-likelihood scale, the size of these differences between models in flexibility penalties is such that they will frequently alter the outcome of comparisons between models based on only the log-likelihood values, that is, on model fit. This is also true in relation to comparisons based on AIC and BIC.

Table 3 presents the parameters of an interpolation function that allows one to estimate NML values within this range (between $n = 30$ and $n = 10{,}000$) for six-point rating scales and values of n that are not tabulated. An analogous table for seven-point rating scales is provided in the online supplement (see Appendix B). Table 3 and the one in the online supplement for seven-point rating scales allow one to compute NML indices without implementing the algorithm described in the previous section.

To understand the interpolation function, note that NML is approximated by the so-called Fisher information approximation (FIA) as n becomes large. Like NML, FIA sums minus the maximum of the logarithmized likelihood of the data given the model and a penalty term FIA_P:

$$\text{FIA} = -\log f(\boldsymbol{x} \mid \hat{\theta}) + \text{FIA}_\text{P}.$$

FIA_P is the sum of two terms:

$$\text{FIA}_\text{P} = \frac{p}{2}\log\frac{m}{2\pi} + \text{FIA}_\text{f},$$

where p and m are the number of model parameters and the sample size of the data (i.e., $m = 2n$ in the present case), respectively. As can be seen, the first term in FIA_P takes the number of parameters and sample size into account in a fashion almost identical to that involved in BIC. The second term is given by:

$$\text{FIA}_\text{f} = \log\int\sqrt{\det I(\theta)}\,d\theta,$$

where I is the so-called Fisher information matrix of the model for a sample of size one. The Fisher information matrix is the matrix of the expected second partial derivatives of the log-likelihood function. FIA_f is independent of the parameterization of the model and can be seen as a measure of the model's flexibility due to its functional form.

Let the penalty term in NML be termed NML_P, that is $\text{NML}_\text{P} = \log\sum_y f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y}))$. FIA is an asymptotic approximation of NML, meaning that $\text{NML}_\text{f} = \text{NML}_\text{P} - \frac{p}{2}\log\frac{m}{2\pi}$ converges to FIA_f as m increases (e.g., Su et al., 2005). The NML_f values for $n = 30, 50, 100, 300, 500, 1000, 3000, 5000$, and $n = 10{,}000$ can be directly computed from the NML_P values shown in Table 2.

For each model, we fitted a power function, $\text{IP}(a, b, c)$, with three parameters a, b, and c to the values of NML_f for $n = 30$, 100, 300, 1000, 3000, and 10,000 using a least-squares criterion. IP was defined as:

$$\text{IP}(a, b, c) = an^{-b} + c.$$

The values of NML_f for $n = 50$, 500, and 5000 were used for cross-validation purposes. Table 3 shows the fitted parameters a, b, and c for each model as well as the root-mean-square error (*RMSE*) measure of goodness of fit and the goodness of cross-validation also in terms of *RMSE*. As can be seen the function both fits the NML_f points used for fitting very well (all $RMSE < 0.024$) and provides a satisfactory interpolation of points within the covered range (all *RMSE* for cross validation <0.04). Given the parameter values for a, b, and c in Table 3 and a value n within the range from 30 to 10,000, the penalty term of NML can thus be interpolated as:

$$\text{NML}_\text{P} = \frac{p}{2}\log\frac{2n}{2\pi} + an^{-b} + c.$$

The estimate of parameter c provides a further opportunity for cross-validating the interpolation function and at the same time for validating the NML algorithm that underlies it. Parameter c describes an asymptote of $\text{IP}(a, b, c)$ as n becomes large. To the extent to which the interpolation function correctly extrapolates NML_P even for values larger than $n = 10{,}000$ and in fact for infinite n, parameter c should estimate the limit approached by NML_f and thus, FIA_f.

On the other hand, FIA_f can be independently computed at least for the discrete models, all of which are members of the class of multinomial processing tree models (Klauer et al., 2015) using results by Wu et al. (2010a,b) implemented in MPTinR (Singmann & Kellen, 2013). In the rightmost column of Table 3 these FIA_f values are remarkably close to parameter c despite the fact that the implied extrapolation is as extreme as it can get (i.e., to infinity). This correspondence considerably increases our confidence in the NML algorithm given that the FIA values approximated by it are computed in a completely different and independent fashion (see also Table 3 in the online supplement (see Appendix B)).

7. Recovery study

The MDL principle aims to "identify a model family that permits the tightest compression of a dataset by effectively filtering out random noise and attending to all of the 'useful' information in the data" (Myung et al., 2006, p. 173). This is related to identifying the model that is most generalizable, leading to the smallest errors in predicting new data (Myung et al., 2006). A related question researchers in recognition memory are often interested in is which model most likely generated a given dataset (Klauer & Kellen, 2011a).

Assessing the suitability of NML for addressing this question requires model-recovery studies in which data are generated from a set of models and then fitted by the models. This allows one to count how frequently the generating model is correctly selected from among the set of considered models when the NML index is used for model selection and to see whether NML performs as well as and perhaps better than the more traditional AIC and BIC indices.

AIC, BIC, and NML are based on the same evidence, the models' maximum-likelihood values. They differ in how this evidence is used and in particular, how model complexity is quantified and weighed against the evidence. Models with larger penalties due to functional form will thereby be selected less frequently and models with smaller penalties more frequently under NML than under AIC and BIC, all else being equal. In other words, complex models will tend to be selected less frequently, entailing fewer correct selections if and when the complex model in fact generated the data and more correct selections if and when the simple model in fact generated the data. The question is whether the reduction in correct selections is less in the former case than the increase in correct selections in the latter case so that the overall number of correct decisions is increased when NML is used (Klauer & Kellen, 2011a).

This is not a trivial question given that NML was not developed with the goal of optimizing model recovery. For instance, as is clear from the definition of NML (for a comprehensive introduction, see Myung et al., 2006), the MDL principle quantifies complexity with respect to all datasets that can occur irrespective of whether they were generated from one of the models under consideration or not, whereas in model recovery models are contrasted with respect to datasets generated from each of these models (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). With respect to this subset of datasets, NML may or may not appropriately quantify the relative ability of the models to fit these data.

For the model-recovery analyses, six-point and seven-point rating scales were simulated; the results for the seven-point scales (which include an "unsure" midpoint) were similar to those for the six-point scales and are presented in the online supplement (see Appendix B). We considered two values for the numbers n of old items and of new items: $n = 60$ representing the smallest sizes usually found in the literature for an individual's dataset, and $n = 300$ approaching the size of the largest datasets usually reported in the literature. Each subset of the 25 models defines a new model-selection problem, and we looked at a range of those as detailed below. For each model in a given set, we generated 10,000 datasets and summarize results in terms of recovery rates, that is the percentage of simulated datasets in which the model-selection index in question selected the generating model. Overall model-recovery performance is evaluated in terms of the percentage of correct selections across generating models.

A difficult question, requiring judicious choice, is how to sample from each model. Klauer and Kellen (2011a) and Myung et al. (2007) sampled parameter values from Jeffrey's noninformative distribution, which assigns equal prior probability to every distinguishable probability distribution that is consistent with the model (Balasubramanian, 1997). Here, we took a different approach because we wanted to ensure that the generated datasets would be ones that could plausibly arise in a normal recognition memory experiment. For that reason, we sampled first a dataset of the desired size from the distribution with probability function proportional to the maximum-likelihood function of a supermodel defined by the mixture of the models UVSDT$_{\sigma\geq 1}$, 2HTM$_{D_o \geq D_n}$, DPSDT, and MSDT that generate the kind of curved asymmetric ROC typically seen in recognition-memory experiments. That is, if $p_i(\boldsymbol{y}|\boldsymbol{\theta_i})$ is the probability of observing dataset $\boldsymbol{y}$ given parameters $\boldsymbol{\theta_i}$ for the above four models indexed by $i = 1, \ldots, 4$, then the supermodel is defined by $p(\boldsymbol{y}|(\boldsymbol{\theta_i}, \lambda_i)_{i=1,\ldots,4}) = \sum_{i=1}^{4} \lambda_i p_i(\mathrm{y}|\boldsymbol{\theta_i})$ using additional mixture parameters $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$. The maximum likelihood for $\boldsymbol{y}$ under the supermodel is the maximum of the maximum likelihoods of the four models computed separately. We sampled a dataset from the probability function proportional to the maximum-likelihood function of this supermodel using rejection sampling with the density proportional to the maximum likelihood of the saturated model as dominating proposal density. To sample from a specific model, the model was then fitted to this dataset and the resulting parameter values were used to generate a new dataset of the desired size from the model.

Consider first selections from pairs of models. For pairwise comparisons, it is possible to assess the optimal level of overall model-recovery performance by identifying the optimal penalty difference between the two models through a simple grid search (Klauer & Kellen, 2011a). Table 4 shows a number of pairwise comparisons involving models that we consider most viable a priori, contrasting (a) 2HTM$_{D_o=D_n}$ and EVSDT, (b) 2HTM and UVSDT, (c) UVSDT$_{\sigma\geq 1}$ and DPSDT, (d) DPSDT and MSDT, (e) UVSDT$_{\sigma\geq 1}$ and MSDT and (f) UVSDT and MSDT. For each of the discrete-state models and DPSDT, we did this separately for the variants without response mapping (subscript "wo"), with order-restricted response mapping restricted to the two highest confidence levels (subscript "r2"), and with order-restricted response mapping (subscript "r").

As can be seen in Table 4, NML is almost never outperformed by AIC and BIC and in fact does better than these as a rule, sometimes by a considerable margin (e.g. in contrasting UVSDT and MSDT or 2HTM and UVSDT), approximating optimal performance. However, the pair DPSDT and MSDT presents a very difficult selection problem for which performance of all indices was below or close to chance. Even optimal performance hardly exceeded the 50% chance level.[4] In three of these cases, AIC and/or BIC performed relatively better than NML. Otherwise, overall recovery rates tend to be well above the 50% chance level. Not surprisingly, overall recovery rates increase as sample size n increases.

Tables 5 and 6 present recovery results for selecting from larger sets of models; Table 5 considers two sets of four models each; Table 6 selection problems comprising ten models. NML performs better than AIC and BIC, providing an advantage in model recovery that is sometimes modest and sometimes considerable in size. There is only one case in which NML and BIC are tied (in Table 6, bottom).

In the set of ten models shown in Table 6, simple models such as 1HTM, EVSDT, and 2HTM$_{D_o=D_n}$ are included, but much of the literature is focused on more complex models such as UVSDT, DPSDT, and MSDT. One motivation for these models is that they account for discrepancies between the predictions of the simple models and ROC data as well as for similar discrepancies found with other tasks (Bayen, Murnane, & Erdfelder, 1996; Yonelinas & Parks, 2007). When the simple models are taken out of the selection set, diagnosticity of the data becomes an important concern. For example, as discussed by Kellen et al. (2013; see also Jang, Wixted, & Huber, 2011 and Myung & Pitt, 2009), it could be the case that the ROC points are too close to each other or to the diagonal to provide reliable information on the function's shape, in which case the datasets are well accounted for by a common restricted model such as EVSDT. In such cases, the data can be seen as non-diagnostic. They would be well fit by all complex

[4] The difficulty in selecting from the pair of DPSDT and MSDT suggests that both models make very similar predictions for ROC data, and it underlines that model selection criteria like AIC, BIC, and NML quantify model flexibility, but not model mimicry (Navarro, Pitt, & Myung, 2004).

Table 4
Model recovery contrasting pairs of models, recovery rates per model (1 vs. 2), and overall recovery in percent.

n	Index	2HTM$_{D_o=D_n}$ vs. EVSDT			2HTM vs. UVSDT			UVSDT$_{\sigma\geq 1}$ vs. DPSDT			DPSDT vs. MSDT			UVSDT vs. MSDT		
		1	2	Tot.	1	2	Tot.	1	2	Tot.	1	2	Tot.	1	2	Tot.
		Without response mapping (wo)[a]												$\sigma \geq 1$[b]		
60	AIC	52	75	63	31	94	62	71	43	57	96	5	51	97	4	50
	BIC	52	75	63	31	94	62	71	43	57	99	3	51	99	2	50
	NML	86	49	67	68	76	72	39	84	61	92	7	50	35	83	59
	Opt.	85	49	67	76	70	73	34	89	61	99	3	51	43	75	59
		With response mapping r2[a]														
	AIC	41	85	63	38	90	64	86	9	48	50	35	42	98	4	51
	BIC	30	91	61	26	95	60	92	4	48	50	35	42	99	2	51
	NML	88	44	66	74	69	71	38	82	60	90	7	48	65	57	61
	Opt.	85	46	66	77	66	72	32	87	60	99	2	50	53	71	62
		With response mapping r[a]														
	AIC	41	86	63	40	89	65	83	8	46	9	64	36			
	BIC	24	94	59	24	95	60	90	3	47	5	72	38			
	NML	89	42	65	78	63	71	38	77	58	15	56	36			
	Opt.	86	45	66	81	61	71	32	86	59	99	1	50			
		Without response mapping (wo)[a]												$\sigma \geq 1$[b]		
300	AIC	67	80	73	55	96	75	70	52	61	94	14	54	93	12	53
	BIC	67	80	73	55	96	75	70	52	61	98	9	54	98	7	53
	NML	89	63	76	75	88	82	38	89	64	92	15	54	31	89	60
	Opt.	88	65	76	76	87	82	42	85	64	95	12	54	34	86	60
		With response mapping r2														
	AIC	64	85	75	67	91	79	85	10	48	51	43	47	95	12	54
	BIC	56	90	73	56	94	75	92	4	48	51	43	47	99	7	53
	NML	91	59	75	82	80	81	36	88	62	84	19	51	69	64	67
	Opt.	88	62	75	81	81	81	34	90	62	96	11	53	66	68	67
		With response mapping r[a]														
	AIC	70	82	76	73	86	79	81	10	45	10	68	39			
	BIC	59	89	74	58	92	75	88	5	46	4	76	40			
	NML	77	73	75	85	75	80	39	78	58	12	65	39			
	Opt.	71	81	76	88	72	80	30	92	61	96	10	53			

Note. Tot. = overall recovery rate. Opt. = optimal recovery rate.

[a] Regards the 2HTM variants and the DPSDT.

[b] Regards the UVSDT model.

Table 5
Model recovery contrasting sets of four models, recovery rates per model (1–4), and overall recovery in percent.

n	Index	2HTM$_{D_o=D_n}$, EVSDT, 2HTM, UVSDT					2HTM$_{D_o\geq D_n}$, UVSDT$_{\sigma\geq 1}$, DPSDT, MSDT				
		1	2	3	4	Tot.	1	2	3	4	Tot.
		Without response mapping (wo)[a]									
60	AIC	30	41	18	75	41	20	85	10	15	33
	BIC	38	53	12	66	43	20	85	10	15	33
	NML	61	30	28	61	45	65	60	24	15	41
		With response mapping r2[a]									
	AIC	22	45	29	72	42	21	85	1	20	32
	BIC	19	65	17	66	42	14	87	0	22	31
	NML	60	25	42	57	46	69	58	5	20	38
		With response mapping r[a]									
	AIC	20	46	34	71	43	22	83	1	20	32
	BIC	14	66	18	66	41	13	87	0	22	31
	NML	60	24	50	54	47	53	55	2	40	37
		Without response mapping (wo)[a]									
300	AIC	41	47	41	84	53	37	84	19	24	41
	BIC	53	61	34	74	55	37	84	19	24	41
	NML	67	46	43	72	57	71	65	37	22	49
		With response mapping r2[a]									
	AIC	41	50	59	79	57	41	83	2	28	39
	BIC	45	69	45	72	58	33	86	0	31	38
	NML	71	42	60	67	60	58	64	7	49	44
		With response mapping r[a]									
	AIC	44	48	67	75	59	46	80	2	27	38
	BIC	45	69	50	71	58	35	84	1	31	38
	NML	63	52	69	62	61	40	59	6	54	40

Note. Tot. = overall recovery rate.

[a] Regards the 2HTM variants and the DPSDT.

Table 6
Model recovery contrasting a set of ten models, recovery rates per model (1–10), and overall recovery in percent.

n	Index	1HTM, 2HTM$_{D_o=D_n}$, 2HTM$_{D_o\geq D_n}$, 2HTM, EVSDT, UVSDT$_{\sigma\geq 1}$, UVSDT, DPSDT, MSDT$_0$, MSDT										
		1	2	3	4	5	6	7	8	9	10	Tot.
		Without response mapping (wo)[a]										
60	AIC	27	23	2	8	36	18	53	2	7	0	18
	BIC	36	30	1	5	46	15	47	0	5	0	18
	NML	25	55	6	9	24	31	27	3	11	1	19
		With response mapping r2[a]										
	AIC	23	19	2	11	43	18	51	0	8	0	18
	BIC	21	17	0	6	63	16	47	0	7	0	18
	NML	35	54	7	13	19	31	24	1	7	0	19
		With response mapping r[a]										
	AIC	26	17	1	11	44	18	50	0	12	0	18
	BIC	20	11	0	5	64	16	47	0	10	0	17
	NML	40	54	6	15	18	31	21	1	6	1	19
		Without response mapping (wo)[a]										
300	AIC	36	35	6	22	43	18	59	9	16	1	25
	BIC	47	45	4	16	56	16	53	5	12	0	25
	NML	35	62	13	16	40	28	35	10	17	2	26
		With response mapping r2[a]										
	AIC	33	39	8	28	48	18	56	1	16	1	25
	BIC	35	43	4	19	68	16	52	0	15	0	25
	NML	45	67	16	21	36	28	32	2	12	2	26
		With response mapping r[a]										
	AIC	40	42	7	29	46	17	53	0	19	1	25
	BIC	40	42	2	19	67	16	50	0	19	0	26
	NML	40	60	14	23	46	28	28	1	19	3	26

Note. Tot. = overall recovery rate.
[a] Regards the models 1HTM, the 2HTM variants, and DPSDT.

Table 7
Model recovery excluding non-diagnostic data, recovery rates per model (1–7), and overall recovery in percent.

n	Index	2HTM$_{D_o\geq D_n}$, 2HTM UVSDT$_{\sigma\geq 1}$, UVSDT, DPSDT, MSDT$_0$, MSDT							
		1	2	3	4	5	6	7	Tot.
		Without response mapping (wo)[a]							
60	AIC	8	20	29	68	10	19	0	22
	BIC	8	20	29	68	10	19	0	22
	NML	19	22	57	38	9	25	1	24
		With response mapping r2[a]							
	AIC	8	26	30	68	1	15	1	21
	BIC	5	18	31	70	0	18	0	20
	NML	22	30	59	36	3	18	1	24
		With response mapping r[a]							
	AIC	7	27	32	67	0	16	1	21
	BIC	4	16	32	70	0	19	0	20
	NML	22	35	62	33	2	16	2	25
		Without response mapping (wo)[a]							
	AIC	17	39	29	67	24	33	2	30
	BIC	17	39	29	67	25	34	1	30
	NML	36	31	49	45	26	38	4	33
		With response mapping r2[a]							
	AIC	21	49	29	66	3	26	3	28
	BIC	18	42	30	68	1	32	1	27
	NML	46	40	50	42	6	30	5	31
		With response mapping r[a]							
	AIC	17	48	29	62	2	36	3	28
	BIC	13	39	31	66	1	41	1	27
	NML	37	41	50	37	3	39	7	31

Note. Tot. = overall recovery rate.
[a] Regards the 2HTM variants and the DPSDT.

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 implemented in model selection by NML.

Because we screen out non-diagnostic data in some of the meta-analyses reported below, Table 7 shows how this affects model recovery from among the more complex models 2HTM$_{D_o\geq D_n}$, 2HTM, UVSDT$_{\sigma\geq 1}$, UVSDT, DPSDT, MSDT$_0$, and MSDT considered in Table 6. For each selection problem in Table 7, datasets were excluded from the recovery counts if NML preferred one of the simpler models EVSDT, 1HTM, or 2HTM$_{D_o=D_n}$. This excludes ROCs that can be accounted for by common restricted models and in particular ROCs with points that are too close to each other or to the diagonal to be diagnostic. As can be seen, the performance advantage of the minimum-description length approach is preserved.

8. Meta-analysis

8.1. Datasets

We reanalyzed confidence-rating ROC data published in the literature. These stem from Benjamin, Tullis, and Lee (2013), Dube and Rotello (2012), Heathcote, Ditton, and Mitchell (2006), Jaeger, Cox, and Dobbins (2012), Jang, Wixted, and Huber (2009), Koen, Aly, Wang, and Yonelinas (2013), Koen and Yonelinas (2011), Onyper et al. (2010), and Smith and Duncan (2004).

There were datasets from 15 studies and 850 individuals, 459 individual datasets based on six-point rating scales, 391 on eight-point rating scales, with size n of each category system ranging from 60 to 384. Table 8 presents a few descriptive statistics for each study. Although these datasets do by no means exhaust the set of confidence-rating data in existence, they stem from multiple laboratories using different procedures and materials, making it unlikely that they are systematically biased in favor of one of the different models under study here.[5]

[5] Participants of Study 7 (see Table 8) also provided Remember/Know judgments (Tulving, 1985) but did not receive any further instructions besides being encouraged to use the whole scale when expressing their confidence.

Table 8
Description of studies in the meta-analysis.

Study	Id	N	n
6-point ROCs			
Dube and Rotello (2012, Exp. 1B, pictures)	1	27	200
Dube and Rotello (2012, Exp. 1B, words)	2	22	200
Heathcote et al. (2006, Exp. 1)	3	16	280
Heathcote et al. (2006, Exp. 2)	4	23	280
Jaeger et al. (2012, Exp. 1, no cue)	5	63	60
Jang et al. (2009)	6	33	70
Koen and Yonelinas (2010, pure study)	7	32	160
Koen and Yonelinas (2011)	8	20	300
Koen et al. (2013, Exp. 2, full attention)	9	48	100
Koen et al. (2013, Exp. 4, immediate test)	10	48	150
Pratte et al. (2010)	11	97	240
Smith and Duncan (2004, Exp. 2)	12	30	70
8-point ROCs			
Benjamin et al. (2013)	13	124	60
Onyper et al. (2010, Exp.1, pictures)	14	136	384
Onyper et al. (2010, Exp.1, words)	15	131	384

Note. Id = Study number (see Table 9); N = number of participants; n = number of test trials with old items (equals number of trials with new items).

8.2. Overall analyses

We computed the NML indices for each individual dataset and the 25 models considered here. Table 9 shows the NML values summed across individual datasets per study, the total sums, and the number of times each model was selected as best model (i.e., was associated with the smallest NML value) across these datasets. The summed NML value (see column "Total") is simply the NML index of the model fitted to all individual datasets simultaneously, with different parameter values permitted for each individual dataset.[6]

The rows of Table 9 are ordered so that the total summed NML values increase from top to bottom. As can be seen, in terms of summed NML values, models from the DPSDT family, the MSDT family, and the 2HTM$_{D_o \geq D_n}$ family perform best. In terms of the number of individual datasets for which each model is selected as best model (vote counting), the six models with best summed NML values also perform well (see rightmost column of Table 9). The classical DPSDT$_{\text{wo}}$ without response mapping stands out from among these models receiving 120 votes, the next highest total being 72 votes for 2HTM$_{D_o \geq D_n,\text{r}}$.

There are, however, a number of models with high vote counts as well as high summed NML values. Such models are thus associated with the smallest NML values for a sizeable number of datasets, yet on average produced NML values larger than those of many other models for the remaining datasets. This phenomenon occurred in particular for many of the simpler models, that is for models from the 1HTM family, the 2HTM$_{D_o = D_n}$ family, and the EVSDT. As already sketched, a simple explanation is that there are many non-diagnostic datasets that are fit well by one of the simpler models. For such datasets, fit values cannot further profit much from the complexity added by the DPSDT, UVSDT, MSDT, and 2HTM models. In consequence, NML selects one of the simpler models as most parsimonious description for the non-diagnostic datasets. For diagnostic datasets, on the other hand, the simple models typically receive poor fit values, accounting for their poor performance in terms of summed NML.

Another pattern worth mentioning is that the models with unconstrained response mappings, DPSDT, 2HTM, 1HTM, 2HTM$_{D_o \geq D_n}$, and 2HTM$_{D_o = D_n}$ are consistently outperformed by constrained versions of them (models with subscripts "r" and "r2") both in terms of summed NML values as well as in terms of vote counting. Models with constrained mappings also outperformed the versions without response mapping (subscript "wo"), an exception being DPSDT$_{\text{wo}}$ that was the best model in terms of vote counting. Note, however, that DPSDT$_{\text{r}}$ and DPSDT$_{\text{r2}}$ still outperformed DPSDT$_{\text{wo}}$ in terms of summed NML values. Taken together, these result patterns constitute evidence for the appropriateness of the constrained response mappings.

A final observation on Table 9 concerns the popular UVSDT and UVSDT$_{\sigma \geq 1}$ models. They are beaten by many models in terms of summed NML values and take places at the bottom of the rank order in terms of vote counting.

Statistically, a Friedman rank sum test reveals that the differences between the models in terms of NML are significant (with individual dataset as unit of analysis): $\chi^2(24) = 6084.39$, $p < 0.001$. The same is true of the differences in vote counts: $\chi^2(24) = 733.76$, $p < 0.001$. We computed pairwise two-tailed Wilcoxon tests for models adjacent in the rank order of summed NML values as shown in Table 9 for the six models with smallest summed NML values (we restricted ourselves to the first six models because only these also received high vote counts). All pairwise comparisons were significant with $p < 0.05$ (Bonferroni–Holm corrected), excepting that between DPSDT$_{\text{wo}}$ and 2HTM$_{D_o \geq D_n,\text{r}}$ with $p = 0.07$ (Bonferroni-Holm corrected). In terms of vote counting, the analogous analysis revealed that only the difference between DPSDT$_{\text{wo}}$ and 2HTM$_{D_o \geq D_n,\text{r}}$ was significant (120 and 72 selections, respectively, $p = 0.003$ according to a binomial test with Bonferroni-Holm correction).

8.3. Analyses restricted to diagnostic datasets

For a second set of analyses, we excluded all 380 individual datasets that were best fit by one of the simple models. that is by one of the models from the 1HTM and 2HTM$_{D_o = D_n}$ families as well as by EVSDT. Such datasets are non-diagnostic for discriminating between the remaining, more complex models. Excluding these datasets leaves 470 datasets in the analysis. The exclusion cannot alter the vote-counting results for the more complex models, but it affects their summed NML values. Table 10 presents the summed NML values.

As can be seen, MSDT advances by one rank place to second-best in the rank order by summed NML values (see the column labeled "Total"). Again, the differences between models are significant ($\chi^2(15) = 2672.68$, $p < 0.001$, according to a Friedman rank-sum test). In addition, in pairwise two-tailed Wilcoxon tests between models with adjacent ranks among the uppermost six models in Table 10, all differences were significant with $p < 0.05$ (Bonferroni-Holm corrected).

8.4. Analyses restricted to non-diagnostic datasets

For the sake of completeness, we considered the 380 datasets previously excluded to test which of the simple models performed best in describing these. Again, the vote-counting results are of course the same as in Table 9, but we nevertheless add a column with vote counts to Table 11, which summarizes the results for the non-diagnostic datasets, for the sake of readability. In terms of vote counts, most models were frequently selected as best, the bottom places being taken by the models with unconstrained response mappings and EVSDT. In terms of summed NML values, models from the 1HTM family (excepting 1HTM$_{\text{wo}}$) perform best, followed by EVSDT, and models from the 2HTM$_{D_o = D_n}$ family.

[6] It is possible to quantify the evidence conveyed by NML through the computation of NML weights (e.g., (Vandekerckhove, Matzke, & Wagenmakers, 2015) We computed the NML weights for the summed NML values for each of the analyses reported here and found that the weights expressed virtually complete preference for the model with the smallest NML, reflecting the large differences between summed NML values on the log-likelihood scale.

Table 9
Summed NML indices per study and across studies (total) and selection frequencies ordered so that total decreases from top to bottom.

Model	Study															Total	Freq.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
DPSDT$_{\mathrm{r2}}$	507	**410**	322	472	934	505	578	406	**768**	**852**	**1889**	454	2351	**3677**	**3558**	**17 684**	58
DPSDT$_{\mathrm{r}}$	509	413	320	**463**	927	503	**572**	406	773	855	1903	454	2348	3765	3606	17 818	25
MSDT	**505**	417	**318**	465	947	505	575	**404**	791	863	1898	454	2387	3728	3583	17 842	24
DPSDT$_{\mathrm{wo}}$	524	415	332	485	946	509	605	413	797	852	1906	452	2367	3774	3662	18 040	120
2HTM$_{D_o \geq D_n,\mathrm{r}}$	528	420	336	504	**926**	**502**	574	415	777	854	1957	**448**	**2329**	3763	3712	18 047	72
MSDT$_0$	514	428	330	492	956	513	584	411	798	870	1965	460	2423	3878	3614	18 235	45
2HTM$_{\mathrm{r}}$	530	433	345	515	957	518	592	428	806	877	2005	462	2393	3846	3783	18 492	1
DPSDT	542	439	341	491	974	532	608	427	824	904	2016	481	2514	4061	3879	19 032	4
2HTM$_{D_o \geq D_n}$	565	450	357	530	978	538	612	434	838	911	2054	483	2516	4060	3930	19 255	2
2HTM$_{D_o \geq D_n,\mathrm{r2}}$	543	429	399	713	1023	517	610	452	809	881	2077	462	2442	3919	4083	19 360	54
UVSDT$_{\sigma \geq 1}$	541	448	346	499	1013	546	632	429	845	915	2063	492	2555	4191	3849	19 365	0
1HTM$_{\mathrm{r}}$	636	441	365	611	1006	508	588	458	808	887	2221	466	2422	4070	3984	19 471	60
2HTM	567	462	366	540	1006	552	629	445	865	932	2102	495	2577	4144	4003	19 684	0
2HTM$_{\mathrm{r2}}$	545	442	409	724	1054	535	627	464	833	907	2124	479	2497	3996	4147	19 783	0
UVSDT	557	461	356	513	1041	563	651	441	874	942	2123	510	2626	4279	3931	19 869	0
1HTM	674	472	386	634	1054	544	624	477	867	943	2317	500	2601	4355	4171	20 620	1
1HTM$_{\mathrm{r2}}$	684	461	517	1038	1165	532	650	537	888	947	2545	493	2607	4474	4753	22 292	59
EVSDT	629	487	455	601	981	573	727	497	903	942	2230	523	2551	5539	4739	22 378	32
2HTM$_{D_o \geq D_n,\mathrm{wo}}$	711	530	617	1370	1301	556	789	635	1026	1005	2887	513	2584	4759	5293	24 574	63
2HTM$_{\mathrm{wo}}$	710	541	627	1382	1322	571	806	645	1049	1030	2934	527	2628	4829	5350	24 953	2
2HTM$_{D_o = D_n,\mathrm{r}}$	733	673	545	934	1196	660	837	918	1149	1472	2529	601	2989	6247	5519	27 001	56
2HTM$_{D_o = D_n}$	766	695	565	946	1232	685	863	917	1199	1520	2597	630	3128	6505	5681	27 930	5
2HTM$_{D_o = D_n,\mathrm{r2}}$	743	677	591	1092	1271	673	871	940	1174	1495	2621	612	3113	6402	5871	28 146	48
1HTM$_{\mathrm{wo}}$	1043	586	755	1760	1495	575	848	770	1146	1097	3551	560	2782	5507	6121	28 595	67
2HTM$_{D_o = D_n,\mathrm{wo}}$	936	793	851	1816	1585	728	1050	1133	1415	1650	3524	685	3314	7364	7176	34 020	52

Note. Freq. = Number of times the model is selected as best. In each column, the smallest NML value is set in bold.

Table 10
Summed NML indices per study and across studies (total) ordered by total for the diagnostic datasets.

Model	Study															Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
DPSDT$_{\mathrm{r2}}$	285	207	242	350	299	198	250	289	**321**	426	1203	219	1044	**2548**	**2295**	**10 175**
MSDT	**282**	207	**239**	**343**	299	197	253	**284**	328	429	1204	219	1059	2583	2315	10 242
DPSDT$_{\mathrm{r}}$	284	207	242	346	295	199	**250**	287	325	430	1215	218	1036	2624	2339	10 296
DPSDT$_{\mathrm{wo}}$	289	**206**	249	363	300	**195**	268	296	329	**419**	**1203**	216	1046	2611	2377	10 368
2HTM$_{D_o \geq D_n,\mathrm{r}}$	298	216	257	384	**294**	199	250	295	326	426	1265	**215**	**1029**	2640	2445	10 539
MSDT$_0$	289	218	253	368	308	206	260	290	337	439	1264	225	1087	2726	2355	10 626
2HTM$_{\mathrm{r}}$	299	222	264	392	305	206	258	304	338	438	1294	220	1058	2698	2489	10 784
DPSDT	302	219	257	365	308	211	266	300	347	454	1290	230	1104	2830	2511	10 994
2HTM$_{D_o \geq D_n}$	318	231	273	402	311	214	266	306	351	454	1330	231	1103	2842	2559	11 191
UVSDT$_{\sigma \geq 1}$	303	227	258	370	325	223	285	303	356	460	1326	238	1139	2941	2502	11 258
2HTM	320	237	279	409	320	221	274	314	362	465	1358	234	1130	2901	2605	11 430
UVSDT	313	234	266	381	336	231	294	311	368	474	1365	246	1170	3003	2555	11 545
2HTM$_{D_o \geq D_n,\mathrm{r2}}$	317	230	317	574	366	208	276	328	343	443	1376	223	1122	2790	2756	11 670
2HTM$_{\mathrm{r2}}$	318	236	324	582	376	215	284	337	354	457	1404	231	1143	2844	2795	11 899
2HTM$_{D_o \geq D_n,\mathrm{wo}}$	436	316	509	1157	502	223	399	480	450	505	2014	249	1199	3516	3660	15 613
2HTM$_{\mathrm{wo}}$	435	321	516	1167	511	229	406	487	460	518	2042	256	1215	3565	3696	15 823

Note. In each column, the smallest NML value is set in bold.

Table 11
Summed NML indices per study and across studies (total) and selection frequencies ordered by total for the non-diagnostic datasets.

Model	Study															Total	Freq.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
1HTM$_{\mathrm{r}}$	288	212	**77**	155	695	**301**	**324**	141	**457**	**434**	**753**	**236**	**1346**	**1171**	**1313**	**7 901**	60
1HTM	304	227	82	163	727	322	344	149	491	462	786	255	1452	1263	1414	8 443	1
1HTM$_{\mathrm{r2}}$	295	**207**	85	219	761	311	344	155	497	459	815	247	1402	1200	1493	8 488	59
EVSDT	275	229	126	**152**	**662**	333	401	**127**	528	476	817	279	1409	1557	1657	9 029	32
2HTM$_{D_o = D_n,\mathrm{r}}$	253	249	120	181	751	382	388	292	674	759	764	311	1566	1656	1555	9 900	56
2HTM$_{D_o = D_n,\mathrm{r2}}$	**249**	247	124	200	775	390	400	291	683	768	779	317	1601	1674	1634	10 133	48
1HTM$_{\mathrm{wo}}$	424	226	113	326	941	337	409	197	626	533	1054	279	1487	1344	1867	10 161	67
2HTM$_{D_o = D_n}$	268	262	125	189	775	395	405	299	704	783	787	327	1651	1736	1645	10 351	5
2HTM$_{D_o = D_n,\mathrm{wo}}$	307	264	152	274	938	417	456	330	809	845	961	349	1690	1811	1951	11 553	52

Note. Freq. = Number of times the model is selected as best. In each column, the smallest NML value is set in bold.

Again, the differences between models in summed NML values are significant according to a Friedman rank sum test: $\chi^2(8) = 233.83$, $p < 0.001$. In addition, in pairwise two-tailed Wilcoxon tests between models with adjacent ranks among the uppermost four models in Table 11, all differences were significant with $p < 0.05$ (Bonferroni-Holm corrected).

8.5. Summary

Taken together, the meta-analysis suggests that models from the DPSDT family perform best in describing traditional confidence-rating based ROC data. In fact, the classical DPSDT model without response mapping, DPSDT$_{\text{wo}}$, performed very well both in terms of the frequency of being selected as best model in the minimum-description length framework as well as in terms of being associated with low overall NML values. In weighing the good performance of the DPSDT family, it is important to note that to the best of our knowledge none of the datasets used strict instructions such as in the remember–know paradigm (Tulving, 1985) that could be argued to favor DPSDT (see Footnote 4). Close competitors of this model were, however, the DPSDT models with constrained response mappings, DPSDT$_{\text{r2}}$ and DPSDT$_{\text{r}}$, MSDT$_0$ and MSDT, as well as 2HTM$_{D_o \geq D_n,\text{r}}$. In line with previous findings, these models accommodate asymmetric ROCs. This is also true of the popular UVSDT$_{\sigma \geq 1}$; but its flexibility along with that of the UVSDT model itself was not outweighed by goodness of fit across these datasets.

Both models with unconstrained response mappings as well as models without probabilistic response mappings (subscript "wo") were generally dominated by models from the same family with constrained response mapping (subscripts "r2" and "r") in terms of both summed NML values and vote counting (excepting DPSDT$_{\text{wo}}$ which was best in terms of vote counting), indicating that there is evidence for the appropriateness of constrained response mappings. Remember that such mappings permit the probabilistic choice of less than highest confidence levels from detect and recollection states, but constrains the likelihood of such choices to decrease as confidence level decreases or to zero for confidence levels below the second-highest one.

9. Discussion

In this manuscript, we described a new algorithm for computing the NML index for models of categorical data with multinomial or product-multinomial distributions. The speed of convergence of the method decreases as the model in question becomes simpler, as roughly quantified by the degrees of freedom available for assessing model fit.

We applied the new method to computing penalties for major models of ROC data based on confidence ratings. We present tables of NML penalties for datasets of typical sizes as well as interpolation functions that allow one to interpolate penalties for datasets with size between the tabulated ones. Some of these models are also applied in work on perception and reasoning so that the current results should also be useful for researchers working in these domains. Beyond these domains, the present method for computing NML values is applicable in principle to the entire vast field of model-based categorical data analysis (e.g., Agresti, 1990; Bishop et al., 1975).

NML allows one to compare the major models in terms of their flexibility weighing in functional form. For example, for datasets of typical sizes, the discrete models from the 2HTM$_{D_o \geq D_n}$ family with order-constrained response mappings were least flexible, followed by the hybrid models MSDT and then DPSDT with order-constrained response mapping, followed by the continuous UVSDT$_{\sigma \geq 1}$.

This also exemplifies one of the strengths of the MDL approach: It accounts for the loss in model flexibility that stems from inequality constraints imposed upon parameter estimates. Having inequality constrained models such as UVSDT$_{\sigma \geq 1}$ is often well justified on theoretical grounds a priori and helps to increase the robustness of the model and its parameter estimates when the constraints are met. Yet, AIC and BIC do not react to inequality constraints and thereby implicitly penalize models with constraints relative to the unconstrained versions, because constrained and unconstrained versions employ the same numbers of parameters. In relatively large model-recovery studies, we generated artificial datasets with characteristics that are typical of empirical ROC data from confidence-rating studies. These studies showed that the use of NML will regularly lead to improvements in model-selection performance relative to the use of AIC and BIC.

We capitalized on NML's capability to account for the loss in model flexibility related to inequality constraints by considering a number of constrained models. These constraints concern the mapping of detect or recollection states on confidence ratings and ensure that psychologically implausible state-response mappings are ruled out. In a meta-analysis of existing confidence-based ROC data using the modern model-selection techniques, we found that the original DPSDT model and constrained versions of it fared best, followed by MSDT, MSDT$_0$, and a constrained version of 2HTM$_{D_o \geq D_n}$, whereas there was little support for UVSDT$_{\sigma \geq 1}$ and UVSDT.

This partially agrees with the results of a meta-analysis based on binary OLD/NEW ROC data in which ROCs are generated by real manipulations of response bias (Kellen et al., 2013). The former meta-analysis likewise did not favor the UVSDT family, and 2HTM$_{D_o \geq D_n}$ fared relatively well in both meta-analyses. On the other hand, the binary data did not favor the DPSDT as clearly as the confidence-rating data, and the binary data did not support models from the MSDT family. Future research might address the question whether these results imply that the two methods by which ROC data are collected, via confidence ratings or via experimental manipulations of response bias, are not really equivalent or whether there is some simpler alternative account of the discrepancies. For instance, it is possible that contrary to common wisdom, response-bias manipulations as implemented for the binary data also affect memory performance (e.g., Bröder et al., 2013; Van Zandt, 2000) and not only response bias. In a similar vein, the tacit assumption of many confidence-rating based signal-detection analyses that the transition from low-confidence NEW judgments to low-confidence OLD judgments is of the same nature psychologically as the transition between two adjacent confidence levels on the NEW or OLD side might be wrong. As it stands, it may be wise to cross-validate findings obtained with one method (e.g., through the use of the confidence-rating paradigm) by a replication using an alternative method (e.g., the second-choice paradigm; Kellen & Klauer, 2011), thereby also avoiding what has been termed mono-operationalism bias (e.g., Shadish, Cook, & Campbell, 2002, Chapter 3).

NML has a simple interpretation: It weighs the goodness of fit of a model against its ability to fit data in general. NML thus penalizes flexibility by integrating the maximum likelihood function across the set of possible data. In a similar vein, Bayes factors penalize flexibility by integrating the likelihood function across each model's parameter space. One interesting avenue for future research concerns the relationship between NML and Bayes factors (Karabatsos & Walker, 2006; Zhang, 2011). Despite the just-mentioned differences, both methods often agree, and there are recent efforts to integrate them (Shiffrin, 2014; Zhang, 2011). One practical advantage of NML over Bayes factors is that in the case of the former the penalties only have to be computed once for a given experimental design, while Bayes factors have to be computed for each obtained dataset anew. Also, contrary to Bayes factors, NML penalties do not depend on the specification of prior parameter distributions (see Liu & Aitkin, 2008).

The development and comparison of models is an endeavor that requires thoughtful and judicious decisions on the part of the modeler. The criteria by which models are compared largely depend on the nature of the models and the goals which one hopes to achieve. This means that model-selection indices such as

NML should be seen as one tool among many that aid researchers in the pursuit of their goals rather than as arbiters of truth. For example, one of the outcomes of the present study is that it is difficult to discriminate between MSDT and DPSDT on the basis of confidence-rating based ROC data (see the recovery-study results), implying that more complex datasets with manipulations targeted at maximizing potential differences between the two models and their predictions need to be collected. Nevertheless, in assessing the models relative to such datasets, criteria that integrate fit and flexibility are likely to again play a role.

Acknowledgments

We thank Aaron Benjamin, Ian Dobbins, Chad Dube, Andrew Heathcote, Marc Howard, Yoonhee Jang, Joshua Koen, and Mike Pratte for making their raw data available. The research reported in this paper was supported by grant Kl 614/32-2 from the Deutsche Forschungsgemeinschaft to the first author.

Appendix A. Convergence of Gibbs sampler, refined algorithm, normalizing constant, approximation error, numerical issues

Convergence of the Gibbs sampler

We diagnosed the convergence of the Gibbs sampler using the R statistic proposed by Gelman et al. (2004, Chapter 11). We generated frequencies using the Gibbs sampler in several independent parallel streams (between 8 and 32 depending on the machine we were working on) starting with frequencies generated from a Dirichlet-multinomial distribution with Dirichlet parameters $\alpha_i = 1$. The R statistic assesses whether the variability within streams equals that to be expected between streams once convergence has been obtained. Values close to 1.0 are taken to indicate convergence.

We computed R statistics for the generated response frequencies separately for each confidence level and trials with new or old items as well as for the maximum likelihood of the frequency data under the saturated model. The burn-in phase of the Gibbs sampler was considered completed once all of these R values were smaller than 1.05 as checked every 100 cycles per stream. This occurred very quickly, usually after the first 100 cycles were completed, indicating that the Gibbs sampler converged very fast.

A refinement of the algorithm for computing NML

As already discussed by Klauer and Kellen (2011a), the speed of convergence of the algorithm can be increased by measures that make the sampling density more similar to the integrands. The sampling density described in the body of the paper is proportional to the maximum likelihood of data patterns under the saturated model. Departing from this sampling scheme, it is possible to sample relatively efficiently, using rejection sampling, from a more restricted model that respects a restriction on hit rates and false alarm rates motivated by the UVSDT model. Sampling from the restricted model led to faster convergence, especially for the more restrictive recognition-memory models considered here that are relatively dissimilar to the saturated model on the one hand, but respect the UVSDT restriction just mentioned on the other hand. At the same time, the restricted sampling model was still sufficiently flexible so that convergence for the more complex recognition-memory models was not noticeably slowed.

Specifically, we noted that many of the models considered impose restrictions on each single pair of false-alarm and hit-rate probability defined for each confidence level $r = 1, \ldots, M+1$. Let $p_n(r)$ and $p_o(r)$ be defined as

$$p_n(r) = \sum_{x=r}^{M+1} P(R = x \mid \text{new}),$$

$$p_o(r) = \sum_{x=r}^{M+1} P(R = x \mid \text{old}).$$

Many models imply that $p_o(r) \geq p_n(r),\ r = 1, \ldots, M+1$, but the UVSDT is less restrictive in that it only implies that if $p_n(r) \geq \frac{1}{2}$ then $p_o(r) \geq \frac{1}{2}$ as is easy to see. This is in fact necessary and sufficient for the single pair of probabilities to be consistent with the UVSDT for binary data (including the boundary case with $\sigma = \infty$ as an instance of the model). In the following, we describe a rejection-sampling scheme that imposes a slightly relaxed version of this restriction: For all r, if $p_n(r) > \frac{1}{2}$ then $p_o(r) \geq \frac{1}{2}$. We will refer to this restriction as the UVSDT restriction. For use in rejection sampling, we define a function g for each data pattern that is close to the maximum-likelihood l of the data pattern under the saturated model with $g \leq l$ and that defines a distribution (after normalization) respecting the UVSDT restriction for each data pattern.

Let $y_{r,t}$ be the frequency of response r, $r = 1 \ldots, M+1$, given trials with new ($t = n$) or studied items ($t = o$), and let q_t be the number of trials of kind t that were observed. Finally, let l_t be the maximum likelihood of the saturated model for the frequencies of responses $r = 1, \ldots, M+1$ for trial type t,

$$l_t = \binom{q_t}{y_{1,t} \ldots y_{M+1,t}} \prod_{r=1}^{M+1} \left(\frac{y_{r,t}}{q_t}\right)^{y_{r,t}}.$$

Define the frequencies $y_n(r)$ and $y_o(r)$ corresponding to $p_n(r)$ and $p_o(r)$ as follows:

$$y_n(r) = \sum_{x=r}^{M+1} y_{r,n},$$

$$y_o(r) = \sum_{x=r}^{M+1} y_{r,o}.$$

Let furthermore r_1 be the smallest r, $1 \leq r \leq M+1$, for which $\frac{y_o(r)}{q_o} < \frac{1}{2}$ and set $r_1 = M+2$, if no such r exists. Note that $r_1 \geq 2$. Let r_2 be the largest r, $1 \leq r \leq M+1$, for which $\frac{y_n(r)}{q_n} > \frac{1}{2}$. If $r_1 > r_2$, then the above UVSDT restriction is satisfied for the relative hit and false-alarm rates, $\frac{y_o(r)}{q_o}$ and $\frac{y_n(r)}{q_n}$, for each confidence level r, $1 \leq r \leq M+1$. If $r_1 \leq r_2$, then the restriction is violated for all r between r_1 and r_2, including r_1 and r_2.

In the first case ($r_1 > r_2$), the function g of the model that imposes the above UVSDT restrictions on the underlying probabilities $P(R = r \mid \text{new})$ and $P(R = r \mid \text{old})$ is set equal to the maximum likelihood under the saturated model and thus $g = l = l_n l_o$.

Otherwise, g is defined by

$$g = l_n \binom{q_o}{y_{1,o} \ldots y_{M+1,o}} \left(\frac{1}{2}\right)^{q_o} \prod_{r=1}^{r_2-1} \left(\frac{y_{r,o}}{q_o - y_o(r_2)}\right)^{y_{r,o}} \times \prod_{r=r_2}^{M+1} \left(\frac{y_{r,o}}{y_o(r_2)}\right)^{y_{r,o}}$$

or

$$g = \binom{q_n}{y_{1,n} \ldots y_{M+1,n}} \left(\frac{1}{2}\right)^{q_n} \prod_{r=1}^{r_1-1} \left(\frac{y_{r,n}}{q_n - y_o(r_1)}\right)^{y_{r,n}} \times \prod_{r=r_1}^{M+1} \left(\frac{y_{r,n}}{y_n(r_1)}\right)^{y_{r,n}} l_o$$

whichever is larger.

To motivate this setting note that for a fixed r', $2 \leq r' \leq M+1$, the probability of the frequencies $(y_{r,t})_{r=1,\ldots,M+1}$ for a given kind of trial t under the multinomial distribution with parameters $(p_{r,t})_{r=1,\ldots,M+1}$ and q_t can be decomposed into the product of (a) a binomial distribution with parameters $p_t(r')$ and q_t and frequency count $y_t(r')$ and (b) two multinomial distributions, one with frequency counts $(y_r)_{r=1,\ldots,r'-1}$ and parameters $(\frac{p_{r,t}}{1-p_t(r')})_{r=1,\ldots,r'-1}$ and $q_t - y_t(r')$, and one with frequency counts $(y_r)_{r=r',\ldots,M+1}$ and parameters $(\frac{p_{r,t}}{p_t(r')})_{r=r',\ldots,M+1}$ and $y_t(r')$:

$$\binom{q_t}{y_{1,t}\ldots y_{M+1,t}}\prod_{r=1}^{M+1}\left(\frac{y_{r,t}}{q_t}\right)^{y_{r,t}}$$
$$= \binom{q_t}{y_t(r')} p_t(r')^{y_t(r')}(1-p_t(r'))^{q_t-y_t(r')}\binom{q_t-y_t(r')}{y_{1,t}\ldots y_{r'-1,t}}$$
$$\prod_{r=1}^{r'-1}\left(\frac{p_{r,t}}{1-p_t(r')}\right)^{y_{r,t}}\binom{y_t(r')}{y_{r',t}\ldots y_{M+1,t}}\prod_{r=r'}^{M+1}\left(\frac{p_{r,t}}{p_t(r')}\right)^{y_{r,t}}.$$

Furthermore, to enforce the UVSDT restriction, the maximum-likelihood estimates under the saturated model $\hat{p}_{r,t} = \frac{y_{r,t}}{q_t}$ need to be modified if $r_1 \leq r_2$. To enforce the restriction, it is sufficient to enforce $\hat{p}_o(r_2) \geq \frac{1}{2}$ or $\hat{p}_n(r_1) \leq \frac{1}{2}$, because p_o and p_r are non-increasing in r. In terms of the just-mentioned binomial distribution (a) with $r' = r_1$ or $r' = r_2$, the supremal likelihood if either of these restrictions is enforced is the one with parameter $\hat{p} = \frac{1}{2}$ (remember that the frequency counts violate the restriction) and it is sufficient to enforce one of them, leading to the two choices for l above. The parameters of the two multinomial distributions (b) are chosen so that their likelihood is simply maximized by using the appropriate relative frequencies.

To summarize, we first generate frequencies from a density proportional to the maximum likelihood of the saturated model using the Gibbs sampler discussed in the body of the paper. The set of frequencies $\boldsymbol{y} = (y_{r,t})_{r=1,\ldots,M+1,t=n,o}$ is accepted with probability given by $\frac{g}{l}$ in a second step of rejection sampling. The finally accepted frequencies are thereby sampled from a model that respects the UVSDT restriction. The resulting sampling density is more similar to the integrands defined by the more restrictive models considered here than the maximum likelihood of the saturated model, and it led to a sizeable overall acceleration of the basic algorithm described in the body of the text.

Estimating the normalizing constant

The NML algorithm as described so far estimates NML up to an additive constant Q (remember that we defined NML on a logarithmic scale). The additive constant that needs to be removed is minus the logarithm of the integral of the maximum-likelihood function of the model with UVSDT restrictions characterized in the previous section.

We estimate this integral in two steps corresponding to the two-step sampling procedure for sampling from the model with UVSDT restriction. First, we approximate the normalizing constant X for the maximum-likelihood function of the saturated model, g, that is the inverse of the sum of g over all possible frequency patterns so that $X\sum_{\boldsymbol{y}} g(\boldsymbol{y}) = 1$. To do so, consider the probability function that assigns each frequency pattern $\boldsymbol{y} = (y_{r,t})_{r=1,\ldots,M+1,t=n,o}$ the equal probability $f(\boldsymbol{y}) = \binom{q_o+M}{M}^{-1}\binom{q_n+M}{M}^{-1}$ and note that $\sum_{\boldsymbol{y}} f(\boldsymbol{y}) = 1$. Furthermore it is not difficult to see that g dominates f. It follows as an instance of so-called inverse importance sampling (Evans & Swartz, 2000, Chapter 7.4) that the desired constant X is approximated by $T_m = \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{y_i})/g(\boldsymbol{y_i})$, where $\boldsymbol{y_i}$ are sampled from the probability function proportional to g via the (non-refined) Gibbs sampler.

Monitoring the acceptance rates in the second rejection sampling step of the algorithm allows one to estimate the normalizing constant for the probability function of the model with UVSDT restriction. This uses the following result (e.g., Evans & Swartz, 2000, Chapter 3). If p is a proposal density and d is proportional to a density from which we wish to sample with $d \leq p$, then the probability p_a of accepting samples from p in rejection sampling based on d equals $\int d$. Set $p = Xg$ and let h be the maximum likelihood under the model with UVSDT restriction from the previous section. It follows (because $h \leq g$ and hence, $Xh \leq Xg$) that $p_a = X\int h$. Hence, the desired constant Q is given by $\log(X) - \log(p_a)$.

Approximation error

Consider first the approximation of NML via $T_m = \frac{1}{m}\sum_i f(\boldsymbol{y_i} \mid \hat{\theta}(\boldsymbol{y_i}))/h(\boldsymbol{y_i})$, where h is the maximum-likelihood of the model with the UVSDT restrictions discussed above, the $\boldsymbol{y_i}$ are patterns of frequency counts sampled from that model, and f is the model function for one of the candidate models of recognition memory. As m becomes large, T_m approximates T where $\log(T)$ is the NML penalty up to an additive constant Q. If s_m is the standard deviation of a set of m independent, identically distributed sampled $f(\boldsymbol{y_i} \mid \hat{\theta}(\boldsymbol{y_i}))/h(\boldsymbol{y_i})$, then an asymptotically valid $(1-\alpha)$ confidence interval for T is given by

$$\left(T_m - z_{(1-\frac{\alpha}{2})}\frac{s_m}{\sqrt{m}}, T_m + z_{(1-\frac{\alpha}{2})}\frac{s_m}{\sqrt{m}}\right).$$

This means that for large m, the true value T is contained in the confidence interval with half-lengths $\pm 3\frac{s_m}{\sqrt{m}}$ with virtual certainty. Monte Carlo methods yield, following convergence of the chain, identically, but not independently distributed samples $f(\boldsymbol{y_i} \mid \hat{\theta}(\boldsymbol{y_i}))/h(\boldsymbol{y_i})$. There are different methods of estimating the quantity s_m taking this into account, and we used the method known as batching (Evans & Swartz, 2000, Chapter 7.5.2). For this purpose, we divide the sequence of m values $f(\boldsymbol{y_i} \mid \hat{\theta}(\boldsymbol{y_i}))/h(\boldsymbol{y_i})$ into nonoverlapping sequential batches of size $l = 100 \times n_{\text{threads}}$, where n_{threads} is the number of independent threads that we generate (between 8 and 32, depending on the machine on which we were working) for values of m that are multiples of the batch size. The estimate of s_m is $\sqrt{l}$ times the estimated standard deviation of the batchwise means around the grand mean of the entire sequence. Given the estimate of the asymptotic standard error $\frac{s_m}{\sqrt{m}}$ of T, the asymptotic standard error of $\log(T)$ is estimated by $\frac{1}{T_m}\frac{s_m}{\sqrt{m}}$ (Rao, 1973, Chapter 6a).

The constant Q is the sum of two parts: The (logarithm) of the normalizing constant X for g and (minus) the logarithm of the acceptance rate of the rejection step of the sampling scheme for the model with the UVSDT restrictions. An asymptotic estimate of the standard error for the normalizing constant X for g (and of its logarithm) can be computed as for T, given that it was approximated analogously. This is also true for standard errors for p_a and $\log(p_a)$, given that we also computed p_a analogously; that is by sampling from the distribution with density proportional to the maximum likelihood g of the saturated model and estimating p_a as $\frac{1}{m}\sum_i(1_{\boldsymbol{y_i}}g(\boldsymbol{y_i}))/g(\boldsymbol{y_i})$ with $1_{\boldsymbol{y_i}}$ equal one if $\boldsymbol{y_i}$ was accepted and zero otherwise (more precisely, we sampled pairs $(u_i, \boldsymbol{y_i})$, where u_i are samples from a uniform distribution on $(0, 1)$ independent of each other and the $\boldsymbol{y_j}$ and we defined $1_{\boldsymbol{y_i}}$ as a function of $(u_i, \boldsymbol{y_i})$ to equal one, if $u_i \leq h(\boldsymbol{y_i})/g(\boldsymbol{y_i})$ and to equal zero otherwise).

The algorithm began with a phase in which X was estimated to high precision, sampling from the density Xg via the Gibbs sampler until the asymptotic standard error s_1 of $\log(X)$ was smaller than a preselected constant much smaller than 0.1. In a second phase, we continued sampling from the density Xg via the Gibbs sampler estimating p_a this time until the asymptotic standard error s_2 of $\log(p_a)$ was smaller than a preselected constant much smaller than 0.1. In a subsequent third phase, we sampled from the density proportional to the maximum likelihood of the model with UVSDT restriction, monitoring the standard error s_3 of the estimate of $\log(T)$ obtained thereby. Because the samples on which the estimates of s_1, s_2, and s_3 are based are independent of each other,[7] an estimate of the total standard error for estimating $\text{NML} = \log(T) - \log(X) + \log(p_a)$ was $s_E = \sqrt{s_1^2 + s_2^2 + s_3^2}$. Sampling in the third phase proceeded until the breadth of the asymptotic confidence interval for NML with $z_{(1-\frac{\alpha}{2})} = 3$ was smaller than 0.1 (i.e. until $6 \times s_E < 0.1$). This ensures that the NML values are estimated with an accuracy of at least one decimal place.

Numerical issues

The same non-trivial numerical issues require careful attention in the NML computation as in Klauer and Kellen (2011a). These are (a) round-off error in sums over many terms, (b) the problem of extreme parameter values, and (c) local minima in maximum-likelihood estimation. Basically the same remarks apply regarding these issues as stated in Appendix B of Klauer and Kellen (2011a).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmp.2015.05.002.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation*, *9*, 349–368. http://dx.doi.org/10.1162/neco.1997.9.2.349.

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). *Psychological Bulletin*, *139*, 1204–1212. http://dx.doi.org/10.1037/a0033894.

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. http://dx.doi.org/10.1037/0278-7393.22.1.197.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1601–1608. http://dx.doi.org/10.1037/a0031849.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge, MA: MIT Press.

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America*, *53*, 129–160. http://dx.doi.org/10.1364/JOSA.53.000129.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665–678. http://dx.doi.org/10.1037/a0028111.

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, *21*, 916–944. http://dx.doi.org/10.1080/09658211.2013.767348.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 587–606. http://dx.doi.org/10.1037/a0015279.

Burnham, K. P., & Anderson, D. R. (2005). *Model selection and multimodel inference* (2nd ed.). New York: Springer.

Chechile, R. A. (1998). A new method for estimating model parameters for multinomial data. *Journal of Mathematical Psychology*, *42*, 432–471. http://dx.doi.org/10.1006/jmps.1998.1210.

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review*, *15*, 906–926. http://dx.doi.org/10.3758/PBR.15.5.906.

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. http://dx.doi.org/10.1037/0033-295X.109.4.710.

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. http://dx.doi.org/10.1037/a0024957.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *117*, 831–863. http://dx.doi.org/10.1037/a0019634.

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. http://dx.doi.org/10.1037/a0019634.

Dubé, C., Rotello, C. M., & Pazzaglia, A. M. (2013). The statistical accuracy and theoretical status of discrete-state MPT models: Reply to Batchelder and Alexander (2013). *Psychological Bulletin*, *139*, 1213–1220. http://dx.doi.org/10.1037/a0034453.

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General*, *127*, 83–96. http://dx.doi.org/10.1037/0096-3445.127.1.83.

Evans, M., & Swartz, T. (2000). *Approximating integrals via Monte Carlo and deterministic methods*. New York: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, B. D. (2004). *Bayesian data analyses* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

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

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

Heathcote, A., Ditton, E., & Mitchell, K. (2006). Word frequency and word likeness mirror effects in episodic recognition memory. *Memory & Cognition*, *34*, 826–838. http://dx.doi.org/10.3758/bf03193430.

Jaeger, A., Cox, J. C., & Dobbins, I. G. (2012). Recognition confidence under violated and confirmed memory expectations. *Journal of Experimental Psychology: General*, *141*, 282–301. http://dx.doi.org/10.1037/a0025687.

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. http://dx.doi.org/10.1037/a0015525.

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. http://dx.doi.org/10.3758/s13423-011-0096-7.

Karabatsos, G., & Walker, S. G. (2006). On the normalized maximum likelihood and Bayesian decision theory. *Journal of Mathematical Psychology*, *50*, 517–520. http://dx.doi.org/10.1016/j.jmp.2006.07.005.

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology*, *55*, 251–266. http://dx.doi.org/10.1016/j.jmp.2010.11.004.

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1795–1804. http://dx.doi.org/10.1037/xlm0000016.

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

Kellen, D., Singmann, H., & Klauer, K. C. (2014). Modeling source-memory overdistribution. *Journal of Memory and Language*, *76*, 216–236. http://dx.doi.org/10.1016/j.jml.2014.07.001.

Kellen, D., Singmann, H., Vogt, J., & Klauer, K. C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology*, *62*, 40–53. http://dx.doi.org/10.1027/1618-3169/a000272.

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*, 465–478. http://dx.doi.org/10.3758/PBR.17.4.465.

Klauer, K. C., & Kellen, D. (2011a). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, *55*, 430–450. http://dx.doi.org/10.1016/j.jmp.2011.09.002.

Klauer, K. C., & Kellen, D. (2011b). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (2010). *Psychological Review*, *118*, 164–173. http://dx.doi.org/10.1037/a0020698.

Klauer, K. C., Singmann, H., & Kellen, D. (2015). Parametric order constraints in multinomial processing tree models: An extension of Knapp and Batchelder (2004). *Journal of Mathematical Psychology*, 64–65, 1–7.

Koen, J. D., Aly, M., Wang, W. C., & Yonelinas, A. P. (2013). Examining the causes of memory strength variability: Recollection, attention failure, or encoding variability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1726–1741. http://dx.doi.org/10.1037/a0033671.

[7] The impact of possible autocorrelations between the last samples from one phase and the first samples of a subsequent phase is swamped by the tens of thousands of samples obtained in each phase.

Koen, J. D., & Yonelinas, A. P. (2010). Memory variability is due to the contribution of recollection and familiarity, not to encoding variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1536–1542. http://dx.doi.org/10.1037/a0020448.

Koen, J. D., & Yonelinas, A. P. (2011). From humans to rats and back again: Bridging the divide between human and animal studies of recognition memory with receiver operating characteristics. *Learning & Memory*, *18*, 519–522. http://dx.doi.org/10.1101/lm.2214511.

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

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, http://dx.doi.org/10.1016/j.jmp.2008.03.002. 362–275.

Luce, R. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*, 61–79. http://dx.doi.org/10.1037/h0039723.

Macho, S. (2004). Modeling associative recognition: A comparison of two-high-threshold, two-high-threshold signal detection, and mixture distribution models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 83–97. http://dx.doi.org/10.1037/0278-7393.30.1.83.

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

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

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

Moshagen, M., & Hilbig, B. E. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, *21*, 1431–1443. http://dx.doi.org/10.3758/s13423-014-0643-0.

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

Myung, J. I., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11170–11175. http://dx.doi.org/10.1073/pnas.170283897.

Myung, J. I., Forster, M., & Brown, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2. http://dx.doi.org/10.1006/jmps.1999.1273.

Myung, J. I., Navarro, D., & Pitt, M. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179. http://dx.doi.org/10.1016/j.jmp.2005.06.008.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499–518. http://dx.doi.org/10.1037/a0016104.

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*, *14*, 1043–1050. http://dx.doi.org/10.3758/BF03193089.

Navarro, D. J., & Lee, M. D. (2005). An application of minimum description length clustering to partitioning learning curves. In *Proceedings of the 2005 IEEE international symposium on information theory* (pp. 587–591). Piscataway, NJ: IEEE.

Navarro, D. J., Pitt, M. A., & Myung, J. I. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84. http://dx.doi.org/10.1016/j.cogpsych.2003.11.001.

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. http://dx.doi.org/10.1037/a0018926.

Pazzaglia, A. M., Dubé, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*, http://dx.doi.org/10.1037/a0033044.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 224–232. http://dx.doi.org/10.1037/a0017682.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*, 227–259. http://dx.doi.org/10.1207/s15327906mbr4103_1.

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

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley, http://dx.doi.org/10.1002/9780470316436.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535. http://dx.doi.org/10.1037/0033-295X.99.3.518.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1446–1465. http://dx.doi.org/10.1037/a0013646.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47. http://dx.doi.org/10.1109/18.481776.

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. http://dx.doi.org/10.1109/18.930912.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367. http://dx.doi.org/10.1037/0033-295X.107.2.358.

Roos, T. (2008). Monte Carlo estimation of minimax regret with an application to MDL model selection. In *IEEE Information Theory Workshop* (pp. 284–288). http://dx.doi.org/10.1109/ITW.2008.4578670.

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. http://dx.doi.org/10.3758/BF03196750.

Schütz, J., & Bröder, A. (2011). Signal detection and threshold models of source memory. *Experimental Psychology*, *58*, 293–311. http://dx.doi.org/10.1027/1618-3169/a000097.

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

Shiffrin, R.M. (2014). Moving past BMS and MDL: Making model evaluation rational. In *Paper presented at the 47th annual meeting of the society for mathematical psychology*.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models with R. *Behavior Research Methods*, *45*, 560–575. http://dx.doi.org/10.3758/s13428-012-0259-0.

Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 615–625. http://dx.doi.org/10.1037/0278-7393.30.3.615.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50. http://dx.doi.org/10.1037/0096-3445.117.1.34.

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, MA: MIT Press.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340. http://dx.doi.org/10.1037/h0040547.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1–12. http://dx.doi.org/10.1037/h0080017.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Townsend, J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford, UK: Oxford 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. http://dx.doi.org/10.1037/0278-7393.26.3.582.

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

Wagenmakers, E. J., & Waldorf, L. (2006). Editors' introduction. *Journal of Mathematical Psychology*, *50*, 99–100. http://dx.doi.org/10.1016/j.jmp.2005.01.005.

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

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. http://dx.doi.org/10.3758/PBR.17.3.275.

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. http://dx.doi.org/10.1016/j.jmp.2010.02.001.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*, 747–763. http://dx.doi.org/10.3758/BF03211318.

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

Zhang, J. (2011). Model selection with informative normalized maximal likelihood: Data prior and model prior. In E. Dzhafarov, & L. Perry (Eds.), *Descriptive and normative approaches to human behavior* (pp. 303–319). New Jersey: World Scientific.