THEORETICAL AND REVIEW ARTICLES

Minimum description length model selection of multinomial processing tree models

HAO WU AND JAY I. MYUNG
Ohio State University, Columbus, Ohio

AND

WILLIAM H. BATCHELDER
University of California, Irvine, California

Multinomial processing tree (MPT) modeling has been widely and successfully applied as a statistical methodology for measuring hypothesized latent cognitive processes in selected experimental paradigms. In this article, we address the problem of selecting the best MPT model from a set of scientifically plausible MPT models, given observed data. We introduce a minimum description length (MDL) based model-selection approach that overcomes the limitations of existing methods such as the G^2-based likelihood ratio test, the Akaike information criterion, and the Bayesian information criterion. To help ease the computational burden of implementing MDL, we provide a computer program in MATLAB that performs MDL-based model selection for any MPT model, with or without inequality constraints. Finally, we discuss applications of the MDL approach to well-studied MPT models with real data sets collected in two different experimental paradigms: source monitoring and pair clustering. The aforementioned MATLAB program may be downloaded from http://pbr.psychonomic-journals.org/content/supplemental.

Multinomial processing tree (MPT) modeling is a statistical methodology for measuring latent cognitive capacities in selected experimental paradigms (Batchelder & Riefer, 1986, 1990, 1999; Chechile, 2004; Erdfelder et al., 2009; Hu & Batchelder, 1994; Riefer & Batchelder, 1988, 1991, 1995; Riefer, Hu, & Batchelder, 1994). The data structure requires that participants performing a cognitive task make categorical responses to a series of test items. An MPT model parameterizes a subset of probability distributions over the response categories by specifying a processing tree designed to represent hypothesized cognitive steps, such as memory encoding, storage, discrimination, inference, guessing, and retrieval.

Since its introduction in the 1980s, MPT models have been successfully applied to modeling performance in a wide range of cognitive tasks, including associative recall, source monitoring, eyewitness memory, hindsight bias, object perception, speech perception, propositional reasoning, social networks, and cultural consensus. Batchelder and Riefer (1999) listed over 80 applications of MPT models in various areas of cognitive and social psychology. MPT models have also been applied to estimate cognitive deficits in special populations (Batchelder & Riefer, 2007; Chechile, 2007; see Erdfelder et al., 2009, for a review of such applications). The use of MPT

models to assess special populations is often referred to as *cognitive psychometrics*, representing the fact that theoretically motivated models are employed as measurement tools of cognitive functioning (Batchelder, 2009; Batchelder & Riefer, 2007; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002). In all of these applications, MPT models were intended to offer researchers more instructive and informative interpretations of data than those based on the traditional data analytic approaches, such as the ANOVA.

In the present study, we are concerned with the logic of selecting the best MPT model from a set of scientifically plausible MPT models that are available to account for a given data set. A researcher may entertain multiple scientific hypotheses about the underlying processes, each formulated as a distinct MPT model,[1] and may wish to determine which one of these models best describes the observed data in some defined sense; this is the problem of model selection (Myung & Pitt, 1997). By selecting among theoretically motivated models, the researcher is able to identify from alternative theories the one best supported by empirical observations. To illustrate, consider the question of how the different languages of bilingual people are cognitively represented. Several theories addressing this issue differ as to whether information pre-

H. Wu, wu.498@osu.edu

sented in a particular language retains a language-specific tag. Source-monitoring experiments were conducted to differentiate these theories (e.g., Rose, Rose, King, & Perez, 1975; Saegert, Hamayan, & Ahmar, 1975), and these theories, as represented by their corresponding source-monitoring MPT models (to be elaborated on in the next section) that assume different treatment effects on their parameters, can be compared by model selection (Batchelder & Riefer, 1990). Similarly, other theoretical issues in cognitive psychology, such as sequential versus nonsequential processes and automatic versus control processes, can also be addressed by comparing MPT models with different tree structures (e.g., Bishara & Payne, 2008; Schweickert, 1993).

In addition to evaluating multiple scientific theories behind different MPT models, model selection can also be employed as a tool for examining the *validity* of an MPT model. The validity of an MPT model concerns whether it is warranted to interpret a parameter in the model as representing the underlying cognitive process that it is explicitly postulated to represent (see, e.g., Batchelder & Riefer, 1999; Riefer et al., 2002; Schweickert & Chen, 2008). To establish validity, it is necessary to apply experimental treatments that have predictable selective influence on the parameters. For example, if a model has a parameter θ that is postulated to measure the ability to retrieve items from memory, then experimental manipulations that should affect levels of retrievability should result in predictable changes in θ but no change in parameters postulated to measure other things. To determine whether the desired selected influence is present for a particular MPT model, it is necessary to select from among different versions of the model assuming different patterns of treatment effects.

Because of its importance in evaluating scientific theories and establishing the validity of MPT models, model selection is of particular interest in MPT modeling. To perform model selection, one must account for the effect of model complexity. This is because model complexity can affect the predictive capacity or accuracy of a model, which is the hallmark of model selection (Myung, 2000; Myung & Pitt, 1997). In the case of MPT models, it has been shown that they can vary greatly in complexity not only in the number of parameters but also, importantly, in the functional form of the models, such as tree structure and parameter constraints (Wu, Myung, & Batchelder, 2010). However, as will be discussed later in this article, the likelihood ratio test (LRT; Read & Cressie, 1988), currently in wide use for MPT modeling, does not select models on the basis of their predictive accuracy. Other popular selection methods, such as the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwartz, 1978) do not fully account for all dimensions of model complexity. Given these limitations, a model-selection method that fully accounts for model complexity is called for.

In the present article, we introduce such a method for MPT modeling: minimum description length (MDL) model selection. MDL has been successfully applied to addressing various model-selection problems in cognitive

modeling (e.g., Lee, 2001; Lee & Pope, 2006; Myung, Pitt, & Navarro, 2007; Navarro & Lee, 2004; Pitt, Myung, & Zhang, 2002) but is entirely absent in MPT modeling, with the exception of our own work (Wu et al., 2010). To help researchers not familiar with numerical computing, in this article we make available a general-purpose computer program that implements MDL-based model selection for virtually all types of MPT models.

The rest of the article is organized as follows: We first begin with a formal definition of MPT models. We then briefly review the extant methods of model selection, such as LRT, AIC, and BIC, before introducing MDL, the focus of the present work. The discussion then turns to the computer program, and we provide detailed instruction of how to use it in a given situation of MPT modeling. Finally, two application examples of MDL-based model selection with real data sets are presented before we conclude the article.

MPT Models

MPT models assume that the observed categorical responses in an experiment follow from a series of latent cognitive events. These events are represented by a tree structure, in which nonterminating nodes represent the events; branches that follow from a node represent all possible outcomes of the event, with the probabilities of these outcomes being either parameters in the model or known constants; and leaves (terminating nodes) of the tree structure represent the observed responses from participants. Because different sequences of events may lead to the same response, a response category may include more than one leaf in the tree.

To illustrate how an MPT model works, consider the one-high-threshold model (1HTM) for source-monitoring experiments as it is depicted in Figure 1. In a source-monitoring experiment, participants first study a list of items from two sources, A and B, and are then asked to judge the source of test items as *from A*, *from B*, or *new* (N; i.e., a new item from neither source). The 1HTM for such experiments consists of three distinct trees (Batchelder & Riefer, 1990), each modeling hypothetical processes assumed to be involved in responding to a given type of items. A distinguishing feature of this model is that it assumes that old items can be correctly detected with probabilities D_1 and D_2 for items from Sources A and B, respectively. If an old item is correctly detected as *old*, a discriminating decision on its source is made, with success probabilities d_1 and d_2 for the two sources, respectively. If any of the two processes fails, guessing processes follow. For new items, however, the model assumes no detection process, and instead, response selection is governed by guessing processes only. The model postulates three types of guessing processes, represented by parameters b, g, and a (see Figure 1 for details). By putting various constraints on the model parameters, a hierarchy of submodels can be derived from the model, which is shown in Figure 2. For instance, the equality constraints of $D_1 = D_2$ and $d_1 = d_2$, which amount to saying that the detection and discrimination probabilities both stay the same across items from different sources, result in 1HTM 5A. On the
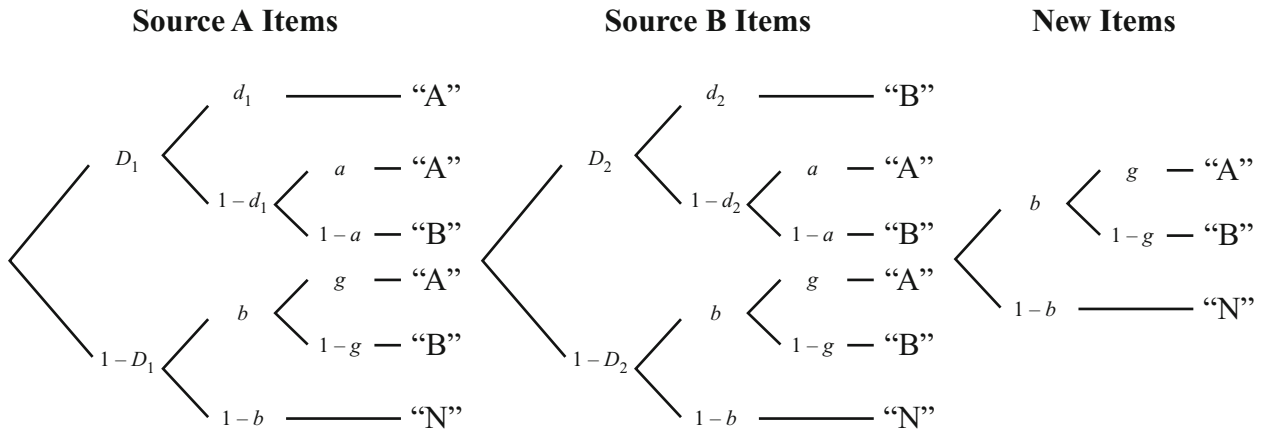
Source A Items   **Source B Items**   **New Items**

Figure 1. The one-high-threshold multinomial processing tree model of source monitoring. The parameters are defined as follows: D_1, detectability of Source A items; D_2, detectability of Source B items; d_1, source discriminability of Source A items; d_2, source discriminability of Source B items; a, guessing that a detected but nondiscriminated item belongs to Source A; b, guessing *old* to a nondetected item; g, guessing that a nondetected item biased as *old* belongs to a Source A category. From "Multinomial Processing Models of Source Monitoring," by W. H. Batchelder & D. M. Riefer, 1990, *Psychological Review, 97*, p. 551. Copyright 1990 by the American Psychological Association. Adapted with permission.

other hand, if we assume that only the source discrimination probabilities are the same for both sources ($d_1 = d_2$), not the detection probabilities ($D_1 \neq D_2$), 1HTM 6B, in which 1HTM 5A is nested, is obtained instead.

Speaking in formal terms, an MPT model parameterizes a subset of multinomial probability distributions over response categories. Because every MPT model can be reparameterized into a binary MPT (BMPT) model, in which every decision node has only two processing possibilities (Hu & Batchelder, 1994), we will only discuss the mathematical formulation of BMPT models. Suppose a BMPT model has S parameters [$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_S)'$] and J categories ($C_1, C_2, \ldots, C_J$), and category C_j includes leaves B_{ij} ($i = 1, 2, \ldots, I_j; j = 1, 2, \ldots, J$). Because of its binary nature, nonconstant probabilities on the branches must be of the form θ_s or $1 - \theta_s$. The probability of taking the decision path to a leaf B_{ij} is given by the product of all probabilities along this path:

$$p_{ij}(\boldsymbol{\theta}) = c_{ij} \prod_{s=1}^{S} \theta_s^{a_{ijs}} \left(1 - \theta_s\right)^{b_{ijs}}, \qquad (1)$$

where a_{ijs} and b_{ijs} are, respectively, the number of times θ_s and $1 - \theta_s$ appear on the path to B_{ij}, and c_{ij} is the product of all constant probabilities along the same path or set to unity if there is no constant probability along this path. The probability of category C_j is the sum of the probabilities of all leaves it includes; that is,

$$p_j(\boldsymbol{\theta}) = \sum_{i=1}^{I_j} p_{ij}(\boldsymbol{\theta}). \qquad (2)$$

For example, each tree in 1HTM discussed above is a BMPT model. The probability for a subject to respond *Source A* given a stimulus from Source A is given by $D_1 d_1 + D_1(1 - d_1)a + (1 - D_1)bg$.

Now let us assume that several participants make categorical responses to the same set of items and that their responses are independently and identically distributed into

the J categories of a model. Let n_j be the number of these responses that fall into category C_j, $\mathbf{n} = (n_1, n_2, \ldots, n_J)'$, and $N = \sum_j n_j$. Then $\mathbf{n}$ is distributed as a multinomial probability distribution given by

$$f(\mathbf{n} \mid \boldsymbol{\theta}) = \binom{N}{n_1, \ldots, n_J} \prod_{j=1}^{J} p_j^{n_j}(\boldsymbol{\theta}), \qquad (3)$$

where the multinomial probability p_j follows the computational rules in Equations 2 and 1.

The above mathematical description of BMPT models with constants a_{ijs}, b_{ijs}, and c_{ij}, although it uniquely and sufficiently specifies the distribution of the data, can be cumbersome as an input to computer programs. For this purpose, Purdy and Batchelder (2009) devised a much more concise and elegant representation of BMPT models. Their string representation scheme exploits the recursive properties of the tree structure and includes only branching probabilities and categories in the model. To illustrate, the string representation of a coin-flipping Bernoulli model is given by *pHT*, where *H* and *T* are outcomes of the process and *p* is the probability of obtaining the outcome *H*. To obtain the string representation for a more complex BMPT model, one begins with a representation of the decision process at the root node and then replaces the two outcomes with the representations of the decision processes that follow those outcomes. To illustrate, take the tree of Source A items in the 1HTM in Figure 1. We first represent the item-detection process with D_1(detected)(undetected). We then replace the outcome *detected* with the representation of the discrimination process d_1A(source unidentified) and the outcome *undetected* with that of the guessing process b(guess as old)N. Now we get $D_1 d_1$A(source unidentified)b(guess as old)N. We continue the replacement until the string contains only branching probabilities and response categories. The string representation of the tree is $D_1 d_1$AaABbgABN. This representation makes the input to computer programs
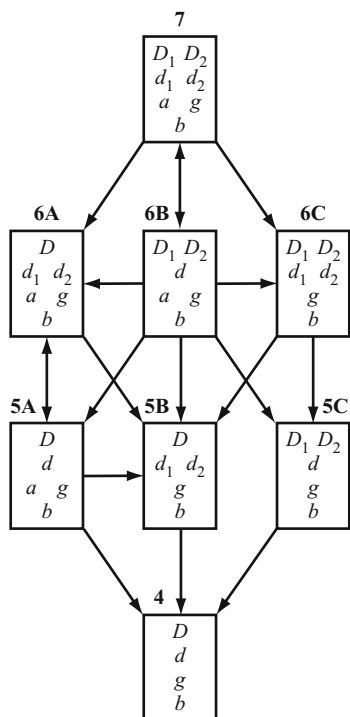
Figure 2. The hierarchy of eight versions of the 1HTM in Figure 1, created by imposing successive constraints on the parameters. In this figure, parameters in each model are listed and a directed arrow from one model to another indicates that the second model is nested in the first. Note that this figure combines the nesting relationships shown in Figures 2 and 3 of Batchelder and Riefer (1990), and some of the nesting relationships are not explicit from the parameter constraints.

much easier and will be exploited in our MATLAB program described later in the present article.

Methods of Model Selection

As was mentioned in the introduction, model selection is a necessary and crucial step in the application of MPT models. Various model-selection methods have been proposed in the past for this purpose. In the following, we review some of these methods, including the MDL method.

LRT. The G^2-based LRT is the most commonly used method of inference in MPT modeling (e.g., Hu & Batchelder, 1994; Hu & Phillips, 1999; Riefer & Batchelder, 1988). At the center of this approach is the likelihood-ratio-based test statistic by which the adequacy of a model is evaluated in the null hypothesis significance testing framework. Specifically, LRT requires the setting of two models, full and reduced, such that the reduced model is nested in the full model with a reduction in the number of parameters. The G^2 test statistic is then defined as $G^2 = -2 \ln LR$, where LR is the ratio of the maximum likelihood of the reduced model to that of the full model. Under the null hypothesis that the reduced model is correct, when the sample size N is large enough, the sampling distribution of G^2 is shown to follow a χ^2 distribution, with

the degrees of freedom equal to the difference in numbers of parameters between the models, provided that certain regularity conditions are satisfied (e.g., Read & Cressie, 1988). If the value of G^2 is large enough to fall in the rejection region of the sampling distribution, the null hypothesis is rejected, and the full model is chosen. Otherwise, the reduced model is chosen over the full model.

The G^2-based LRT is generally a useful method of model evaluation but has several limitations in its use as a model-selection method for MPT models. First, the method can only be used for comparing pairs of nested models, one pair at a time. This effectively excludes its application to the situation in which multiple models with or without nesting relationships are compared. Second, the regularity conditions of the test require that the maximum likelihood estimate (MLE) under either model should not be on the boundary of the parameter space (see Shapiro, 1988, for an alternative procedure). This implies that LRT is not able to take into account inequality constraints in the models. To see this, because the inclusion of inequality constraints does not change the degrees of freedom of the LRT, it changes the result of the test only when parameters of either model are estimated on the boundary defined by those constraints, but that would violate the regularity conditions mentioned above and would render the test invalid. For the same reason, LRT cannot be employed to compare two nested models with the same number of parameters but different functional forms, such as 1HTMs 6A and 6B in Figure 2.

Besides the above issues, note that the goal of LRT is to assess the *descriptive adequacy* of a given null model in the null hypothesis significance test framework, not to choose from a set of candidate models the one that best captures the regularities underlying the data (Myung & Pitt, 1997). As such, LRT does not necessarily help identify the best-approximating model, which is what model selection is about. This latter criterion is known as *generalizability* in statistics (e.g., Myung, 2000; Myung & Pitt, 1997). In the rest of this section, we discuss various model-selection criteria proposed as generalizability measures and the importance of model complexity in determining a model's generalizability.

Generalizability and model complexity. The *generalizability* of a model refers to how well the conclusion from the current observed data can be applied to future, not yet observed data (Myung, 2000). By definition, the model with best generalizability gives the closest approximation to the underlying mechanism of the data and should therefore be preferred in model selection. Models that generalize well should first provide a good fit to the present data; however, generalizability is more than goodness of fit and is significantly affected by model complexity.

Model complexity or flexibility has to do with a model's intrinsic capability to fit a wide range of data patterns. Generally speaking, a model with many parameters is more complex than a model with fewer parameters. Furthermore, models with the same number of parameters but different equation forms can also differ in complexity. This is called the *functional form* dimension of model complex-

ity (Myung & Pitt, 1997). To give an example, two psychophysics models, $y = ax^b + \varepsilon$ and $y = a \log(x + b) + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2)$, may differ in complexity, despite the fact that they both have two parameters. Because of their flexibility, complex models tend to overfit the current data, thereby generalizing poorly to future observations, and should therefore be penalized in model selection.

In the case of MPT models, differences in complexity due to functional form can arise in a variety of ways: from different tree structures, different parametric constraints, and/or different category assignments to the leaves of a tree. For example, consider 1HTMs 5A, 5B, and 5C, shown in Figure 2, each of which imposes a distinct set of equality constraints on the parameters of the largest model, 1HTM 7. Although all three models have five parameters, their complexity may be quite different from one another. This is in fact what Wu et al. (2010) found. Their results showed that the difference in complexity between 1HTMs 5A and 5B is larger than that between 1HTMs 5B and 4. An implication is that the complexity difference due to functional form of MPT models can be even greater than that due to the number of parameters. Findings such as these point to the importance of accounting for the functional form dimension of complexity—in particular, in model selection with MPT models.

MDL. Various model-selection methods that estimate a model's generalizability have been proposed in statistics.[2] Among them, AIC (Akaike, 1973) and BIC (Schwartz, 1978) have been used in MPT modeling, although MDL (Grünwald, 2007; Rissanen, 1996, 2001) has not, to our knowledge. In what follows, we briefly review AIC and BIC before turning our discussion to MDL.

Unlike LRT, AIC (Akaike, 1973) and BIC (Raftery, 1999; Schwartz, 1978; Wagenmakers, 2007; Weakliem, 1999) can be used to compare multiple nested or nonnested models. They are defined as

$$\text{AIC} = -2\text{LML} + 2S, \qquad (4)$$
$$\text{BIC} = -2\text{LML} + S \ln N, \qquad (5)$$

where LML$\{= \ln f[\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})]\}$ denotes the natural logarithm of the maximized likelihood, $\mathbf{x}$ is the current data, S is the number of parameters, and N is the sample size. For both model-selection criteria, among a set of competing models, the model with the smallest criterion value is judged to best generalize and is thus preferred. We can see that in both equations, the first term is related to the fit of the model, while the second term represents a complexity penalty, thereby formalizing Occam's razor (Myung & Pitt, 1997). However, both AIC and BIC penalize complex models only by their number of parameters, neglecting their functional form complexity. Consequently, neither criterion is appropriate for selecting among models with inequality constraints.

In what follows, we discuss MDL-based model selection, which presents itself as an attractive alternative method, because it overcomes all of the aforementioned problems of LRT, AIC, and BIC and is particularly appropriate for MPT model selection, given its ability to fully capture model complexity.

The principle of MDL originates from algorithmic coding theory in computer science. According to this principle, statistical modeling is viewed as data compression, and the best model is the one that compresses the data as tightly as possible. A model's ability to compress the data is measured by the shortest code length with which the data can be coded with the help of the model. The resulting code length is related to generalizability such that the shorter the code length is, the better the model generalizes (Grünwald, 2000, 2007; Grünwald, Myung, & Pitt, 2005; Myung, Navarro, & Pitt, 2006).

The Fisher information approximation (FIA; Rissanen, 1996) represents a formal implementation of the MDL principle for model selection. It gives the shortest code length with which a model can code the data.[3] This criterion is defined as

$$\text{FIA} = -\text{LML} + C_{\text{FIA}}, \qquad (6)$$

with

$$C_{\text{FIA}} = \frac{S}{2} \ln \frac{N}{2\pi} + \ln \int \sqrt{|\mathbf{I}(\boldsymbol{\theta})|} \, d\boldsymbol{\theta}, \qquad (7)$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix (e.g., Casella & Berger, 2001) of sample size 1 with its elements given by

$$I(\boldsymbol{\theta})_{ij} = -E\left[\frac{\partial^2 \ln f(x_1 | \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right].[4]$$

A smaller criterion value indicates better generalization, and thus, the model that minimizes the criterion should be chosen.

There are several observations that one can make about FIA. First, in this selection criterion, generalizability is measured as a trade-off between goodness of fit (LML) and complexity (C_{FIA}). Second, regarding the complexity measure of FIA, C_{FIA}, its first term captures the effects of the number of parameters (S), and its second term captures the functional form effects through the Fisher information matrix $[\mathbf{I}(\boldsymbol{\theta})]$. Note, especially, that the functional form complexity in C_{FIA} is expressed as an integral over the parameter space. As such, it would therefore be straightforward to represent inequality constraints on parameters in C_{FIA}, because the constraints simply reduce the size of the parameter space. Third, FIA is related to BIC in that both criteria can be viewed mathematically as approximations to -2 times the log marginal likelihood in Bayesian statistics when Jeffreys's prior $\pi(\boldsymbol{\theta}) \propto \sqrt{|\mathbf{I}(\boldsymbol{\theta})|}$ is assumed, but FIA gives a better approximation than BIC (Grünwald, 2007, chap. 8).

To summarize, FIA overcomes the practical and theoretical shortcomings of LRT in that the former is based on generalizability and can be applied to multiple nested or nonnested models. Furthermore, the ability of MDL to capture functional form complexity and also to account for the effects of parametric inequality constraints provide MDL with unique advantages over LRT, as well as over AIC and BIC. Given the great variability in functional form complexity among MPT models, MDL is ideally suited for model selection among these models.

MDL Model Selection of MPT Models

A computer program for C_{FIA} computation. Applying the FIA criterion to MPT models requires the computation of C_{FIA} in Equation 7. As no analytic solution is available in general for the integral in the expression of C_{FIA}, the solution must be sought by numerical integration, which may be too cumbersome for most researchers who are interested in applying FIA. To help ease some of the computational burden, we have written a computer program that can be used to compute the complexity measure.

The general-purpose MATLAB program for computing the quantity C_{FIA} for BMPT models is available for download as a supplement from the *Psychonomic Bulletin & Review* Web site (http://pbr.psychonomic-journals.org/content/supplemental). The program evaluates the integral using a Monte Carlo algorithm. The technical details of this numerical integration algorithm are described in Wu et al. (2010). The scope of the program is general enough to compute the complexity of any BMPT model. Given that every MPT model can be reparameterized into an equivalent BMPT model, the program is applicable to all MPT models. The program can also incorporate inequality constraints on parameters insofar as the constrains are of the form $\theta_i < \theta_j$ or its combinations.[5]

The program assumes that the BMPT model has a single tree structure. When K trees are present in one model, these trees should be combined into a single tree with multinomial probabilities $p_k = N_k/N$, where N_k is the sample size for tree k in the experimental design, and N is the total sample size. To illustrate how this is done, consider model 1HTM, shown in Figure 1, and suppose that the sample sizes for items from Sources A and B and new items are 250, 250, and 500, respectively. The three trees should then be joined to form a single tree with numerical probabilities $p_A = p_B = .25$ and $p_N = .5$. Although all three trees in 1HTM are BMPT models, the new tree that we obtained by joining them is not, because there are three branches from the root node. It needs to be turned into a BMPT model through reparameterization. To reparameterize, one first joins the two trees for Sources A and B to a single node with branching probability .5, and then joins the resulting binary tree with the tree for new items with branching probability .5. This is shown in Figure 3.

The MATLAB program involves a function, BMPTFIA, with six input arguments and six output arguments: `function [CFIA,CI,lnInt,CI1,lnconst, CI2] = BMPTFIA(s,parameters,ineq0,category,N,Sample)`. The input and output arguments are described below.

The first input argument, `s`, is related to the string representation of the BMPT model, as was discussed earlier. It can be obtained by replacing all categories in the string by the capital letter "C" and all branching probabilities, including parameters and fixed constants, by the lowercase letter "p." For example, for model 1HTM shown in Figure 1, this argument should be s= "ppppCpCCppCCCppCpCCppCCCppCCC."

The second input argument, `parameters`, is a row vector that assigns parameters or constants to the "p"s in
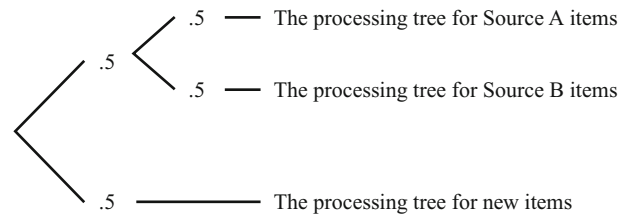

Figure 3. Example of combining the three processing trees in the 1HTM (shown in Figure 1) into one BMPT model. The sample sizes are assumed to be 250, 250, and 500 for Source A items, Source B items, and new items, respectively.

the string `s`. Its length should be the same as the number of "p"s in `s`, and its elements correspond to the "p"s according to their order in `s`. Positive integer elements in `parameters` assign parameters to the corresponding "p"s, with the same integer denoting the same parameter. Constants are assigned to the "p"s using the negation of their values. For Model 5C with multinomial probabilities .25, .25, and .5 for Sources A and B and new items, respectively, this input argument should be parameters= $[-.5, -.5, 1, 3, 5, 4, 5, 2, 3, 5, 4, 5, 4, 5]$. In this vector, the five parameters (D_1, D_2, d, b, g) are coded using the integers 1 to 5, respectively, and the first two elements of the vector (the $-.5$s) are the probability constants that we used to join the three trees into a single tree.

The third input argument, `ineq0`, assigns inequality constraints imposed on the parameters. It is a matrix with two columns. Each element denotes a parameter coded in the same way as in `parameters`. For each row, the parameter on the left column is constrained to be smaller than that on the right column. The number of rows is determined by the total number of simple inequality constraints of the form $\theta_1 < \theta_2$ in the model. For example, if we were to impose an inequality constraint $D_1 > D_2$ for Model 5C, the matrix would be set to ineq0= $[2,1]$. If no inequality constraints are assumed, we set it to an empty matrix, ineq0= $[]$.

The fourth input argument, `category`, assigns categories to the "C"s in the string `s` in the same way in which `parameters` assigns branching probabilities, except that only positive consecutive integers from 1 to J, the total number of categories, are allowed. For model 1HTM, this argument should be set to category = $[1, 1, 2, 1, 2, 3, 5, 4, 5, 4, 5, 6, 7, 8, 9]$. Note that with three different responses in each of the three conditions, there are nine different categories in total.

Finally, the fifth input argument, `N`, specifies the total sample size, and the last input argument, `Sample`, specifies the number of random samples to be drawn in the Monte Carlo algorithm.

The first output argument, `CFIA`, gives C_{FIA}. Given the stochastic nature of the Monte Carlo algorithm, the output value `CFIA` changes from one run of the program to another. The second output argument, `CI`, gives the Monte Carlo confidence interval for `CFIA`. The remaining four output arguments are optional. They are described in detail in the program file.

We now provide an example of the application of the computer program. Consider again 1HTM 5C in Figure 2 with an inequality constraint of $D_1 > D_2$ and sample sizes 250, 250, and 500 for the three kinds of stimuli. In a Monte Carlo run with Sample = 200,000, we obtained CFIA = 12.6182, CI = [12.6113, 12.6251].

The MATLAB program described above gives only the complexity term, C_{FIA}. As is shown in Equation 6, to obtain the value of FIA for a given MPT model, one must also compute the goodness of fit term, $-$LML. This term can be obtained from a user-friendly program called GPT .EXE (Hu & Phillips, 1999), which is available for free download from www.xiangenhu.info. This program performs maximum likelihood estimation and outputs best-fit parameter values and the value of $-$LML for any MPT model with and without inequality constraints.

In what follows, we demonstrate the use of the MATLAB program for model selection of MPT models in two different experimental paradigms: source monitoring and pair clustering.

Modeling source-monitoring data. Our first example concerns the source-monitoring experiment of Rose et al. (1975, Experiment 1). The purpose of their study was to examine whether accurate source memory of language could occur at the semantic level of language processing. In their experiment, participants studied a mixed list of English and Spanish sentences before being tested on recognition and source memory performance. Contextual relationships between the sentences were manipulated in the experiment such that in one condition, the sentences were semantically related to a common topic, whereas in the other condition, all of the sentences were semantically unrelated. Rose et al. reasoned that if language source information is available at the semantic level of processing, because the contextual

relationship among the sentences makes them semantically less distinguishable, language source memory would be poorer for related sentences than for unrelated sentences.

On the basis of an ANOVA of the data, Rose et al. (1975) concluded that there was no significant difference between the two treatment conditions. Batchelder and Riefer (1990, Tables 7 and 8) reanalyzed the same data with model 1HTM 4, shown in Figure 2, which can distinguish the effect of item detection from that of source discrimination. Their LRT results suggested that recognition memory (D) was significantly poorer for related than for unrelated sentences, but there was no significant difference in source monitoring (d). On the basis of this finding, Batchelder and Riefer (1990) concluded that contextual relationships are detrimental to item detection but not to source discrimination, indicating that language source information is not available at the semantic level.

We reanalyze the data from Batchelder and Riefer (1990, Table 7) in a model-selection framework using MDL, as well as AIC and BIC. The results are shown in Table 1. The first model, M_0, has eight parameters, four (D, d, b, g) for each treatment condition, without any equality constraints. A total of 15 additional MPT models were created by imposing various equality constraints on the parameters. For the rest of the models, a subscript notation is used to indicate how the parameters are constrained to be equal across the two treatment conditions, related and unrelated. For example, in M_{Dd}, D and d are set as equal across the two conditions, whereas b and g are allowed to vary across the conditions. In addition to such equality constraints, inequality constraints are also considered for models where either of the two parameters, D or d, is allowed to differ between the two conditions, because theories underlying those models predict a particular direction of difference: Both parameters

Table 1
Summary of Model-Selection Results for Source-Monitoring Data
From Rose et al. (1975, Experiment 1)

Model	S	$-$LML	C_{AIC}	AIC	C_{BIC}	BIC	C_{FIA}	FIA	C'_{FIA}	FIA$'$
M_0	8	36.17	8	44.17	30.24	66.41	22.2	58.4	20.9	57.1
M_g	7	36.61	7	43.61	26.46	63.07	19.7	56.3	18.4	55.0
M_b	7	38.23	7	45.23	26.46	64.69	19.4	57.6	18.0	56.2
M_d	7	36.18	7	43.18	26.46	62.64	20.7	56.9	20.0	56.2
M_D	7	41.16	7	48.16	26.46	67.62	20.4	61.6	19.7	60.9
M_{bg}	6	38.66	6	44.66	22.68	61.34	16.9	55.6	15.5	54.2
M_{dg}	6	36.61	6	**42.61**	22.68	59.29	18.2	54.8	17.5	54.1
M_{Dg}	6	41.60	6	47.60	22.68	64.28	17.8	59.4	17.1	58.7
M_{db}	6	38.30	6	44.30	22.68	60.98	17.8	56.1	17.1	55.4
M_{Db}	6	41.84	6	47.84	22.68	64.52	17.3	59.1	16.6	58.4
M_{Dd}	6	41.55	6	47.55	22.68	64.23	18.5	60.1	18.5	60.1
M_{dbg}	5	38.73	5	43.73	18.90	**57.63**	15.1	**53.8**	14.4	**53.1**
M_{Dbg}	5	42.28	5	47.28	18.90	61.18	14.6	56.9	13.9	56.2
M_{Ddg}	5	41.98	5	46.98	18.90	60.88	15.8	57.8	15.8	57.8
M_{Ddb}	5	42.31	5	47.31	18.90	61.21	15.3	57.6	15.3	57.6
M_{Ddbg}	4	42.74	4	46.74	15.12	57.86	12.5	55.2	12.5	55.2

Note—The data can be found in Batchelder and Riefer (1990, Table 7). The total sample size is $N = 1,920$. See the main text for the description of the models. Inequality constraints apply only to parameters D and d (not to b or g) when the corresponding equality constraints are not present. The FIA$'$ column shows Fisher information approximation (FIA) values if inequality constraints are taken into account, and the FIA column gives FIA values if those constraints are neglected. S, the number of parameters; LML, the natural logarithm of the maximized likelihood; C, complexity; AIC, Akaike information criterion; BIC, Bayesian information criterion. The minimum value for each of the selection criteria is shown in bold.

are expected to be smaller for related sentences than for unrelated sentences. No inequality constraints on the guessing parameters, b and g, are imposed, even if they are allowed to differ across the two conditions. Because parameter estimates do not violate these inequality constraints in the data, incorporating inequality constraints does not change the values of LML, AIC, and BIC. In contrast, inequality constraints do change the value of C_{FIA} and thus the value of FIA. In Table 1, the latter two values obtained under inequality constraints are denoted by C'_{FIA} and FIA′.

From Table 1, we can observe that as the number of parameters decreases, −LML increases, whereas C_{FIA} decreases, as expected, exhibiting a trade-off between goodness of fit and complexity. Of particular interest is the observation that among models with the same number of parameters, C_{FIA} varies greatly, indicating the effects of functional form on complexity. The difference in complexity due to functional form between two models can sometimes be even greater than the difference in goodness of fit. The case in point is the comparison between M_d and M_g. In terms of −LML, M_d fits the data better than M_g (36.18 vs. 36.61) but is more complex (C_{FIA} = 20.7 vs. 19.7), thus yielding an overall larger FIA value than M_g (56.9 vs. 56.3). As a result, M_g is preferred to M_d under FIA, although the latter would be selected if we were to treat the two models as equally complex, as we could in AIC and BIC. On the same token, inequality constraints can have similar effects on model complexity. For example, M_{Ddbg} (FIA = 55.2) is preferred to M_{bg} (FIA = 55.6) if no inequality constraints in M_{bg} are considered, but the preference is reversed if the constraints are considered (FIA′ = 54.2 for M_{bg}).

Turning the discussion to model selection, we first consider the results in Table 1 obtained when no inequality constraints are considered. FIA selects M_{dbg} with FIA = 53.8 as the best model among the 16 models under consideration; so does BIC. On the other hand, AIC selects M_{dg}, a model with one additional parameter. Now let us consider the inequality constraints. Obviously, there would be no changes in the conclusion for AIC and BIC, because inequality constraints do not change the fit of the models for this data set, and the complexity measures in both criteria are blind to inequality constraints. FIA still favors M_{dbg} with inequality constraints. In summary, from among a total of 32 models compared, including the ones with inequality constraints, we conclude that M_{dbg} with inequality constraints is the best generalizing model of all.

The model-selection analysis discussed so far is conducted for models obtained by considering all possible combinations of constraints, equality and inequality, on the four parameters (D, d, b, g). In addressing the theoretical issue raised in Rose et al. (1975) and Batchelder and Riefer (1990), however, one only needs to consider the two parameters of main interest, D and d. Under this more focused scope, there are four relevant models to compare: M_0, M_D, M_d, and M_{Dd}, along with their inequality constraints. Among these four, M_d is favored under all three criteria, AIC, BIC, and FIA. According to this best-generalizing model, the two treatment conditions with related and unrelated sentences differ in item recognition (D) but not in source monitoring (d). An

implication of this conclusion is that semantic information is useful in the recognition task but does not include any information about language source.

Essentially the same conclusion as ours was reached by both Rose et al. (1975) and Batchelder and Riefer (1990). Although this particular example may be somewhat disappointing, one should note that generally speaking, FIA-based model-selection analysis allows one to entertain and evaluate all types of theoretical hypotheses of interest and that the effort is in turn likely to generate much richer and deeper insights into the underlying cognitive processes than analyses based on traditional methods such as LRT and ANOVA.

Modeling pair-clustering data. Our second example demonstrating the application of FIA-based model selection is related to the pair-clustering experiment of Batchelder and Riefer (1980, Experiment 1A). The purpose of that experiment was to examine the effects of within-category spacing on recall performance. They hypothesized that a small lag between a pair of categorically related words facilitates the formation and storage of a pair cluster, whereas a large lag facilitates the retrieval process. In the experiment, participants first studied a word list that consisted of both paired words and singletons and were then tested in a free recall task. The paired words were two words that were categorically related. In the word list, each pair of words occupied positions that were separated by J = 0, 4, 12, or 24 words unrelated to the pair. This study–recall cycle was repeated for five trials. The data set that we reanalyze in this article was from Batchelder and Riefer (1986, Table 1).

Batchelder and Riefer (1986) used LRTs to analyze the data with the pair-clustering MPT model. The original version of this model is shown in Figure 4. The model assumes three parameters: c, the probability of pairs being clustered and stored in memory; r, the (conditional) probability of a stored pair being retrieved from memory; and u, the probability of a single item being stored and retrieved from memory for either pairs or singletons. Accordingly, response category E_1 indicates recalling adjacently both items of a studied pair; E_2 indicates nonadjacently recalling both items of a pair; E_3 indicates recalling only one item; E_4 indicates recalling neither item in a pair; and finally, F_1 and F_2 indicate successful and unsuccessful recall of a singleton, respectively. Because there were four conditions for category pairs in the experiment, the MPT model for each trial consists of four trees for category pairs, one for each lag condition, and another tree for singletons. If the parameter u is assumed to be different for pairs and singletons and for different lag conditions, there will be 13 parameters for each trial, with 12 for category pairs and 1 for singletons, and 65 parameters for the five trials. A typical pair-clustering model assumes a single u, thereby reducing the number of parameters to 9 for each trial and 45 for the entire data set.

The results from separate LRTs for the five trials performed by Batchelder and Riefer (1986) showed that parameter c was significantly different across lag conditions for the data from all five trials. However, r was significant for Trials 1 and 3, but not for the other trials. Another LRT with data from all five trials combined revealed that c was
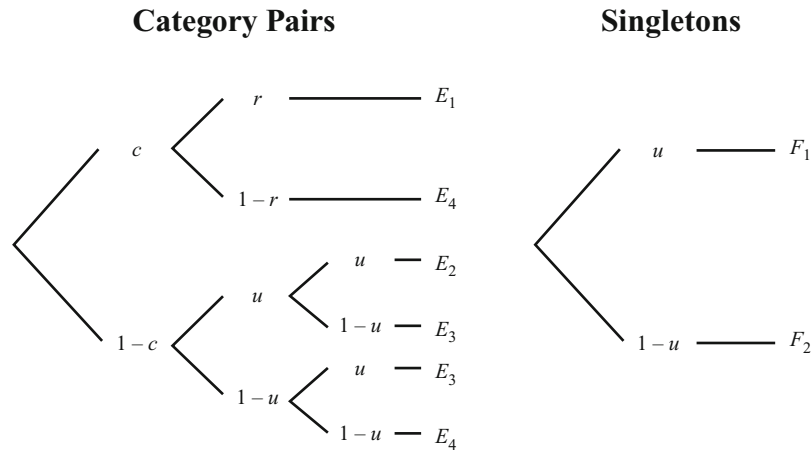
Category Pairs   ## Singletons

Figure 4. Batchelder and Riefer's (1980, 1986) MPT model of pair clustering.

significantly different across lag conditions, but r was only marginally significant. These results, taken together, support the hypothesis that small lag between category pairs facilitates the formation and storage of pair clusters, but offer only marginal support for the hypothesis that long lag facilitates the retrieval of pairs. Furthermore, given that boundary maximum likelihood estimates have been obtained from the inequality constraints when fitting the model, the LRTs' results would be uninterpretable.

Now we reanalyze the data by FIA-based model selection. Table 2 summarizes the results. There are eight models to be compared. M_0 is the model with 65 parameters described above. M_u, with 45 parameters, assumes a single u for each trial, as in typical pair-clustering models. From this model, various equality constraints on c and r are applied to form the rest of the models. Models M_{ur} and M_{uc} assume the same r and c, respectively, across the four conditions. In M_{ucr}, both c and r are assumed to be the same throughout the lag conditions. The three primed models, M'_u, M'_{ur}, and M'_{uc}, differ from the unprimed ones in that in the former, additional inequality constraints are imposed on the relevant parameters across the four lag conditions, such as $c_{J=0} > c_{J=4} > c_{J=12} > c_{J=24}$, $r_{J=0} <$

$r_{J=4} < r_{J=12} < r_{J=24}$, or both, to embody the theoretical hypotheses concerning the order of those parameters. Such constraints do not change the number of parameters in the model, but they may lead to a larger $-$LML value, because the maximum likelihood is searched over the restricted and thus smaller parameter space. This is indeed observed in Table 2 for all three pairs of models. For the same reason, inequality constraints reduce model complexity.

From Table 2, one can observe the trade-off between goodness of fit and model complexity; the smallest $-$LML value (141.3) is achieved by the most complex model M_0 with $C_{FIA} = 137.0$. At the other end of the complexity spectrum, M_{ucr}, with the least number of parameters (15), gives the largest $-$LML value (206.2) and the smallest C_{FIA} value (43.9). We also note that models with the same number of parameters can differ greatly in their complexity. For example, the four models M_{uc}, M'_{uc}, M_{ur}, and M'_{ur} all have 30 parameters, but their C_{FIA} complexity varies from the smallest (59.9) to the largest (79.5), due to the combination of tree structure and inequality constraints on parameters. Such complexity difference between two models with the same number of parameters can be larger

Table 2
Summary of Model-Selection Results for Pair-Clustering Data
From Batchelder and Riefer (1980, Experiment 1A)

Model	S	$-$LML	C_{AIC}	AIC	C_{BIC}	BIC	C_{FIA}	FIA
M_0	65	141.3	65	206.3	262.51	403.8	137.0	278.3
M_u	45	155.2	45	200.2	181.74	337.0	106.1	261.3
M'_u	45	157.7	45	202.7	181.74	339.5	73.7	**231.4**
M_{uc}	30	196.5	30	226.5	121.16	317.6	75.8	272.3
M'_{uc}	30	200.5	30	230.5	121.16	321.7	59.9	260.4
M_{ur}	30	168.0	30	**198.0**	121.16	289.1	79.5	247.5
M'_{ur}	30	169.5	30	199.5	121.16	290.7	63.6	233.1
M_{ucr}	15	206.2	15	221.2	60.58	**266.7**	43.9	250.1

Note—The data can be found in Batchelder and Riefer (1986, Table 1). The total sample size is $N = 3,220$. See the main text for the description of the models. S, the number of parameters; LML, the natural logarithm of the maximized likelihood; C, complexity; AIC, Akaike information criterion; BIC, Bayesian information criterion; FIA, Fisher information approximation. The minimum value for each of the selection criteria is shown in bold.

than the difference in LML, thus affecting model-selection results. Such a pattern of result is indeed observed in the table. That is, among the same four models with 30 parameters, both AIC and BIC select M_{ur}, whereas FIA picks M'_{ur}.[6] This is because although M_{ur} fits the data better than M'_{ur} ($-$LML $=$ 168.0 vs. 169.5), the model is more complex ($C_{FIA} =$ 79.5 vs. 63.6) and thus yields a larger FIA than its counterpart (FIA $=$ 247.5 vs. 233.1).

Turning the discussion to model selection, among the eight models[7] compared, AIC favors the 30-parameter M_{ur}, and BIC favors the most restrictive model, M_{ucr}, whereas FIA selects M'_u. Note that the model M'_u imposes inequality constraints on both c and r in the directions consistent with the experimental hypotheses of Batchelder and Riefer (1980). In other words, our FIA-based reanalysis of the data supports the hypotheses in their ordered form. As was discussed earlier, the LRT results by Batchelder and Riefer (1986) indicated that the hypothesized within-pairs spacing effect on parameter r was inconclusive, whereas the hypothesized effect on parameter c was supported. The specific order relationships among the hypotheses were not, however, examined. Model M'_{ur} embodies this interpretation of the data and, interestingly, turns out to be the second best model after M'_u. Especially if the inequality constraints were not considered in model selection, FIA would choose M_{ur}, leading to a different conclusion. This shows that the hypothesized order relationship of parameters, which may restrict the parameter space and reduce the complexity of the model, can lead to different model-selection conclusions and, as such, should not be neglected.

To summarize, we demonstrated the application of the FIA-based model-selection approach for selecting among MPT models of pair clustering for the Batchelder and Riefer (1980) data set. The flexibility of the approach allowed us to construct and test a variety of MPT models, including models with inequality constraints on parameters. We compared the results from our analysis with those from the LRT-based analysis of the same data reported in Batchelder and Riefer (1986). By and large, the same scientific conclusions were drawn from either analysis, although our FIA-based analysis provides more definitive support for the hypotheses in their ordered form originally formulated and tested in Batchelder and Riefer (1980, 1986).

Conclusion

MPT modeling represents a theoretically motivated and statistically justified methodology for evaluating cognitive capacities for various experimental paradigms. Selecting among different MPT models is especially important, both in addressing theoretical issues and in validating an MPT model in a particular experimental paradigm. In this article, we have introduced the MDL-based model-selection method to the practitioners of MPT modeling. Especially to facilitate the use of this new methodology, we provide a general purpose computer program in MATLAB that can be exploited to compute FIA for any MPT model. Two example applications of MDL model selection with real data sets selected from different experimental paradigms were also discussed.

MDL's flexibility of application to a wide range of model comparison situations that may arise in MPT modeling makes it an attractive alternative to traditional methods, such as LRT, AIC, and BIC. First, instead of using a series of null hypothesis significance tests such as LRTs, MDL represents a model-selection approach in which the models in contention are ranked by their generalizability, or, equivalently, predictive accuracy, which is the hallmark of model selection. Also, unlike AIC and BIC, MDL considers not only the effects of the number of parameters and sample size on model complexity but also, importantly, the effect of functional form, which alone can significantly contribute to complexity and sometimes even more so than the number of parameters. As a result, another advantage of MDL over the other three methods is that MDL allows one to incorporate inequality constraints on model parameters. Last but not least, with the freely available MATLAB program, FIA is now entirely within the reach of everyday practitioners.

As is usually the case with any new methodology, MDL as presented in this article is not without shortcomings. For example, it does not address the issues of model misspecification and individual differences. To allow for model misspecification, exact equality constraints should be replaced by fuzzy equality constraints, as was done in an elegant sampling-based method known as *population-parameter mapping* proposed by Chechile (1998). Regarding individual differences, one way of incorporating this important factor is through hierarchical modeling in which parameter values corresponding to different individuals are assumed to be sampled from a common distribution (see, e.g., Klauer, 2010; Smith & Batchelder, 2010). Although it is possible in theory to address these issues within the MDL framework, it is beyond the scope of the present work.

In conclusion, model selection lies at the core of the scientific inference process. Accordingly, a theoretically well-justified and widely applicable methodology can help advance science. We believe that MDL represents such a methodology and that it provides versatile yet powerful tools for assessing the validity of MPT models in a way that goes beyond the shortcomings of current methods such as LRT, AIC, and BIC.

One final note: Statistical model-selection techniques alone, however sophisticated, are not a panacea for all inference problems. Other, nonstatistical means of model evaluation, such as plausibility, interpretability, and explanatory adequacy, are equally—if not more—important. Instead of automatic tools implemented in software, statistical model-selection methods can be most useful if they are combined with the judicious use of sound subjective but scientific judgment.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings from the Second Symposium on Information Theory* (pp. 267-281). Budapest: Akadémiai Kiadó.

Batchelder, W. H. (2009). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71-93). Washington, DC: American Psychological Association.

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, **87**, 375-397.

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical & Statistical Psychology*, **39**, 129-149.

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

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

Batchelder, W. H., & Riefer, D. M. (2007). Using multinomial processing tree models to measure cognitive deficits in clinical populations. In R. W. J. Neufeld (Ed.), *Advances in clinical cognitive science: Formal modeling of processes and symptoms* (pp. 19-50). Washington, DC: American Psychological Association.

Bishara, A. J., & Payne, B. K. (2008). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology*, **45**, 524-534.

Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). Belmont, CA: Duxbury Press.

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

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

Chechile, R. A. (2007). A model-based storage–retrieval analysis of developmental dyslexia. In R. W. J. Neufeld (Ed.), *Advances in clinical cognitive sciences: Formal modeling of processes and symptoms* (pp. 51-80). Washington, DC: American Psychological Association.

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

Grünwald, P. [D.] (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.

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

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (Eds.) (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

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

Hu, X., & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments, & Computers*, **31**, 220-234.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, **75**, 70-98. doi:10.1007/s11336-009-9141-0

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multitrial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, **48**, 215-229.

Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, **45**, 131-148.

Lee, M. D., & Pope, K. J. (2006). Model selection for the rate problem: A comparison of significance testing, Bayesian, and minimum

description length statistical inference. *Journal of Mathematical Psychology*, **50**, 193-202.

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

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Guest editors' introduction: Special issue on model selection. *Journal of Mathematical Psychology*, **44**, 1-2.

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

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

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

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*, **11**, 961-974.

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

Purdy, B. P., & Batchelder, W. H. (2009). A context-free language for binary multinomial processing tree models. *Journal of Mathematical Psychology*, **53**, 547-561.

Raftery, A. E. (1999). Bayes factors and BIC: Comment on "A critique of the Bayesian information criterion for model selection." *Sociological Methods & Research*, **27**, 411-427.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

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

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

Riefer, D. M., & Batchelder, W. H. (1995). A multinomial modeling analysis of the recognition-failure paradigm. *Memory & Cognition*, **23**, 611-630.

Riefer, D. M., Hu, X., & Batchelder, W. H. (1994). Response strategies in source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 680-693.

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

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

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

Rose, R. G., Rose, P. R., King, N., & Perez, A. (1975). Bilingual memory for related and unrelated sentences. *Journal of Experimental Psychology: Human Learning & Memory*, **1**, 599-606.

Saegert, J., Hamayan, E., & Ahmar, H. (1975). Memory for language of input in polyglots. *Journal of Experimental Psychology: Human Learning & Memory*, **1**, 607-613.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, **21**, 168-175.

Schweickert, R., & Chen, S. (2008). Tree inference with factors selectively influencing processes in a processing tree. *Journal of Mathematical Psychology*, **52**, 158-183.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review*, **56**, 49-62.

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, **54**, 167-183.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, **14**, 779-804.

WAGENMAKERS, E.-J., & WALDORP, L. (2006). Editors' introduction. *Journal of Mathematical Psychology*, **50**, 99-100.

WEAKLIEM, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, **27**, 359-397.

WU, H., MYUNG, J. I., & BATCHELDER, W. H. (2010). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, **54**, 291-303.

NOTES

1. In this article, an MPT model may refer either to a model for a particular experimental paradigm (e.g., source monitoring, with three trees) or to a set of such models, each representing a different experimental condition in an experiment.

2. The interested reader is directed to two recent *Journal of Mathematical Psychology* special issues on model selection for discussion and example applications of these and other methods of model selection (Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006).

3. More precisely, it is a large-sample approximation to normalized maximum likelihood (NML; Myung et al., 2006; Rissanen, 2001), which can be considered the shortest code length that a model can achieve in coding the current data in the worst case of its true distribution. NML can be computationally intensive and will not be discussed in this article.

4. It should be noted that Equation 7 is valid only when the model is globally identified (i.e., if different parameter values generate different category probabilities). For a nonidentified model, one should reparameterize it into an equivalent but identified model before applying this formula.

5. Knapp and Batchelder (2004) showed that BMPT models with such inequality constraints can be reparameterized into BMPT models without inequality constraints. Our program computes the integral of the original model directly, without invoking such reparameterization.

6. Technically AIC and BIC cannot favor M'_{ur} over M_{ur}, because these models have the same C_{AIC} and C_{BIC} values but the fit of M'_{ur} cannot be better than M_{ur}.

7. We include all eight models for demonstration purposes only. To address the theoretical issues of interest, only the following five models are to be compared: M_0, M'_u, M'_{uc}, M'_{ur}, and M_{ucr}. The model-selection conclusion based on FIA will not change.

SUPPLEMENTAL MATERIALS

A MATLAB .m file that computes the FIA complexity C_{FIA} for BMPT models in general, with or without inequality constraints, may be downloaded from http://pbr.psychonomic-journals.org/content/supplemental.