

# On the Minimum Description Length Complexity of Multinomial Processing Tree Models

Hao Wu and Jay I. Myung

The Ohio State University

William H. Batchelder

University of California, Irvine

## Abstract

Multinomial processing tree (MPT) modeling is a statistical methodology that has been widely and successfully applied for measuring hypothesized latent cognitive processes in selected experimental paradigms. This paper concerns model complexity of MPT models. Complexity is a key and necessary concept to consider in the evaluation and selection of quantitative models. A complex model with many parameters often overfits data beyond and above the underlying regularities, and therefore, should be appropriately penalized. It has been well established and demonstrated in multiple studies that in addition to the number of parameters, a model's functional form, which refers to the way by which parameters are combined in the model equation, can also have significantly effects on complexity. Given that MPT models vary greatly in their functional forms (tree structures and parameter/category assignments), it would be of interest to evaluate their effects on complexity. Addressing this issue from the minimum description length (MDL) viewpoint, we prove a series of propositions concerning various ways in which functional form contributes to the complexity of MPT models. Computational issues of complexity are also discussed.

## Introduction

The issue of model complexity is of fundamental importance in the evaluation and selection of statistical models and has received much attention recently in the field of mathematical psychology (see, e.g., Myung, Forster & Browne, 2000; Myung, 2000; Grünwald, 2000; Myung, Navarro & Pitt, 2006; Pitt & Myung, 2002). Model complexity refers to a model's inherent flexibility that allows the model to fit diverse data patterns. A model that gives good fit to a wide range of data patterns is more complex than one that can only fit a limited range of data patterns. Generally speaking, the more parameters a model has, the more complex it is.

The relevance of model complexity in model selection has to do with the overfitting phenomenon. The flexibility of a model is a double edged sword. The flexibility allows the model to readily capture the regularities underlying the observed data but also enables it to improve model fit by capitalizing on random noise and thereby result in over-fitting the data beyond and above the regularities. Consequently, choosing among models based solely on goodness of fit (i.e., how well each model fits observed data) can lead to misleading conclusions about the underlying process, unless the overfitting effect is appropriately taken into account. This is realized in model selection by defining a selection criterion that trades off a model's goodness of fit for its simplicity so as to avoid overfitting. The resulting criterion, known as generalizability (or predictive accuracy), quantifies how well a model can predict future yet unseen data patterns from the same underlying process that has generated the current data pattern.

Of particular importance in estimating a model's generalizability is to accurately measure its complexity considering all relevant dimensions of model complexity. This is especially true for multinomial processing tree (MPT) models, for reasons detailed in a later part of this section. MPT modeling is a statistical methodology introduced in the 1908s for measuring latent cognitive

---

Word count of text and appendices: 10000  
Running head: Complexity of Multinomial Processing Tree models

This paper is based on Hao Wu's Master of the Arts thesis submitted to the Ohio State University in July 2006. It is supported in part by National Institute of Health Grant R01-MH57472 to JIM. Work on this paper by the third author was supported in part from a grant from the National Science Foundation: SES-0616657 to X.Hu and W.H. Batchelder (Co-PIs). We wish to thank Mark A. Pitt and Michael W. Browne for valuable feedbacks provided for the project. Correspondence concerning this article should be addressed to Hao Wu, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH 43210. E-mail: wu.498@osu.edu. Tel: 614-292-5510

capacities in selected experimental paradigms (Batchelder & Riefer, 1980; Riefer & Batchelder, 1988; Batchelder & Riefer, 1986). MPT models have been successfully applied to modeling performance in a 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 & Riefer, 1999; Hu & Batchelder, 1994; Chechile, 2004; Riefer & Batchelder, 1988,9; Riefer, Hu & Batchelder, 1994). The data structure requires that participants make categorical responses to a series of test items, and an MPT model parameterizes a subset of probability distributions over the response categories by specifying a processing tree designed to represent hypothesized cognitive steps in performing a cognitive task, such as memory encoding, storage, discrimination, inference, guessing and retrieval.

To give a concrete example of MPT modeling, consider a source monitoring experiment in which participants, after having studied a list of items from two sources, A and B, are asked to judge the source of a test item as either from A, from B, or new (i.e., stimulus from neither source). MPT models for such experiments typically consist of three distinct trees (Batchelder & Riefer, 1990; Bayen, Murnane & Erdfelder, 1996), each of which models the hypothetical processes a participant might employ to select a response to a given type of item with three possible responses, A, B, or N. One such model is depicted on the left panel of Figure 1. A distinguishing feature of this one-high-threshold model (1HTM) is that it assumes only thresholds for old items to be correctly detected with probability  $D_1$  and  $D_2$  for sources A and B, respectively. If an old item is correctly detected as old, a discriminating decision on its source is made and the parameters  $d_1$  and  $d_2$  represents this process for items from sources A and B, respectively. If either the detection or the discrimination process fails, one or more guessing processes involving parameters  $b$ ,  $g$  and  $a$  follows. For new items, however, the model assumes no detection process, and instead response selection is determined solely by guessing represented by the parameter  $g$ . By imposing constraints successively on the model parameters, a hierarchy of sub-models can be derived from the original model. This is shown in the right panel of Figure 1. Likewise, new processes can be added to the original model. For example, the two-high-threshold model (2HTM) as depicted in Figure 2 assumes a separate detection threshold  $D_3$  for the new items.

One prominent aspect of these MPT models of source monitoring is that they differ from one another not only in terms of the number of parameters but also in terms of functional form. For example, all three models, 6a, 6b and 6c, in Figure 1 have the same number of parameters (6) but each has different functional form distinct from the others. The same can be said about the three 5-parameter models, 5a, 5b and 5c in the figure. This is also true

in general for MPT models and will be discussed in the next section. In addition to these, it is not uncommon that researchers develop and validate various MPT models with processing assumptions represented by different structures (e.g., Chechile, 2004; Bayen, Murnane & Erdfelder, 1996). In selecting among such MPT models, it would be of particular interest to accurately measuring the contributions of model complexity due to functional form, as well as number of parameters. Importance of the former factor in model selection is well documented and demonstrated for models of information integration, retention and categorization (e.g., Myung & Pitt, 1997; Pitt, Myung & Zhang, 2002; Pitt & Myung, 2002), but the issue remains to be explored in the context of MPT modeling.

In this paper we investigate the effects of model structure on the complexity of MPT models. The particular approach we take here is that of minimum description length (MDL Grünwald, 2000; Grünwald, Myung & Pitt, 2005; Myung, Navarro & Pitt, 2006; Grünwald, 2007). The desirability and success of MDL in addressing model selection problems for various types of cognitive models are well documented (e.g., Lee, 2001; Pitt, Myung & Zhang, 2002; Navarro & Lee, 2004; Lee & Pope, 2001; Myung, Pitt & Navarro, 2007). Importantly, MDL is well suited for the present purpose; among other things, the MDL complexity metric (defined in the next section) not only is theoretically well justified, intuitively interpretable and readily computable, but also importantly, takes into account both the number of parameters and functional form dimensions of model complexity. As a first step toward applying MDL to the evaluation and selection of MPT models, we address the issue of model complexity for this class of models from the standpoint of MDL.

The rest of the paper is organized as follows. We first define the class of binary MPT (BMPT) models and show how they can be constructed recursively from elementary decision nodes. A BMPT model, by definition, allows only binary choices at each decision node. Since any MPT model can be reparameterized into an equivalent BMPT model (Hu & Batchelder, 1994), it is sufficient to restrict our attention to this class of models. This is followed by a formal definition of MDL and a brief discussion of its statistical properties in relation to the issue of model complexity. The next section presents the main results of our theoretical investigations. Here we prove various propositions regarding MDL complexity of BMPT models. Some of these results are possible because of the recursive nature of the BMPT models. The section followed discusses computational issues of MDL complexity. Finally, the conclusion summarizes and recaps the contributions of the present work.

## BMPT Models and MDL

*Formal Definition of BMPT Models*

Binary multinomial processing tree (BMPT) models are a subclass of MPT models that involve exactly two processing possibilities at each decision node of the tree. They have been defined in detail in other papers (e.g. Batchelder & Riefer, 1999; Knapp & Batchelder, 2004; Purdy & Batchelder, in press), so our definition will be succinct and emphasize some of the properties of the class that turn out to play a role in establishing some of the propositions involving model complexity. Any BMPT model is built out of a set of  $J > 1$  mutually exclusive and exhaustive observable categories,  $\mathbf{C} = \{C_1, C_2, \dots, C_J\}$ , and a set  $\Theta$  of  $S$  latent parameters arrayed for convenience in a vector  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_S)'$ . Each parameter  $\theta_s$  represents the probability of the occurrence, and  $(1 - \theta_s)$  the non-occurrence, of some latent cognitive event, such as storing an item in memory, retrieving a stored item, guessing a response given imperfect memory, making a particular inference, and the like. The parameters are functionally independent and each is free to vary in  $(0,1)$ , so the parameter space for the model is given by  $\Omega = \{\boldsymbol{\theta} \in (0, 1)^S\}$ .

Each BMPT model has a structural component and a computational component. The structural component is specified by the assignment of observable categories and latent parameters to the nodes of a full binary tree (FBT). A FBT is a special type of digraph  $\mathbf{D} = (V, R)$ , where  $V$  is a full set of nodes,  $R \subseteq V \times V$ , and  $(v, v') \in R$  represents a directed edge from  $v$  to  $v'$ . In order to define the class of FBTs, we need the concepts of the parents and children of a digraph node. For any node  $v \in V$ , the set of parents of  $v$  is defined as  $\{v' | (v', v) \in R\}$ , and the set of children of  $v$  is defined as  $\{v' | (v, v') \in R\}$ . A FBT is a directed graph with a single root node with no parents, a set of terminal nodes with no children (called leaves), and satisfying the properties that every node but the root has exactly one parent and every non terminal node has exactly two children. In this paper, FBTs are oriented with the root on the left, so every nonterminal node has an upper child and a lower child. The structural component of a BMPT model is completed by specifying the assignment of a parameter to each nonterminal node and a category to each leaf of the FBT. It is possible to assign a category to more than one leaf and a parameter to more than one nonterminal node, and also a fixed number  $x \in (0, 1)$  can be assigned to a nonterminal node instead of a parameter. The left panel of Figure 1 exhibits three BMPT models. Note that the root is on the left and the leaves are on the right with their assigned categories. To the right of every nonterminal node is a parameter, for example  $D_1$  written next to the upper child of the node and the 'complement' of the parameter, for example,  $1 - D_1$ , written next to the

lower child of the node. By convention this notation is designed to depict the case that the parameter  $D_1$  is assigned to the node in question. This convention will be explained more fully when we discuss the computational component of a BMPT model.

The computational component of a BMPT model provides the set of probability distributions over the categories of the model in terms of the parameters and numbers assigned to the nonterminal nodes of the model's FBT. These probability distributions are determined by the branch probabilities of the BMPT model. Starting at the root, a branch leading to a leaf is probabilistically selected by a series of binary choices governed by the parameters (or numbers) associated with the nonterminal nodes of the tree. Return again to Figure 1 (left panel) and consider any nonterminal node  $v$ . The parameter (or number) assigned to that node represents the conditional probability that if node  $v$  is reached by prior binary decisions, the directed edge to the upper child is taken with probability  $\theta_s$  and the edge to the lower child is taken with probability  $1 - \theta_s$ .

In general, a BMPT tree may have  $I_j$  leaves assigned to category  $C_j$ . Let  $p_{ij}(\theta)$  represents the probability of selecting the branch  $B_{ij}$  leading to  $C_j$ , as a function of the parameters  $\theta \in \Omega$ . Then the computational rules specify that the probability of category  $C_j$  is given by summing the  $I_j$  probabilities,

$$p_j(\theta) = \sum_{i=1}^{I_j} p_{ij}(\theta) \tag{1}$$

Further from the computational rules, these branch probabilities take a particular form given by (see Hu & Batchelder, 1994, for more details)

$$p_{ij}(\theta) = c_{ij} \prod_{s=1}^S \theta_s^{a_{ijs}} (1 - \theta_s)^{b_{ijs}} \tag{2}$$

where  $a_{ijs}$  and  $b_{ijs}$  are, respectively, the number of times  $\theta_s$  and  $1 - \theta_s$  that appear on the path  $B_{ij}$  of category  $C_j$ , and  $c_{ij}$  is the product of the numbers along the same path or set to unity if there are no numbers along that path. From equations (1) and (2) it is seen that any BMPT model  $M_J$  with  $J$  categories parametrically specifies a subset  $\Lambda(M_J)$  of the simplex of possible pdfs over  $J$  categories given by  $\Lambda_J = \{\mathbf{p} = (p_1, p_2, \dots, p_J) | p_j \geq 0, \sum_j p_j = 1\}$ .

In the present paper it is assumed that several participants each make categorical responses to the same set of items and that these responses are independent and identically distributed into the  $J$  categories of a model. Let  $n_j$  be the number of these responses that fall into category  $C_j$ ,  $\mathbf{n} = (n_1, n_2, \dots, n_J)$

and  $N = \sum_j n_j$ . Then from the computational rules in equations (1) and (2),  $\mathbf{n}$  is distributed as a structured multinomial distribution given by

$$f(\mathbf{n}|\theta) = \binom{N}{n_1, \dots, n_J} \prod_{j=1}^J p_j^{n_j}(\theta) \tag{3}$$

The Fisher information matrix of a BMPT model in terms of the representation of equations (1), (2) and (3) is given in Lemma 2 in the Appendix.

It is particularly important to note the role of the structural component of a BMPT model in this paper. First, it gives rise to the functional form differences among different BMPT models, which is a central issue in this article. Because the structural component involves more than the structure of the FBT, functional form differences may still arise for BMPT models with the same FBT. In particular, how the leaves of the FBT are combined into categories and how the parameters are assigned to the nonterminal leaves may change a model’s functional form. To see the different sources of functional form differences, we note the 1HTM and 2HTM described in Introduction differ in their tree structures (though they may have the same number of parameters after assuming appropriate constraints), while the functional difference between models 5a and 5b in Figure 1 is entirely due to the different assignment of parameters to the nodes. To avoid possible confusion in referring to identical BMPT models, we use the following definitions in this paper.

**Definition 1** (functionally identical). *BMPT models are called functionally identical if*

1. *they share the same FBT tree structure,*
2. *their parameter assignment gives the same partition of their non-terminal nodes, and*
3. *their leafs are combined into categories in the same way.*

It is evident that functionally identical BMPT models can be reparameterized to each other by properly mapping their category sets and parameter sets.

**Definition 2** (functionally identical with identical parameter assignment). *BMPT models are called functionally identical with common parameter set if they are functionally identical with identical parameter assignment from the same parameter set.*

It should be noted that the definition requires more than functionally identical BMPT models with common parameter sets, in which the way parameters are assigned to the nodes may be different.

Furthermore, the structural component of a BMPT model satisfies several recursive properties that are useful in understanding the model selection properties of BMPT models developed in the next section. First suppose  $\mathcal{A}_1$  and  $\mathcal{A}_2$  are two BMPT models where some of the categories and some of the parameters may be shared between models. Let  $p$  be a parameter with space  $(0,1)$  which can either be one of the parameters in the two models or functionally independent of those parameters. Then we can construct a new BMPT model, denoted by  $p\mathcal{A}_1\mathcal{A}_2$ , by introducing a root node assigned to  $p$  and associating  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , respectively, to the upper and lower children of the new root node. The number of leaves in the new model is the sum of the leaves in the two component models, and the parameters consist of all the parameters from the two models plus the new parameter  $p$  (unless it is already one of the old parameters). In fact Purdy & Batchelder (in press) have shown that all BMPT models can be built up by joining pairs of models in this way starting with elemental BMPT models consisting of a single category. A second recursive property of BMPT models is that one can select one or more categories in a BMPT model  $\mathcal{A}$  and replace each of them with another BMPT model, and the result is a BMPT model. In this case the number of leaves is the sum of the leaves in all models minus the number of BMPT models attached to  $\mathcal{A}$ . The first two panels of Figure 6 illustrates these two ways that new BMPT models can be constructed from other BMPT models.

*Minimum Description Length*

The principle of minimum description length (MDL) originates from algorithmic coding theory in computer science. According to the 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, the better the model generalizes (Grünwald, Myung & Pitt, 2005; Grünwald, 2007).

There are currently two implementations of the MDL principle for model selection, Fisher Information Approximation (FIA: Rissanen, 1996) and Normalized Maximum Likelihood (NML: Rissanen, 2001). For a model with probability density  $f(y|\theta)$  and observed data  $y$ , they are defined as an additive combination of goodness of fit and model complexity terms:

$$FIA = -LML + C_{FIA} \tag{4}$$

$$NML = -LML + C_{NML} \tag{5}$$

where

$$\begin{aligned}
 LML &= \ln f(y|\hat{\boldsymbol{\theta}}(y)) \\
 C_{\text{FIA}}(N) &= \frac{S}{2} \ln \frac{N}{2\pi} + \ln \int_{\Theta} \sqrt{|I(\boldsymbol{\theta})|} d\boldsymbol{\theta} \tag{6}
 \end{aligned}$$

$$C_{\text{NML}}(N) = \ln \int_{\mathcal{X}} f(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})) d\mathbf{x} \tag{7}$$

In the above equations, LML standing for the logarithm of the maximized likelihood ( $\ln f(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))$ ) represents a goodness of fit measure,  $S$  is the number of parameters,  $N$  is the sample size,  $\Theta$  is the parameter space,  $\mathcal{X}$  is the set of all possible sample with sample size  $N$ , and  $I(\boldsymbol{\theta})$  is the Fisher information matrix (e.g. Casella & Berger, 2001) of sample size 1 defined as  $I(\boldsymbol{\theta})_{ij} = -E \left[ \frac{\partial^2 \ln f(x_1|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right]$ ,  $i, j = 1, \dots, S$ . When data is discrete, as in the case of MPT modeling, the integration in Equation (7) is replaced by summation. Note that both complexity measures are functions of sample size  $N$ . Under each selection method, a smaller criterion value indicates better generalization, and thus, the model that minimizes the criterion should be chosen.

In both FIA and NML, generalizability is measured as a trade-off between goodness of fit and simplicity, thus formalizing the Occam’s razor (Myung & Pitt, 1997). Specifically, both methods share the same goodness of fit measure (i.e., LML) but differ from each other in how complexity is measured, represented by  $C_{\text{FIA}}$  and  $C_{\text{NML}}$ , respectively. As mentioned earlier, models with the same number of parameters but with different equation forms can differ in complexity. This is called the *functional form* dimension of model complexity. To illustrate, two psychophysics models in perception,  $y \sim N(ax^b, \sigma^2)$  and  $y \sim N(a \ln(x + b), \sigma^2)$ , may have different complexity values, despite the fact that they both have two parameters.

As can be seen in Equation (6), the FIA complexity measure,  $C_{\text{FIA}}$ , takes into account the number of parameters ( $S$ ), the sample size ( $N$ ) and importantly, functional form captured through the Fisher information matrix ( $I(\boldsymbol{\theta})$ ). This is unlike two selection criteria most commonly in use, namely, the Akaike Information Criterion (AIC: Akaike, 1973) and the Bayesian Information Criterion (BIC: Schwartz, 1978), neither of which considers functional form. Regarding the NML complexity measure,  $C_{\text{NML}}$  in Equation (7), note that this complexity term is defined as the logarithm of *the sum of maximum likelihoods* the model can provide across all possible data (of sample size  $N$ ) that could potentially be observed in a given experimental design. Accordingly, a model that can fit well almost every data pattern, human or non-human, would be more complex than another model that fits well a few data patterns but fits poorly other

data patterns, thereby nicely capturing our intuition about model complexity. It is worth noting that between  $C_{\text{FIA}}$  and  $C_{\text{NML}}$ , the former is derived as an asymptotic approximation of the latter (Rissanen, 1996, 2001), and as such, all three dimensions of complexity that  $C_{\text{FIA}}$  captures are also represented, though implicitly, in  $C_{\text{NML}}$ .

It should be noted that Equation (6) is only applicable for statistically identified<sup>1</sup> models. If a model is not identified,  $C_{\text{FIA}}$  can be taken as that of a statistically equivalent but identified model.

In what follows, we present results from our theoretical investigation of the properties of  $C_{\text{FIA}}$  and  $C_{\text{NML}}$  for MPT models and also discuss some computational issues that may arise in practical implementations of these measures.

### Theoretical Investigation of MDL Complexity

In this section we prove some important properties of the MDL complexity measures that are valid for the entire class of BMPT models, with a particular focus on the effects of functional form on complexity.

#### *Complexity of Nested and Equivalent Models*

In exploring the issue of model complexity for BMPT models, it is useful to note two observations about MDL that apply to any family of models. Our first observation concerns complexity relationship among nested MPT models. If model  $\mathcal{A}$  is nested within model  $\mathcal{B}$ , then the complexity of model  $\mathcal{A}$  is no greater than that of model  $\mathcal{B}$ , or formally,  $C_{\text{NML},\mathcal{A}} \leq C_{\text{NML},\mathcal{B}}$ . This observation is self-evident. If model  $\mathcal{A}$  is nested within model  $\mathcal{B}$ , then the parameter space of model  $\mathcal{A}$  is a subset of that of model  $\mathcal{B}$ . Consequently, for every data set, the maximum likelihood for model  $\mathcal{A}$  would be equal to or smaller than that of model  $\mathcal{B}$ . Therefore, according to the definition of  $C_{\text{NML}}$  in equation (5), model  $\mathcal{A}$ 's  $C_{\text{NML}}$  value is not larger than that of model  $\mathcal{B}$ . Our second observation concerns models that are reparameterizations of one another. Such models are statically equivalent in the sense that they entail exactly the same set of possible probability distributions. Since the maximized likelihood function for each possible data set is one of these distributions, reparameterized models generate identical values of  $C_{\text{NML}}$  and therefore are equally complex. The observation that equivalent models have the same complexity implies that model complexity is an *intrinsic* property of the model, independent of the model's parameterization and identifiability.

---

<sup>1</sup>An MPT model is statistically identified if different parameter values produce different category probabilities. Strictly speaking, equation (6) is applicable if a model is identified excluding a subset of its parameter space that has zero measure.

These observations are helpful in understanding the fact that an apparently more complicated model may turn out to have the same or even smaller complexity value than one that looks much simpler. One good example is given by the following proposition concerning two source monitoring models: the two-high-threshold five parameter model (2HTM-5) and the one-high-threshold four parameter model (1HTM-4). Both 1HTM and 2HTM have been described in the Introduction and they are depicted in Figure 1 and 2 respectively. In particular, 2HTM-5 assumes the same constraints to 1HTM-4, with one extra parameter  $D^*$  for the threshold for the new item. This model is known to be not identified (e.g. Bayen, Murnane & Erdfelder, 1996).

**Proposition 1.** *The 1HTM-4 and 2HTM-5 models are statistically equivalent.*

*Proof.* We only need to prove 2HTM-5 is nested in 1HTM-4. To distinguish between the parameters in the two models, we use  $(\tilde{D}, \tilde{D}^*, \tilde{d}, \tilde{b}, \tilde{g})$  for parameters of 2HTM-5 and  $(D, d, b, g)$  for 1HTM-4. The following set of equations reparameterizes the 2HTM-5 into 1HTM-4.

$$\begin{aligned} g &= \tilde{g} \\ b &= \tilde{b}(1 - \tilde{D}^*) \\ D &= \frac{\tilde{b}\tilde{D}^* + \tilde{D}(1 - \tilde{b})}{\tilde{b}\tilde{D}^* + (1 - \tilde{b})} \\ d &= \left( \frac{\tilde{b}\tilde{D}^* + (1 - \tilde{b})}{\tilde{b}(\tilde{D}^*/\tilde{D}) + (1 - \tilde{b})} \right) \tilde{d} \end{aligned}$$

It is self evident that for all values of  $(\tilde{D}, \tilde{D}^*, \tilde{d}, \tilde{b}, \tilde{g})$  within  $[0, 1]$ , the parameters  $(D, d, b, g)$  determined by the above equations are always within  $[0, 1]$ .  $\square$

Because the two models are statistically equivalent, they must have the same complexity, though 2HTM-5 looks to be more complex by allowing an extra threshold for the new items and assuming an extra parameter. Especially, the well-known result that 2HTM-4 is nested in 1HTM-4 (e.g. Bayen, Murnane & Erdfelder, 1996) is implied by this proposition. Consequently 2HTM-4 actually has smaller complexity than 1HTM-4 though it assumes one more threshold.

Another example concerns BMPT models with inequality constraints. In many cases, theoretical considerations such as the desired order of treatment effects are incorporated into BMPT models as inequality constraints on the parameters. Such a model with inequality constraints is nested in the original model without such constraints and therefore has smaller complexity value. Especially, Knapp & Batchelder (2004) showed that when the inequality constraints are in the form of non-overlapping linear orders, for instance,

$0 < \theta_1 < \theta_2 < \dots < \theta_k < 1$ , the BMPT model can be reparameterized into an equivalent BMPT model with the same number of parameters and categories but without inequality constraints. The second BMPT model looks more complex than the original model. It should be noted, however, since the new, unconstrained model is statistically equivalent to the original model with the inequality constraints, the complexity of the two models are the same, and both smaller than that of the original model without those constraints. This, again, indicates sometimes a model that looks more complex may turn out to have smaller complexity value than one that looks simpler.

A third example concerning BMPT models with unique parameters and categories (uBMPT) is discussed below.

*Complexity of uBMPT Models*

A special class of BMPT models results when unique parameters and categories are assigned, respectively, to the non-terminal nodes and leaves of a FBT. These uBMPT models are of special interest because every BMPT model is nested in one or more of these models. Each such model has  $J$  categories and  $J - 1$  parameters; however, for each value of  $J$ , there are many different uBMPT models that may differ greatly in their tree structures. The following proposition shows that these different models obtain the same maximal possible complexity for a categorical model. Thus for the special case of uBMPT models, the shape of the tree does not affect the complexity of the model.

**Proposition 2.** *Let  $\mathcal{M}_J$  be a unconstrained multinomial model on  $J$  categories with parameter space  $\mathbf{\Lambda}_J = \{\mathbf{p} = (p_1, p_2, \dots, p_J) | p_j \leq 1, \sum p_j = 1\}$ . Then every uBMPT model with  $J$  categories is statistically equivalent to  $\mathcal{M}_J$  and therefore has maximal possible complexity.*

*Proof.* Let  $\mathcal{M}$  be a uBMPT model with  $J$  categories. Clearly  $\mathbf{\Lambda}(\mathcal{M}) \subseteq \mathbf{\Lambda}_J$  because  $\mathcal{M}$  entails a set of probability distributions over  $J$  categories given by equations (1) and (2), where  $\forall j, I_j = 1, c_{1j} = 1$ . Below we show  $\mathbf{\Lambda}_J \subseteq \mathbf{\Lambda}(\mathcal{M})$ , i.e., all pdf on the simplex can be implied by a member in  $\mathcal{M}$ .

Suppose  $\mathbf{p} \in \mathbf{\Lambda}_J$ . Note that the parameter  $\theta_r$  assigned to the root node of  $\mathcal{M}$  partitions the  $J$  categories into two sets  $C_R, C_L$ , which are, respectively, the categories that are below the right and left children of the root node. Assign the parameter  $\theta_r = \sum_{j|C_j \in C_R} p_j$ . Next note that the two continuations of the tree from the root node of  $\mathcal{M}$ , denoted by  $\mathcal{M}'$  and  $\mathcal{M}''$ , are themselves uBMPT models on the categories in  $C_L$  and  $C_R$ , respectively, whose root nodes are the left and right children of the root of  $\mathcal{M}$ . One proceeds to assign values to the parameters at these root nodes in the same way above. The process

continues until only leaves remain. The result is a numerical assignment of all  $J - 1$  parameters in  $\mathcal{M}$ . Now each category  $C_j$  has a unique branch in the uBMPT whose probability is given from equation (2) as a product of the assigned parameters and their complements (one minus a parameter) that occur on the branch. This product after canceling matching denominators and numerators will equal  $p_j$ . Hence  $\mathbf{p} \in \mathbf{\Lambda}(\mathcal{M})$  and therefore  $\mathbf{\Lambda}_J \subseteq \mathbf{\Lambda}(\mathcal{M})$ .  $\square$

The proposition implies that all uBMPT models with  $J$  categories have the same  $C_{\text{FIA}}$  (and  $C_{\text{NML}}$ ) complexity value, which is also the maximal complexity value for any MPT model with  $J$  categories. This value is given by  $C_{\text{FIA}} = \frac{J-1}{2} \ln \frac{N}{2\pi} + \frac{J}{2} \ln \pi - \ln \Gamma(\frac{J}{2})$  (see Rissanen, 1996; Grünwald, Myung & Pitt, 2005, Chapter 16), which is a *non-linear* function of the number of parameters  $J$  due to the presence of  $\ln \Gamma(\frac{J}{2})$ . In Grünwald, Myung & Pitt (figure 16.3 of 2005, Chapter 16), both  $C_{\text{FIA}}$  and  $C_{\text{NML}}$  are plotted against  $J$ . Both curves are concave, increasing slower as  $J$  increases, different from the complexity measure in BIC and AIC, which are linear in  $J$ .

*Effects of Combining Response Categories on Complexity*

The following proposition concerns what happens to  $C_{\text{NML}}$  of an BMPT model when two or more of response categories are combined into one.

**Proposition 3.** *For a given MPT model  $\mathcal{A}$ , if some of its categories are combined to create a new model  $\mathcal{B}$ , then we have  $C_{\text{NML},\mathcal{B}} \leq C_{\text{NML},\mathcal{A}}$ . The equality holds if and only if the probabilities associated with the to-be-combined categories, say  $p_k(\boldsymbol{\theta})$ ,  $k = 1, 2, \dots, K$ , satisfy the relationship  $p_k = c_k p(\boldsymbol{\theta})$ , where  $c_k$  are constants not depending on parameter. The same conclusion holds for  $C_{\text{FIA}}$  if both models are identified.*

*Proof.* We first prove the NML part. Suppose the probability mass function of the original model is:  $P_{\mathcal{A}}(\bar{\mathbf{n}}, \mathbf{n}|\boldsymbol{\theta})$  where  $\mathbf{n} = (n_1, n_2, \dots, n_k)$  is the vector of counts in the  $k$  categories to be combined, and  $\bar{\mathbf{n}} = (n_{k+1}, \dots, n_J)$  includes counts of the rest categories. Denote by  $\mathbf{1}_k$  the column vector of  $k$  1's. The probability mass function of the new model is  $P_{\mathcal{B}}(\bar{\mathbf{n}}, n_0) = \sum_{\mathbf{n}\mathbf{1}_k=n_0} P_{\mathcal{A}}(\bar{\mathbf{n}}, \mathbf{n}|\boldsymbol{\theta})$ , which follows from the law of total probability. Summing over both sides of the equation, we have

$$\begin{aligned} C_{\text{NML},\mathcal{A}} &= \sum_{\bar{\mathbf{n}}\mathbf{1}+n_0=N} \sum_{\mathbf{n}\mathbf{1}=n_0} P_{\mathcal{A}}(\bar{\mathbf{n}}, \mathbf{n}|\hat{\boldsymbol{\theta}}_{(\bar{\mathbf{n}},\mathbf{n})}) \\ &\geq \sum_{\bar{\mathbf{n}}\mathbf{1}+n_0=N} \sum_{\mathbf{n}\mathbf{1}=n_0} P_{\mathcal{A}}(\bar{\mathbf{n}}, \mathbf{n}|\hat{\boldsymbol{\theta}}_{(\bar{\mathbf{n}},n_0)}) \\ &= \sum_{\bar{\mathbf{n}}\mathbf{1}+n_0=N} P_{\mathcal{B}}(\bar{\mathbf{n}}, n_0|\hat{\boldsymbol{\theta}}_{(\bar{\mathbf{n}},n_0)}) = C_{\text{NML},\mathcal{B}} \end{aligned}$$

where  $\hat{\boldsymbol{\theta}}_{(\bar{\mathbf{n}}, \mathbf{n})}$  and  $\hat{\boldsymbol{\theta}}_{(\bar{\mathbf{n}}, n_0)}$  denote the MLEs obtained from  $P_{\mathcal{A}}$  and  $P_{\mathcal{B}}$ , respectively. The equality holds if and only if

$$\forall \bar{\mathbf{n}}, \mathbf{n} \quad P_{\mathcal{A}}(\bar{\mathbf{n}}, \mathbf{n} | \hat{\boldsymbol{\theta}}_{(\bar{\mathbf{n}}, \mathbf{n})}) = P_{\mathcal{A}}(\bar{\mathbf{n}}, \mathbf{n} | \hat{\boldsymbol{\theta}}_{(\bar{\mathbf{n}}, n_0)}) \quad (8)$$

or the maximizer of  $P_{\mathcal{A}}$  is a function of  $\mathbf{n}$  through  $n_0$  only. Note  $P_{\mathcal{A}}(\bar{\mathbf{n}}, \mathbf{n} | \boldsymbol{\theta}) = P_{\mathcal{B}}(\bar{\mathbf{n}}, n_0 | \boldsymbol{\theta}) P(\mathbf{n} | n_0, \boldsymbol{\theta})$ . Condition (8) is equivalent to  $P(\mathbf{n} | n_0, \boldsymbol{\theta})$  does not depend on  $\boldsymbol{\theta}$ . Further note that  $P(\mathbf{n} | n_0, \boldsymbol{\theta}) \propto \prod_i (p_i(\boldsymbol{\theta}) / p_0(\boldsymbol{\theta}))^{n_i}$  and we can see (8) is equivalent to  $p_i(\boldsymbol{\theta}) = c_i p_0(\boldsymbol{\theta})$ .

For the FIA part, we note from Lemma 2 that the Fisher information matrix of model  $\mathcal{A}$ ,  $\mathbf{I}^{\mathcal{A}}$ , is given by  $I_{rs}^{\mathcal{A}} = \sum_j p_j^{-1} P_{jr} P_{js}$ , where  $P_{js} = \frac{\partial p_j}{\partial \theta_s}$ . Similarly, that of model  $\mathcal{B}$  is given by  $I_{rs}^{\mathcal{B}} = p_0^{-1} P_{0r} P_{0s} + \sum_{j=k+1}^J p_j^{-1} P_{jr} P_{js}$ . Let  $\mathbf{a} = (a_1, a_2, \dots, a_S)'$  be an arbitrary nonzero vector, and let  $\alpha_j = \sum_s a_s P_{js}$ ,  $j = 0, 1, \dots, J$ . We have

$$\begin{aligned} \mathbf{a}^T \mathbf{I}^{\mathcal{A}} \mathbf{a} &= \sum_j p_j^{-1} \alpha_j^2 = \sum_{j=1}^k p_j^{-1} \alpha_j^2 + \sum_{j=k+1}^J p_j^{-1} \alpha_j^2 \\ &\geq \left( \sum_{j=1}^k p_j \right)^{-1} \left( \sum_{j=1}^k \alpha_j \right)^2 + \sum_{j=k+1}^J p_j^{-1} \alpha_j^2 = p_0^{-1} \alpha_0^2 + \sum_{j=k+1}^J p_j^{-1} \alpha_j^2 = \mathbf{a}^T \mathbf{I}^{\mathcal{B}} \mathbf{a} \end{aligned}$$

which implies  $|\mathbf{I}^{\mathcal{A}}| \geq |\mathbf{I}^{\mathcal{B}}|$ . The FIA inequality can be proved by applying equation (6). The equality holds if and only if  $(p_j)^{-1} (p_j^{-1} \alpha_j^2) = (p_j^{-1} \alpha_j)^2$  does not depend on  $j$ , or  $p_j^{-1} P_{js}$  does not depend on  $j$  for all  $s$ , which is equivalent to  $\frac{\partial (\ln p_j - \ln p_i)}{\partial \theta_s} = p_j^{-1} P_{js} - p_i^{-1} P_{is} = 0$ , or  $p_j(\boldsymbol{\theta}) = c_{ij} p_i(\boldsymbol{\theta})$ , for some constants  $c_{ij}$  not depending on  $\boldsymbol{\theta}$ .

It should be noted that because equation (6) is used, the identification of both models is required. When either model is not identified, the inequality still holds for the quantity on the right hand side of equation (6), but in this case, the  $C_{\text{FIA}}$  value is defined instead as that of an equivalent and identified model, which may not satisfy the inequality.  $\square$

Figure 3 illustrates the proposition. In the top panel, for the model on the left, two of its leaves representing categories  $C_1$  and  $C_2$  are combined to represent a new category  $C_0$ . According to the proposition, the resulting model, shown on the right, will have a smaller complexity value than the original one. For the two models on the bottom panel of the figure, they are both equally complex. This is because the probabilities of two categories  $C_1$  and  $C_2$  of the

model on the left are equal (i.e.,  $pq$ ), with the constant multiplication factor being equal to one.

Figure 4 provides another illustration of Proposition 3 this time for a well studied MPT model of pair-clustering (Batchelder & Riefer, 1986). The pair clustering experiment involves studying a list of two types of items, paired items and singletons, followed by free recall of the list. The model posits three parameters:  $c$ , probability of pairs being clustered and stored in memory;  $r$ , probability of a stored pair being retrieved from memory;  $u$ , 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 the studied pairs, response category  $E_2$  indicating recalling non-adjacently both items of the pairs, response category  $E_3$  indicating recalling only one item, and so on. Figure 5 shows  $C_{\text{NML}}$  and  $C_{\text{FIA}}$  curves as a function of sample size for the model (two upper curves). The two trails closely parallel each other, indicating that  $C_{\text{FIA}}$  provides a good approximation of  $C_{\text{NML}}$  in this case.

Both of categories  $E_4$  and  $E_5$  represent unsuccessful retrieval and as such, cannot be distinguished from each other based on observed responses. It is therefore necessary to combine the two categories into one. What would happen to model complexity if this is done? As shown by the two lower curves in Figure 5, combining the categories has reduced complexity as predicted by Proposition 3. Since the number of parameters (3) remain unchanged, the reduction in complexity must be due to the difference in functional forms between the two models, one uncombined and the other combined.<sup>2</sup>

*Effects of Combining Trees on Complexity*

As we noted in the section of Formal Definition of BMPT Models, the structural component of BMPT models satisfy two recursive properties. In this section we will exploit these recursive properties and focus on the situation in which two or more BMPT models are combined to form a new BMPT model. We are interested in knowing how model complexity is affected by such operations. All results in this subsection is based on Lemma 3 in the Appendix, which gives the form of Fisher information matrix of combined trees in general. We begin with the simplest situation in which two BMPT models are combined with a single binomial parameter.

**Proposition 4.** *Let  $\mathcal{A}_1$  and  $\mathcal{A}_2$  be two BMPT models with disjoint parameter sets  $\Theta_1$  and  $\Theta_2$  and disjoint category sets. Suppose  $\mathcal{C} = p\mathcal{A}_1\mathcal{A}_2$  (see the top*

---

<sup>2</sup>It is worth noting that the two models assume multinomial distributions with different numbers of categories and cannot be applied to the same data set, so direct contrast of their complexity does not have implications for model selection.

panel of Figure 6) where  $p \notin \Theta_1 \cup \Theta_2$ . Then

$$C_{\text{FIA},\mathcal{C}}(N) = C_{\text{FIA},\mathcal{A}_1}(N) + C_{\text{FIA},\mathcal{A}_2}(N) + \frac{1}{2} \ln \frac{N}{2\pi} + \ln \beta \left( \frac{S_1 + 1}{2}, \frac{S_2 + 1}{2} \right)$$

where  $S_1$  and  $S_2$  are the number of parameters in model  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , respectively, and  $\beta$  is the beta function.

This proposition follows directly from Lemma 3 in the Appendix and equation (6). Note that the third and fourth terms of the foregoing equation reflects an increase in complexity due to the binomial parameter  $p$  that joins the two trees. If we were to assume that the addition of the binomial parameter independently contributes to overall complexity, an expected increase in complexity would be  $\frac{1}{2} \ln \frac{N}{2\pi} + \ln \beta \left( \frac{1}{2}, \frac{1}{2} \right)$ , the FIA complexity of binomial distribution according to equation (6). Since  $\ln \beta \left( \frac{S_1+1}{2}, \frac{S_2+1}{2} \right) < \ln \beta \left( \frac{1}{2}, \frac{1}{2} \right)$ , the complexity of tree  $\mathcal{C}$  with  $(S_1 + S_2 + 1)$  parameters would be smaller than the sum of complexities of the two individual trees and the binomial model, in contrast to the AIC and BIC complexity measures which are additive in this case.

The above proposition shows that the FIA complexity value is generally not additive and will be less than the sum of all three parts of the model in a very simple situation. In the following proposition, we further pursue this decrease in complexity and present a result in a more general setting.

**Proposition 5.** *Let  $\mathcal{B}_k^r$ ,  $r = 1, 2, \dots, R_k$ ,  $k = 1, 2, \dots, K$ , be  $\sum_k R_k$  BMPT models that satisfy the following condition: for all  $k$ , the  $R_k$  models  $\mathcal{B}_k^r$ ,  $r = 1, 2, \dots, R_k$  are functionally identical with identical parameter assignment (and therefore have identical complexity value  $C_{\text{FIA},k}$ ). In addition, let  $\mathcal{A}$  be a BMPT model with parameter set  $\Theta_{\mathcal{A}}$  and complexity value  $C_{\text{FIA},\mathcal{A}}$ . Suppose all the  $1 + \sum_k R_k$  models have disjoint category sets and the  $K + 1$  parameter sets  $\Theta_{\mathcal{A}}, \Theta_k, k = 1, 2, \dots, K$  are disjoint. If a new BMPT model  $\mathcal{C}$  is constructed by replacing  $\sum_k R_k$  of  $\mathcal{A}$ 's categories by the  $\sum_k R_k$  models  $\mathcal{B}_k^r$  respectively, then we have*

$$C_{\text{FIA},\mathcal{C}}(N) \leq C_{\text{FIA},\mathcal{A}}(N) + \sum_k C_{\text{FIA},k}(N) + \frac{1}{2} \sum_k S_k \ln S_k - \frac{1}{2} \left( \sum_k S_k \right) \ln \left( \sum_k S_k \right)$$

where  $S_k$  is the number of parameters in  $\mathcal{B}_k^r$ .

*Proof.* From Lemma 3 in the Appendix we can see  $\mathbf{I}_{\mathcal{C}}$  is a block diagonal matrix given by  $\text{diag}\{\mathbf{I}_{\mathcal{A}}, p_1 \mathbf{I}_1, p_2 \mathbf{I}_2, \dots, p_K \mathbf{I}_K\}$ , where for any  $k = 1, 2, \dots, K$ ,  $p_k$  denotes the total probability of the  $R_k$  categories in model  $\mathcal{A}$  replaced by  $\mathcal{B}_k^r$ ,

$r = 1, 2, \dots, R_k$ . Taking the determinant, we have  $|\mathbf{I}_C| = |\mathbf{I}_A| \prod_k (|\mathbf{I}_k| p_k^{S_k})$ . Apply equation (6), note the parameter sets of the  $K + 1$  models are disjoint, and we have

$$\begin{aligned} \int |\mathbf{I}_C(\boldsymbol{\theta})|^{\frac{1}{2}} d\boldsymbol{\theta} &= \left( \int |\mathbf{I}_A(\boldsymbol{\theta}_A)|^{\frac{1}{2}} \prod_k p_k(\boldsymbol{\theta}_A)^{\frac{S_k}{2}} d\boldsymbol{\theta}_A \right) \prod_k \int |\mathbf{I}_k(\boldsymbol{\theta}_k)|^{\frac{1}{2}} d\boldsymbol{\theta}_k \\ &\leq \prod_k \left( \frac{S_k}{\sum_j S_j} \right)^{\frac{S_k}{2}} \left( \int |\mathbf{I}_A(\boldsymbol{\theta}_A)|^{\frac{1}{2}} d\boldsymbol{\theta}_A \right) \prod_k \int |\mathbf{I}_k(\boldsymbol{\theta}_k)|^{\frac{1}{2}} d\boldsymbol{\theta}_k \end{aligned}$$

The inequality follows from the fact that  $\prod_k p_k^{S_k}$  with  $\sum_k p_k \leq 1$  reaches its maximum when  $p_k = \frac{S_k}{\sum_k S_k}$ . Further taking logarithm of both sides and applying equation (6) completes the proof.  $\square$

The model constructed in this proposition is described in the second panel of Figure 6. The proposition shows that the complexity of the new model is always less than the sum of those of its parts, and the amount decrease in complexity has a lower bound of  $-\frac{1}{2} \sum_k S_k \ln S_k + \frac{1}{2} (\sum_k S_k) \ln (\sum_k S_k) \geq 0$  (this inequality can be verified by exponentiating both sides).

The following proposition summarizes the same result for NML complexity.

**Proposition 6.** *Assuming the same conditions in Proposition (5) except for the requirement of disjoint parameter sets, we have*

$$C_{\text{NML},\mathcal{C}}(N) \leq C_{\text{NML},\mathcal{A}}(N) + \max_{\sum_k n_k = N} \sum_k C_{\text{NML},k}(n_k) \leq C_{\text{NML},\mathcal{A}}(N) + \sum_k C_{\text{NML},k}(N)$$

*Proof.* We only prove the proposition for the case where all categories of model  $\mathcal{A}$  are replaced by some  $\mathcal{B}_k^r$ . The other case where some categories of model  $\mathcal{A}$  are also categories of model  $\mathcal{C}$  can be easily taken care of by allowing some  $\mathcal{B}_k^r$  be a degenerated tree with no parameter and a single category with category probability 1.

We index the categories of tree  $\mathcal{A}$  by  ${}^r_k$ ,  $k = 1, 2, \dots, K$ ,  $r = 1, 2, \dots, R_k$ , according to the index of  $\mathcal{B}_k^r$  attached to it. We also index the counts of tree  $\mathcal{C}$  by  $n_{kj}^r$ ,  $j = 1, 2, \dots, J_k$ , as they are always counts in some tree  $\mathcal{B}_k^r$ . Let  $m_k^r = \sum_{j=1}^{J_k} n_{kj}^r$  be the sum of counts of tree  $\mathcal{B}_k^r$ ,  $o_{kj} = \sum_{r=1}^{R_k} n_{kj}^r$  be the sum of the corresponding counts in category  $j$  across the  $R_k$  trees  $\mathcal{B}_k^r$  ( $r = 1, 2, \dots, R_k$ ), and  $l_k = \sum_{r=1}^{R_k} m_k^r = \sum_{j=1}^{J_k} o_{kj}$ . Vector  $\mathbf{n}$  involves all counts of  $\mathcal{C}$ , and  $\mathbf{n}_k^r$  involves all counts that share the same indices  ${}^r_k$ . Other vectors such as  $\mathbf{m}$ ,  $\mathbf{n}_{kj}$ ,  $\mathbf{n}_k$  and  $\mathbf{o}_k$  are similarly defined.

We have

$$\begin{aligned} \sum_{\mathbf{n}} P(\mathbf{n}|\hat{\boldsymbol{\theta}}_{\mathbf{n}}) &= \sum_{\mathbf{m}} P_{\mathcal{A}}(\mathbf{m}|\hat{\boldsymbol{\theta}}_{\mathbf{m}}) \sum_{\mathbf{n}|\mathbf{m}} P(\mathbf{n}|\mathbf{m}, \hat{\boldsymbol{\theta}}_{\mathbf{n}}) \\ &\leq \sum_{\mathbf{m}} P_{\mathcal{A}}(\mathbf{m}|\hat{\boldsymbol{\theta}}_{\mathbf{m}}^A) \sum_{\mathbf{n}|\mathbf{m}} \prod_k \prod_r P_k(\mathbf{n}_k^r | m_k^r, \hat{\boldsymbol{\theta}}_{\mathbf{n}_k}^k) \end{aligned}$$

where  $\sum_{\mathbf{m}}$  denotes  $\sum_{\sum_{k,r} m_k^r = N}$ ,  $\sum_{\mathbf{n}}$  denotes  $\sum_{\sum_{k,r,j} n_{kj}^r = N}$ ,  $\sum_{\sum_j n_{kj}^r = m_k^r}$ , and  $\hat{\boldsymbol{\theta}}_{\mathbf{n}_k}^k$  maximizes  $\prod_r P_k(\mathbf{n}_k^r | m_k^r, \boldsymbol{\theta})$ . Note the conditional distribution  $P_k$  depends only on  $k$  but not  $r$  since  $\mathcal{B}_k^r$ ,  $r = 1, 2, \dots, R_k$  have the same tree structure and parameter assignment.

To simplify the last term, we note

$$\begin{aligned} \prod_r P_k(\mathbf{n}_k^r | m_k^r, \boldsymbol{\theta}^k) &= \prod_r \left\{ \binom{m_k^r}{n_{k1}^r, n_{k2}^r, \dots, n_{kJ_k}^r} \prod_j p_{kj}(\boldsymbol{\theta}^k)^{n_{kj}^r} \right\} \\ &= \left\{ \prod_r \binom{m_k^r}{n_{k1}^r, n_{k2}^r, \dots, n_{kJ_k}^r} \right\} \prod_j p_{kj}(\boldsymbol{\theta}^k)^{o_{kj}} \\ &= \frac{\prod_r \binom{m_k^r}{n_{k1}^r, n_{k2}^r, \dots, n_{kJ_k}^r}}{\binom{l_k}{o_{k1}, o_{k2}, \dots, o_{kJ_k}}} P_k(\mathbf{o}_k | l_k, \boldsymbol{\theta}^k) \end{aligned}$$

We can see immediately that  $\hat{\boldsymbol{\theta}}_{\mathbf{n}_k}^k = \hat{\boldsymbol{\theta}}_{\mathbf{o}_k}^k$  depends on  $\mathbf{o}_k$  only. Back to equation (9) we have

$$\begin{aligned} \sum_{\mathbf{n}} P(\mathbf{n}|\hat{\boldsymbol{\theta}}_{\mathbf{n}}) &\leq \sum_{\mathbf{m}} P_{\mathcal{A}}(\mathbf{m}|\hat{\boldsymbol{\theta}}_{\mathbf{m}}^A) \prod_k \sum_{\mathbf{n}_k|\mathbf{m}_k} \frac{\prod_r \binom{m_k^r}{n_{k1}^r, n_{k2}^r, \dots, n_{kJ_k}^r}}{\binom{l_k}{o_{k1}, o_{k2}, \dots, o_{kJ_k}}} P_k(\mathbf{o}_k | l_k, \boldsymbol{\theta}_{\mathbf{o}_k}^k) \\ &\leq \sum_{\mathbf{m}} P_{\mathcal{A}}(\mathbf{m}|\hat{\boldsymbol{\theta}}_{\mathbf{m}}^A) \prod_k \left\{ \sum_{\mathbf{o}_k} P_k(\mathbf{o}_k | l_k, \hat{\boldsymbol{\theta}}_{\mathbf{o}_k}^k) \sum_{\mathbf{n}_k|\mathbf{m}_k, \mathbf{o}_k} \frac{\prod_r \binom{m_k^r}{n_{k1}^r, n_{k2}^r, \dots, n_{kJ_k}^r}}{\binom{l_k}{o_{k1}, o_{k2}, \dots, o_{kJ_k}}} \right\} \\ &= \sum_{\mathbf{m}} P_{\mathcal{A}}(\mathbf{m}|\hat{\boldsymbol{\theta}}_{\mathbf{m}}^A) \prod_k \sum_{\mathbf{o}_k} P_k(\mathbf{o}_k | l_k, \hat{\boldsymbol{\theta}}_{\mathbf{o}_k}^k) \\ &\leq \left\{ \sum_{\mathbf{m}} P_{\mathcal{A}}(\mathbf{m}|\hat{\boldsymbol{\theta}}_{\mathbf{m}}^A) \right\} \max_1 \prod_k \sum_{\mathbf{o}_k} P_k(\mathbf{o}_k | l_k, \boldsymbol{\theta}_{\mathbf{o}_k}^k) \end{aligned}$$

Taking natural logarithm to both sides completes the proof.  $\square$

Regarding Proposition 5 and Proposition 6, we have the following remarks. First, if all categories of  $\mathcal{A}$  are replaced and all the attached models are statistically equivalent, the middle upper bound can be sharpened to  $C_{\text{NML},\mathcal{A}}(N) + kC_{\text{NML},\mathcal{B}}(N/k)$ . Second, when sample size is large,  $C_{\text{FIA}}$  and  $C_{\text{NML}}$  will be close to each other given the asymptotic nature of  $C_{\text{FIA}}$ . However, in this case the bound given by Proposition 5 is much sharper than the second upper bound in Proposition 6, because the former bound for  $C_{\text{FIA}}$  makes use of the asymptotic order of NML complexity. Third, the second upper bound will be achieved only when  $K = 1$  and all categories of model  $\mathcal{A}$  was replaced. This same condition would be needed for the lower bound of  $C_{\text{FIA}}$  complexity difference to be 0. In this case, for both NML and FIA, the complexity of the new model are the sum of those of model  $\mathcal{A}$  and  $\mathcal{B}_1$ . This special additive case is summarized in the following proposition.

**Proposition 7.** *Suppose trees  $\mathcal{A}$  and  $\mathcal{B}^j$ ,  $j = 1, 2, \dots, J_{\mathcal{A}}$ , represent  $J_{\mathcal{A}} + 1$  BMPT models with disjoint sets of categories. Let the  $J_{\mathcal{A}}$  models  $\mathcal{B}^j$ ,  $j = 1, 2, \dots, J_{\mathcal{A}}$  be functionally identical with identical parameter assignment. Suppose that a new tree  $\mathcal{C}$  is created by replacing the  $J_{\mathcal{A}}$  response categories of tree  $\mathcal{A}$  by the  $J_{\mathcal{A}}$  trees  $\mathcal{B}^j$ . Then the complexity of tree  $\mathcal{C}$  is equal to the sum of complexities of  $\mathcal{A}$  and  $\mathcal{B}$ , that is,  $C_{\text{NML},\mathcal{C}} = C_{\text{NML},\mathcal{A}} + C_{\text{NML},\mathcal{B}}$  and  $C_{\text{FIA},\mathcal{C}} = C_{\text{FIA},\mathcal{A}} + C_{\text{FIA},\mathcal{B}}$ .*

The third panel of Figure 6 illustrates the proposition. Note that the additive-complexity rule of the proposition relies on two conditions: (a) the two models do *not* share common parameters; and (b) one model is added to *every* leaf of the other model. The failure to satisfy any of these conditions invalidates the additivity rule, as has been demonstrated in Proposition 4, 5 and 6. This is in sharp contrast to AIC/BIC viewpoint, in which the complexity is linear in the number of parameters and is therefore always additive.

Now consider the case in which the same tree structure is added recursively to every one of its *own* leaves. The following proposition shows that the complexity increases as a logarithmic function of the total number of layers.

**Proposition 8.** *Let  $\mathcal{A}_k$  be a collection of functionally identical BMPT models with identical parameter assignment. The sequence of trees  $\mathcal{B}_d$ ,  $d = 1, 2, \dots$  is constructed as follows:*

1.  $\mathcal{B}_1 = \mathcal{A}_1$ , which has  $J_{\mathcal{A}}$  categories.
2.  $\mathcal{B}_{d+1}$  is constructed by replacing the  $(J_{\mathcal{A}})^d$  categories of  $\mathcal{B}_d$  by  $\mathcal{A}_k$ ,  $k = 1, 2, \dots, (J_{\mathcal{A}})^d$ , respectively.

*Then the  $C_{\text{FIA},\mathcal{B}_d} = C_{\text{FIA},\mathcal{A}} + \frac{S_{\mathcal{A}}}{2} \ln d$ , where  $S_{\mathcal{A}}$  is the number of parameters in tree  $\mathcal{A}$ . Note  $d$  is the number of layers formed as a result of the recursive operation.*

This proposition follows directly from Lemma 3 in the Appendix. The bottom panel of Figure 6 illustrates the proposition. The significance of this proposition lies in that it enables us to construct two MPT models that have the same number of parameters but differ greatly in their complexity values. This is explained below.

Given the way tree  $\mathcal{B}_d$  is constructed, many of its leaves will have the same probability expression. Therefore, according to Proposition 3, we can combine those leaves and obtain another tree  $\mathcal{C}$  with fewer leaves but with the same complexity as tree  $\mathcal{B}_d$ . Now suppose that we create a new tree  $\mathcal{D}$  by splitting the categories of model  $\mathcal{B}_1$  with constant conditional probabilities such that it has the same number of categories as tree  $\mathcal{C}$ . The same proposition implies that the splitting operation does not change complexity. In the end, both trees  $\mathcal{C}$  and  $\mathcal{D}$  would have the same number of parameters and the same number of response categories and yet, according to Proposition 8, the complexity difference between the two trees is equal to  $\frac{S_A}{2} \ln d$ , which can be quite large for large  $S_A$  and  $d$ .

To conclude the present section, the above analytical results on the complexity of BMPT models provide an illuminating understanding of various ways that tree structure can contribute to the MDL complexity. It is worth noting that we proved most of the analytical results by taking advantage of the recursive nature of the BMPT class, as defined in the section of Formal Definition of BMPT Models. For the great majority of MPT models of practical interest, however, the complexity measures  $C_{\text{NML}}$  and  $C_{\text{FIA}}$  do not usually have analytical form solutions. As such, the complexity must be calculated numerically on computer, to which we now turn our discussion in the following section.

### Computational Issues

In this section we discuss practical implementation issues concerning the MDL complexity, which can be non-trivial to compute. Recall that to compute  $C_{\text{NML}}$ , one must first obtain the maximized likelihood for all possible data sets that could be observed in an experiment. Given the fact that analytic expression for maximum likelihood is generally not available for MPT models, computing  $C_{\text{NML}}$  directly would be out of question unless the sample size is small or the models are simple enough to yield the maximum likelihood in analytic form. Given this, the next best thing to do would be computing  $C_{\text{FIA}}$  instead. Note that  $C_{\text{FIA}}$  represents an asymptotic approximation of  $C_{\text{NML}}$  but does not requiring calculating maximum likelihoods. In the rest of the section, we give a Monte Carlo algorithm for computing  $C_{\text{FIA}}$ .

*A Monte Carlo Algorithm*

A key step in computing  $C_{\text{FIA}}$  is the evaluation of the integral in Equation (6) via Monte Carlo. The rationale of Monte Carlo integration lies as follows. The integral  $\int_{\Theta} \sqrt{|\mathbf{I}(\boldsymbol{\theta})|} d\boldsymbol{\theta}$  can be written as the expected value of  $h(\boldsymbol{\xi}) = \sqrt{|\mathbf{I}(\boldsymbol{\xi})|}/\pi(\boldsymbol{\xi})$ , where random vector  $\boldsymbol{\xi}$  follows a distribution with density  $\pi(\boldsymbol{\xi})$ . Monte Carlo method then approximates the expectation by the sample average  $\frac{1}{N} \sum_{i=1}^N h(\boldsymbol{\xi}_i)$ , where  $\boldsymbol{\xi}_i$ 's are realizations of  $\boldsymbol{\xi}$ . Although the choice of density  $\pi$  is arbitrary, different choices of  $\pi$  may lead to different rate of convergence of the sample average to the integral. To choose an appropriate proposal distribution  $\pi(\boldsymbol{\xi})$ , we need the following proposition.

**Proposition 9.**  $\sqrt{|\mathbf{I}(\boldsymbol{\theta})|} < c \prod_{s=1}^S \frac{1}{\sqrt{\theta_s(1-\theta_s)}}$  for some constant  $c$ .

*Proof.* Consider the matrix  $\mathbf{I}(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})$ , where diagonal matrix  $\mathbf{D}$  has typical element  $D_{ss} = \theta_s(1-\theta_s)$ , and the elements in the Fisher information matrix  $\mathbf{I}(\boldsymbol{\theta})$  is given by Equation (10) in the Appendix. We can see

$$(\mathbf{ID})_{sr} = \sum_{j=1}^J \left\{ \sum_{i=1}^{I_j} p_{ij} \left( \frac{a_{ijs}}{\theta_s} - \frac{b_{ijs}}{1-\theta_s} \right) \right\} \left\{ \sum_{i=1}^{I_j} \frac{p_{ij}}{p_j} (a_{ijr}(1-\theta_r) - b_{ijr}\theta_r) \right\}$$

From the expression of  $p_{ij}$  given by (3) we know that for all  $i, j$  and  $s$ ,  $\frac{a_{ijs}p_{ij}}{\theta_s}$  and  $\frac{b_{ijs}p_{ij}}{1-\theta_s}$  are polynomials of the  $\theta_s$  and are therefore bounded on  $[0,1]$ . In addition,  $p_{ij}/p_j$  is also bounded, so  $(\mathbf{ID})_{sr}$  must also be bounded. The proposition follows immediately.  $\square$

This proposition has two implications. First, it implies that the integral is finite and our Monte Carlo computation is meaningful. Second, it implies that the choice of density  $\pi(\boldsymbol{\theta}) = \prod_{s=1}^S \frac{1}{\pi\sqrt{\theta_s(1-\theta_s)}}$  would lead to a bounded  $h(\boldsymbol{\theta})$  for all MPT models. This is desirable as it gives a finite Monte Carlo standard deviation of the estimate. It should be noted that uniform distribution over  $[0, 1]^S$  does not generally satisfy this requirement and the convergence of Monte Carlo algorithm can be very slow if  $\pi(\boldsymbol{\theta}) = 1$  is chosen.

*Models with Multiple Trees*

Because experiments often involve multiple treatments, most MPT models involves multiple trees, each representing a different treatment. It follows directly from Lemma 3 in the Appendix that the Fisher information matrix of

MPT models with multiple trees  $\mathcal{A}_k$  is the same as that of an MPT model with a single tree constructed by joining  $\mathcal{A}_k$ 's by constant multinomial probabilities  $c_k = \frac{N_k}{N}$ , where  $N_k$  is the sample size for tree  $\mathcal{A}_k$  and  $N$  is the total sample size. This Fisher information matrix is given by  $\sum_{k=1}^K c_k \tilde{\mathbf{I}}_k$ , where  $\tilde{\mathbf{I}}_k$  denotes the Fisher information matrix of tree  $\mathcal{A}_k$  extended to include all parameters in the model as defined in Lemma 3. This property can be exploited for the computation of models with multiple trees.

*Models with Inequality Constraints*

For models with inequality constraints on the parameters, the calculation of  $C_{\text{FIA}}$  is straightforward: one simply needs to restrict the integral in equation (6) to the constrained parameter space. To do this, the algorithm need to be modified to sample  $\xi_i$ 's from the restricted area of the parameter space proportional to proposal distribution  $\pi$ . The  $h(\boldsymbol{\theta})$  function need also be changed to incorporate the new normalizing constant of  $\pi$ . However, due to the symmetry of the proposal distribution, for most inequality constraints in practice such as the full or partial order relations on  $\Theta \cup \{\frac{1}{2}\}$ , the normalizing constant can be easily computed analytically.

In particular, if the original unconstrained model has some symmetry property, the ratio between the constrained and unconstrained integrals can be calculated analytically and a separate run of the Monte Carlo algorithm is not needed. For example, if a BMPT model involves  $k$  functionally identical trees representing  $k$  experimental treatments and parameters  $\theta_1, \theta_2, \dots, \theta_k$  are correspondent parameters for the treatments, a treatment effect on  $\theta$  represented by  $\theta_1 \leq \theta_2 \leq \dots \leq \theta_k$  would reduce the integral to its  $1/k!$ . However, if treatment effects are expected simultaneously on two parameters, this method would not work as the treatments are no longer symmetric after the placement of one of the effects. For example, if either  $D_1 > D_2$  or  $d_1 > d_2$  is assumed in 1HTM-6c (shown in figure (1)), its model complexity will reduce by  $\ln 2$ , as the two sources have symmetric roles in the model, but the inclusion of both at the same time may not reduce the complexity by  $2 \ln 2$ .

*Complexity of Source Monitoring Models*

To provide a concrete example of MDL complexity, here we compute  $C_{\text{FIA}}$  for some well-known MPT models of source monitoring. The MPT models of source monitoring in Figure 1 (also see Figure 2) are among the most widely studied classes of MPT models in cognitive psychology. Application of MDL-based model evaluation to this class of models is therefore of special interest.  $C_{\text{FIA}}$  was computed for the hierarchy of models shown on the right panel of

Figure 1. Each model is defined in terms of three tree structures, one for each item type, so three sample sizes ( $n_A, n_B, n_N$ ) are defined. In a typical source monitoring experiment, the sample sizes for the old items, A or B, are set to be the same ( $n_A = n_B \equiv n_O/2$ ), and only the ratio of new items to the total number of items ( $n_N/(n_O + n_N)$ ) varies from experiment to experiment.

Shown in Figure 7 are  $C_{FIA}$  complexity curves for six selected models, plotted as a function of the percentage of new items for the total sample size of  $n_O + n_N = 1000$ . (Among the eight models in Figure 1, models 7 and 6a are excluded from consideration as they are equivalent to models 6b and 5a, respectively (Batchelder & Riefer, 1990).) The first thing to note in the figure is that complexity is generally ordered according to the number of parameter. The two six-parameter models are the most complex, trailed by the three five-parameter models, and the four-parameter model is the simplest. Also note that among models with the same number of parameters, their complexity values can differ significantly from one another, sometimes even greater than the complexity difference due to the difference in the number of parameters. The case in point is the three models, 4, 5b and 5c. At the 50% value of new items, the complexity difference between model 5c and model 5b is equal to about 1.50, which is greater than the complexity difference (0.62) between models 5b and 4.<sup>3</sup> These results, taken together, again, demonstrate that model complexity is determined not only by the number of parameters but also importantly, by functional form, and sometimes even more so by the latter.

### Conclusions

Model complexity is an integral and key concept in the evaluation and selection of quantitative models of cognition. In this paper we have explored the issue of model complexity in multinomial process tree (MPT) modeling with a special focus on the effects of tree structure on complexity for MPT models. The particular approach we took in the present investigation is that of minimum description length (MDL). The primary contributions of the present study are a series of Propositions we proved concerning the properties of the MDL com-

---

<sup>3</sup>We comment briefly on the unusual crossovers of complexity curves observed at extreme values of the percentage of new items. Note in the figure that model 4, which is nested within model 5b, has greater complexity values than model 5b for the percentage of new items being greater than 95% or smaller than 5%. This violates the complexity order relationship that is supposed to hold between nested models, as stated in Section 3.1.1. Similar “illegitimate” crossovers are also observed between another pair of nested models, 5c and 6c. The observation of these crossovers is not entirely surprising. As noted earlier,  $C_{FIA}$  is an asymptotic approximation of  $C_{NML}$  and therefore, can sometimes exhibit abnormal complexity order relationships, due to the inaccuracy of the approximation (Navarro, 2004).

plexity of this class of models and a general algorithm for the computation of  $C_{\text{FIA}}$ .

Speaking of model complexity, recall that complexity refers to the range of data patterns a model can provide good fits to, in the sense that a complex model fits well a wider range of data patterns than a simpler model. This idea is formalized in the NML complexity measure,  $C_{\text{NML}}$ , which is equal to the logarithmic value of the sum of best fits that the model can provide, by varying its parameter values, for all potential data patterns. Here we highlight a few important insights we have gained about complexity of MPT models from our analytic investigations. First of all, according to the NML complexity measure, what matters in measuring a model's complexity is not the apparent complications of its tree structure (i.e., "functional form") or the number of its parameters but instead the "size" of the family of probability distributions indexed by the model's parameters. Second, insofar as the same family of probability distributions are indexed, complexity remains unchanged regardless of how the model is parameterized. Third, if other things are equal, the more distinct response categories a model assumes, generally, the more complex the model is (Proposition 3). Fourth, complexity is, in general, non-additive with respect to combining two or more models of disjoint parameter sets (Proposition 4, 5 and 6). Finally and related, tree structure can significantly contribute to model complexity, sometimes even more than the number of parameters. As an extreme example of this, it is possible to construct two MPT models with the same number of parameters yet with complexity values different greatly, as described in Proposition 8.

The next step is to implement the MDL complexity measures and apply FIA and NML criteria to addressing actual model selection problems in the field of MPT modeling. The work is currently underway and results will be reported elsewhere.

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Kotz and Johnson (1992): *Breakthroughs in Statistics*. NY: Springer Verlag.
- Balasubramanian, V. (1996). A geometric formulation of Occam's razor for inference of parametric distributions. Available as preprint number adaporg/9601001 from <http://xyz.lanl.gov/> and as Princeton University Physics Preprint PUPT-1588.
- Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, *9*, 349-368.
- Batchelder, W. H. (1990). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331-344.
- Batchelder, W.H. (In Press). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. Embretson and J. Roberts (Eds.). *New directions in psychological measurement with model based approaches*. APA Books.
- Batchelder, W. H. and Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, *87*, 375-397
- Batchelder, W. H. and Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, *39*, 129-149.
- Batchelder, W. H. and Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548-564.
- Batchelder, W. H. and Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review*, *6*, 57-86.
- Batchelder, W.H., & Riefer, D.M. (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, D.C.: American Psychological Association Books.

- Batchelder, W. H., Riefer, D. M. and Hu, X. (1994). Measuring memory factors in source monitoring: Reply to Kinshla. *Psychological Review*, 101, 172-176.
- Bayen, U. J., Murnane, K. and Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models for source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 197-215.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer Verlag.
- Casella, G. and Berger, R. L. (2001) *Statistical Inference. 2nd edition* Duxbury Press;
- Chechile, R.A. (2004). New models for the Chechille-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. 51-79), Washington, D.C.: American Psychological Association Books.
- Grünwald, P. (2000). Model selection based on Minimum Description Length. *Journal of Mathematical Psychology*, 44, 133-152.
- Grünwald, P. (2007). *The Minimum Description Length Principle*. MIT Press.
- Grünwald, P., Myung, I. J., and Pitt, M. A. (2005). *Advances in Minimum Description Length: Theory and Applications*. MIT Press.
- Hu, X. and Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59, 21-47.
- Hu, X. and Phillips, G.A. (1999). GPT.EXE: A powerful tool for visualizing and analysis of general processing tree models. *Behavior Research Methods, Instruments, & Computers*, 31, 220-234.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Knapp, B., & Batchelder, W.H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, 2004, 215-229.
- Lawley, D. N. and Maxwell, A. E. (1971). *Factor analysis as a statistical method*. Elsevier New York.

- Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, *45*, 131-148.
- Lee, M. D. and 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., Balasubramanian, V. and Pitt, M. A. (2000). Counting probability distributions: differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170-11175.
- Myung, I. J., Forster, M. R. and 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. and Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167-179.
- Myung, I. J., and Pitt, M. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79-95.
- Myung, I. J., Pitt, M. and Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review*, *14*(6), 1043-1050.
- Navarro, D. J. (2004). A note on the applied use of MDL approximations. *Neural Computation*, *16*, 1763-68.
- Navarro, D. J. and Lee, M. D. (2004). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychonomic Bulletin & Review*, *11*, 961-974.
- Navarro, D. J., Pitt, M. A. and Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47-84.
- Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Psychology*, *6*, 421-425.
- Pitt, M. A., Myung, I. J. and Zhang, S (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472-491.

- Purdy, B.P. and Batchelder, W.H. (in press). A context free language for binary multinomial processing tree models. *Journal of Mathematical Psychology*.
- Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer-Verlag.
- Riefer, D. M. and Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318-339.
- Riefer, D.M. and Batchelder, W.H. (1991). Statistical inference for multinomial processing tree models. In Jean-Paul Doignon and Jean-Claude Falmagne (Eds.). *Mathematical psychology: Current developments*. New York: Springer-Verlag, 313-335.
- Riefer, D.M., and Batchelder, W.H. (1995) A multinomial modeling analysis of the recognition-failure paradigm. *Memory & Cognition*, *23*, 611-630.
- Riefer, D.M., Hu, X., and Batchelder, W.H. (1994) Response strategies in source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *1994*, *20*, 680-693.
- Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D. and 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.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* *6*, 461-464.
- Schweickert, R. and Chen, S. (2008). Tree inference with factors 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.
- Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, *14*, 779-804.

- Wagenmakers, E. -J., Grünwald, P. and Waldorp, L. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50, 149-166.
- Wagenmakers, E. -J., Ratcliff, R., Gomez, P. and Iverson, G. J. (2006). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48, 28-50.
- Wagenmakers, E. -J. and Waldorp, L. (2006). Editors' introduction. *Journal of Mathematical Psychology*, 50, 99-100.

Appendix: Three Lemmas

Here we present three lemmas used in the proofs of prepositions. Lemma 1 gives an equality of combination numbers used in the proof of Proposition 6. Lemma 2 gives the form of Fisher information matrix for BMPT models in standard representation. Lemma 3 gives the form of Fisher information matrix of complex BMPT models in terms of its components by exploiting the recursive property of BMPT models.

**Lemma 1.** *For all non-negative integer  $N$ ,  $\{G^{(i)}\}_{i=1}^p$ ,  $\{G_j\}_{j=1}^q$  and  $\{g_j^{(i)}\}_{i=1, j=1}^{p, q}$  that satisfy the restriction  $\sum_i G^{(i)} = \sum_j G_j = N$ , the following equality holds:*

$$\sum_{\left\{g_j^{(i)} \mid \begin{matrix} \sum_i g_j^{(i)} = G_j \\ \sum_j g_j^{(i)} = G^{(i)} \end{matrix} \right\}} \frac{\prod_i \binom{G^{(i)}}{g_1^{(i)}, g_2^{(i)}, \dots, g_q^{(i)}}}{\binom{N}{G_1, G_2, \dots, G_q}} = 1 \tag{9}$$

*Proof.* Suppose  $N$  objects belongs to  $p$  groups, with  $G^{(i)}$  objects in the  $i$ th group. Now we would like to randomly re-group all the objects into  $q$  groups, with  $G_j$  objects in the  $j$ th group. Let  $g_j^{(i)}$  be the number of objects that are recruited from the  $i$ th original group to the  $j$ th new group. Then the fraction in this lemma gives the probability for this particular solution, which, after summed up over all possible solutions, yields 1.  $\square$

**Lemma 2.** *The Fisher information matrix of a BMPT model with the representation shown in Equations (3),(1) and (2) is given by*

$$I_{rs} = -E \frac{\partial^2 \ln L}{\partial \theta_s \partial \theta_r} = \sum_{j=1}^J \frac{1}{p_j} \left\{ \sum_{i=1}^{I_j} p_{ij} \left( \frac{a_{ijs}}{\theta_s} - \frac{b_{ijs}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i=1}^{I_j} p_{ij} \left( \frac{a_{ijr}}{\theta_r} - \frac{b_{ijr}}{1 - \theta_r} \right) \right\} \tag{10}$$

*Especially, if every category of the BMPT model includes only a single leaf, its Fisher information matrix is a diagonal matrix given by*

$$I_{ss} = \sum_{j=1}^J \sum_{i=1}^{I_j} p_{ij} \left( \frac{a_{ijs}}{\theta_s^2} + \frac{b_{ijs}}{(1 - \theta_s)^2} \right) \tag{11}$$

*Proof.* Equation (10) follows from Grünwald, Myung & Pitt (2005, equation 16.4 on p.420) with the matrix  $P_{js}$  in the equation given by Hu & Batchelder (1994, equation 36 on p. 40). When the MPT model is a simple one, its Hessian matrix is given by (Hu & Batchelder, 1994, equation 37 on p. 40). Taking expectation of both sides completes the proof of (11).  $\square$

**Lemma 3.** *Let  $\mathcal{A}$  and  $\mathcal{B}$  be BMPT models with disjoint sets of categories  $\mathbf{C}_\mathcal{A}$  and  $\mathbf{C}_\mathcal{B}$  and parameter sets  $\Theta_\mathcal{A}$  and  $\Theta_\mathcal{B}$ , which need not be disjoint. Consider an MPT model  $\mathcal{C}$  constructed by replacing category  $C_0$  of  $\mathcal{A}$  with model  $\mathcal{B}$ . The Fisher information matrix of  $\mathcal{C}$  is given by  $\mathbf{I}^\mathcal{C} = \tilde{\mathbf{I}}^\mathcal{A} + \tilde{\mathbf{I}}^\mathcal{B} p_0^\mathcal{A}$ , where  $p_0^\mathcal{A}$  is the implied category probability of  $C_0$  in  $\mathcal{A}$ , and  $\tilde{\mathbf{I}}^\mathcal{A}$  is the extended Fisher information matrix in which all parameters in  $\mathcal{C}$  are included but the entries corresponding to parameters not in model  $\mathcal{A}$  are set to 0.*

*Remarks:*

1. This proposition can be easily extended to the case in which  $k > 1$  categories in  $\mathcal{A}$  are replaced by trees  $\mathcal{B}_k$  of disjoint sets of categories.
2. When the two parameter sets are disjoint,  $\mathbf{I}^\mathcal{C}$  is block diagonal and its determinant is given by  $|\mathbf{I}^\mathcal{C}| = |\mathbf{I}^\mathcal{A}||\mathbf{I}^\mathcal{B}|(p_0^\mathcal{A})^{S_\mathcal{B}}$ , where  $S_\mathcal{B}$  is the number of parameters in model  $\mathcal{B}$ .

*Proof.* We first prove a version of the proposition with *disjoint* parameter sets  $\Theta_\mathcal{A}$  and  $\Theta_\mathcal{B}$ .

Before diving into algebra, we first give some links between model  $\mathcal{C}$  and models  $\mathcal{A}$  and  $\mathcal{B}$ . For category and leaf probabilities: if  $C_j \in \mathbf{C}_\mathcal{A}$  with  $j \neq 0$ , then we have  $p_{ij}^\mathcal{C} = p_{ij}^\mathcal{A}$  and  $p_j^\mathcal{C} = p_j^\mathcal{A}$ ; if  $C_j \in \mathbf{C}_\mathcal{B}$ , we have  $p_{ij}^\mathcal{C} = p_{i_\mathcal{A}0}^\mathcal{A} p_{i_\mathcal{B}j}^\mathcal{B}$  and  $p_j^\mathcal{C} = p_0^\mathcal{A} p_j^\mathcal{B}$ , where  $i_\mathcal{A}0$  and  $i_\mathcal{B}j$  are the leaves related to leaf  $ij$  in  $C_j$ . For the  $a_{ijs}$  and  $b_{ijs}$ : if  $\theta_s \in \Theta_\mathcal{A}$ , we have  $a_{ijs}^\mathcal{C} = a_{i_\mathcal{A}0s}^\mathcal{A}$  and  $b_{ijs}^\mathcal{C} = b_{i_\mathcal{A}0s}^\mathcal{A}$  for  $C_j \in \mathbf{C}_\mathcal{B}$  and  $a_{ijs}^\mathcal{C} = a_{i_\mathcal{A}js}^\mathcal{A}$  and  $b_{ijs}^\mathcal{C} = b_{i_\mathcal{A}js}^\mathcal{A}$  for  $C_j \in \mathbf{C}_\mathcal{A}$ ; if  $\theta_s \in \Theta_\mathcal{B}$ , we have  $a_{ijs}^\mathcal{C} = b_{ijs}^\mathcal{C} = 0$  for  $C_j \in \mathbf{C}_\mathcal{A}$  and  $a_{ijs}^\mathcal{C} = a_{i_\mathcal{B}js}^\mathcal{B}$  and  $b_{ijs}^\mathcal{C} = b_{i_\mathcal{B}js}^\mathcal{B}$  for  $C_j \in \mathbf{C}_\mathcal{B}$ .

Consider element  $\mathbf{I}_{rs}^\mathcal{C}$  with  $\theta_r, \theta_s \in \Theta_\mathcal{A}$ . Remember that the expression of  $\mathbf{I}_{rs}^\mathcal{C}$  in equation (10) is a sum over the set  $\{j|C_j \in \mathbf{C}_\mathcal{C}\}$  (a short hand notation  $j \in \mathcal{C}$  will be used). We now consider two parts of the sum separately:

$$\begin{aligned}
 & \sum_{j \in \mathcal{B}} \frac{1}{p_j^\mathcal{C}} \left\{ \sum_{i \in j} p_{ij}^\mathcal{C} \left( \frac{a_{ijs}^\mathcal{C}}{\theta_s} - \frac{b_{ijs}^\mathcal{C}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i \in j} p_{ij}^\mathcal{C} \left( \frac{a_{ijr}^\mathcal{C}}{\theta_r} - \frac{b_{ijr}^\mathcal{C}}{1 - \theta_r} \right) \right\} \\
 = & \sum_{j \in \mathcal{B}} \frac{1}{p_0^\mathcal{A} p_j^\mathcal{B}} \left\{ \sum_{i_\mathcal{A} \in 0} \sum_{i_\mathcal{B} \in j} (p_{i_\mathcal{A}0}^\mathcal{A} p_{i_\mathcal{B}j}^\mathcal{B}) \left( \frac{a_{i_\mathcal{A}0s}^\mathcal{A}}{\theta_s} - \frac{b_{i_\mathcal{A}0s}^\mathcal{A}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i_\mathcal{A} \in 0} \sum_{i_\mathcal{B} \in j} (p_{i_\mathcal{A}0}^\mathcal{A} p_{i_\mathcal{B}j}^\mathcal{B}) \left( \frac{a_{i_\mathcal{A}0r}^\mathcal{A}}{\theta_r} - \frac{b_{i_\mathcal{A}0r}^\mathcal{A}}{1 - \theta_r} \right) \right\} \\
 = & \frac{1}{p_0^\mathcal{A}} \left\{ \sum_{i_\mathcal{A} \in 0} p_{i_\mathcal{A}0}^\mathcal{A} \left( \frac{a_{i_\mathcal{A}0s}^\mathcal{A}}{\theta_s} - \frac{b_{i_\mathcal{A}0s}^\mathcal{A}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i_\mathcal{A} \in 0} p_{i_\mathcal{A}0}^\mathcal{A} \left( \frac{a_{i_\mathcal{A}0r}^\mathcal{A}}{\theta_r} - \frac{b_{i_\mathcal{A}0r}^\mathcal{A}}{1 - \theta_r} \right) \right\}
 \end{aligned}$$

$$\begin{aligned}
& \sum_{j \in \mathcal{A}} \frac{1}{p_j^{\mathcal{C}}} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{C}} \left( \frac{a_{ijs}^{\mathcal{C}}}{\theta_s} - \frac{b_{ijs}^{\mathcal{C}}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{C}} \left( \frac{a_{ijr}^{\mathcal{C}}}{\theta_r} - \frac{b_{ijr}^{\mathcal{C}}}{1 - \theta_r} \right) \right\} \\
&= \frac{1}{p_j^{\mathcal{A}}} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{A}} \left( \frac{a_{ijs}^{\mathcal{A}}}{\theta_s} - \frac{b_{ijs}^{\mathcal{A}}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{A}} \left( \frac{a_{ijr}^{\mathcal{A}}}{\theta_r} - \frac{b_{ijr}^{\mathcal{A}}}{1 - \theta_r} \right) \right\}
\end{aligned}$$

Summing up the two parts we have  $I_{rs}^{\mathcal{C}} = I_{rs}^{\mathcal{A}}$ .

For  $\theta_s, \theta_r \in \Theta_{\mathcal{B}}$ , we note in equation (10) the summands with  $j \in \mathcal{A}$  does not contribute to the summation because the  $a$ 's and  $b$ 's are 0, so we only need to consider summation over  $j \in \mathcal{B}$ .

$$\begin{aligned}
I_{rs}^{\mathcal{C}} &= \sum_{j \in \mathcal{B}} \frac{1}{p_j^{\mathcal{C}}} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{C}} \left( \frac{a_{ijs}^{\mathcal{C}}}{\theta_s} - \frac{b_{ijs}^{\mathcal{C}}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{C}} \left( \frac{a_{ijr}^{\mathcal{C}}}{\theta_r} - \frac{b_{ijr}^{\mathcal{C}}}{1 - \theta_r} \right) \right\} \\
&= \sum_{j \in \mathcal{B}} \frac{1}{p_0^{\mathcal{A}} p_j^{\mathcal{B}}} \left\{ \sum_{i_{\mathcal{A}} \in 0} \sum_{i_{\mathcal{B}} \in j} (p_{i_{\mathcal{A}}0}^{\mathcal{A}} p_{i_{\mathcal{B}}j}^{\mathcal{B}}) \left( \frac{a_{i_{\mathcal{B}}js}^{\mathcal{B}}}{\theta_s} - \frac{b_{i_{\mathcal{B}}js}^{\mathcal{B}}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i_{\mathcal{A}} \in 0} \sum_{i_{\mathcal{B}} \in j} (p_{i_{\mathcal{A}}0}^{\mathcal{A}} p_{i_{\mathcal{B}}j}^{\mathcal{B}}) \left( \frac{a_{i_{\mathcal{B}}jr}^{\mathcal{B}}}{\theta_r} - \frac{b_{i_{\mathcal{B}}jr}^{\mathcal{B}}}{1 - \theta_r} \right) \right\} \\
&= p_0^{\mathcal{A}} I_{rs}^{\mathcal{B}}
\end{aligned}$$

For  $\theta_s \in \Theta_{\mathcal{A}}$  and  $\theta_r \in \Theta_{\mathcal{B}}$ , similar to the previous case, we only need to consider  $j \in \mathcal{B}$  for summation. Simple algebra gives

$$\begin{aligned}
I_{rs}^{\mathcal{C}} &= \sum_{j \in \mathcal{B}} \frac{1}{p_j^{\mathcal{C}}} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{C}} \left( \frac{a_{ijs}^{\mathcal{C}}}{\theta_s} - \frac{b_{ijs}^{\mathcal{C}}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i \in j} p_{ij}^{\mathcal{C}} \left( \frac{a_{ijr}^{\mathcal{C}}}{\theta_r} - \frac{b_{ijr}^{\mathcal{C}}}{1 - \theta_r} \right) \right\} \\
&= \sum_{j \in \mathcal{B}} \frac{1}{p_0^{\mathcal{A}} p_j^{\mathcal{B}}} \left\{ \sum_{i_{\mathcal{A}} \in 0} \sum_{i_{\mathcal{B}} \in j} (p_{i_{\mathcal{A}}0}^{\mathcal{A}} p_{i_{\mathcal{B}}j}^{\mathcal{B}}) \left( \frac{a_{i_{\mathcal{A}}0s}^{\mathcal{A}}}{\theta_s} - \frac{b_{i_{\mathcal{A}}0s}^{\mathcal{A}}}{1 - \theta_s} \right) \right\} \left\{ \sum_{i_{\mathcal{A}} \in 0} \sum_{i_{\mathcal{B}} \in j} (p_{i_{\mathcal{A}}0}^{\mathcal{A}} p_{i_{\mathcal{B}}j}^{\mathcal{B}}) \left( \frac{a_{i_{\mathcal{B}}jr}^{\mathcal{B}}}{\theta_r} - \frac{b_{i_{\mathcal{B}}jr}^{\mathcal{B}}}{1 - \theta_r} \right) \right\} \\
&= \left\{ \sum_{i_{\mathcal{A}} \in 0} p_{i_{\mathcal{A}}0}^{\mathcal{A}} \left( \frac{a_{i_{\mathcal{A}}0s}^{\mathcal{A}}}{\theta_s} - \frac{b_{i_{\mathcal{A}}0s}^{\mathcal{A}}}{1 - \theta_s} \right) \right\} \sum_{j \in \mathcal{B}} \left\{ \sum_{i_{\mathcal{B}} \in j} p_{i_{\mathcal{B}}j}^{\mathcal{B}} \left( \frac{a_{i_{\mathcal{B}}jr}^{\mathcal{B}}}{\theta_r} - \frac{b_{i_{\mathcal{B}}jr}^{\mathcal{B}}}{1 - \theta_r} \right) \right\} \\
&= \frac{\partial p_0^{\mathcal{A}}}{\partial \theta_s} \left( \sum_{j \in \mathcal{B}} \frac{\partial p_j^{\mathcal{B}}}{\partial \theta_r} \right) = 0
\end{aligned}$$

because the second term is the derivative of a constant, i.e.,  $\sum p_j^{\mathcal{B}} = 1$ . Summarizing the three parts of Fisher information matrix, we have  $\mathbf{I}^{\mathcal{C}} = \text{diag}\{\mathbf{I}^{\mathcal{A}}, p_0^{\mathcal{A}} \mathbf{I}^{\mathcal{B}}\} = \tilde{\mathbf{I}}^{\mathcal{A}} + \tilde{\mathbf{I}}^{\mathcal{B}} p_0^{\mathcal{A}}$ .

If  $\Theta_{\mathcal{A}}$  and  $\Theta_{\mathcal{B}}$  are not disjoint, we first consider the version of tree  $\mathcal{C}$  in which the two sets are disjoint with Fisher information matrix  $\bar{\mathbf{I}}^{\mathcal{C}}$  of size  $|\Theta_{\mathcal{A}}| + |\Theta_{\mathcal{B}}|$ , where  $|\cdot|$  denote the number of elements in a set. Then we equate the parameters in the two sets that corresponds to parameters shared by  $\mathcal{A}$  and

$\mathcal{B}$ . Suppose the Jacobian matrix of the transformation is  $\mathbf{J}$ , a  $|\Theta_C| \times (|\Theta_A| + |\Theta_B|)$  matrix. We have  $\mathbf{I}^C = \mathbf{J}\bar{\mathbf{I}}^C\mathbf{J}^T = \mathbf{J}(\tilde{\mathbf{I}}^A + \tilde{\mathbf{I}}^B p_0^A)\mathbf{J}^T = \tilde{\mathbf{I}}^A + \tilde{\mathbf{I}}^B p_0^A$ .  $\square$

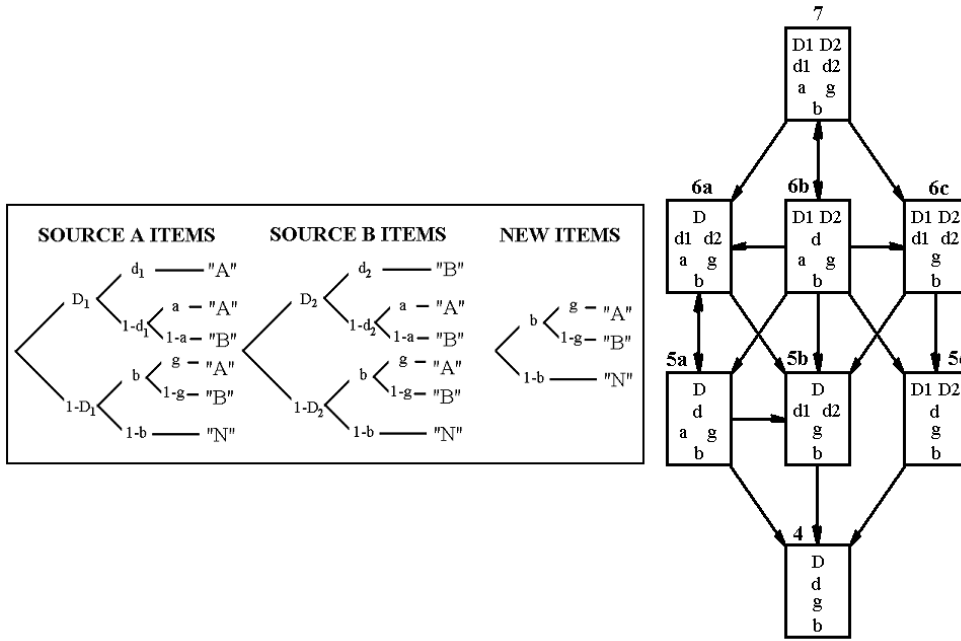
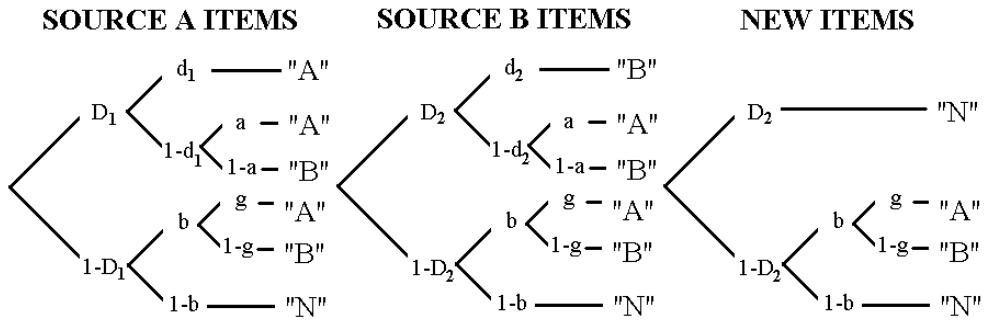


Figure 1. The one-high-threshold (1HT) multinomial processing tree model of source monitoring is shown on the left panel. The parameters are defined as follows:  $D_1$  (detectability of source A items);  $D_2$  (detectability of source B items);  $d_1$  (source discriminability of source A items);  $d_2$  (source discriminability of source B items);  $a$  (guessing that a detected but nondiscriminated item belongs to source A category);  $b$  (guessing "old" to a nondetected item);  $g$  (guessing that a nondetected item biased as old belongs to source A category). The right panel shows a nested hierarchy of eight versions of the model on the left, created by imposing successive constraints on the parameters. In the figure, the model parameters for each model are listed and a directed arrow from one model to another means that the second model is nested in the first.



*Figure 2.* The two-high-threshold (2HT) multinomial processing tree model of source monitoring. Adapted from Bayen, Murnane & Erdfelder (1996, Figure 3). The parameters are defined as follows:  $D_1$  (detectability of source A items);  $D_2$  (detectability of source B items);  $D_3$  (detectability of new 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 category);  $b$  (guessing old to a nondetected item);  $g$  (guessing that a nondetected item biased as old belongs to source A category).

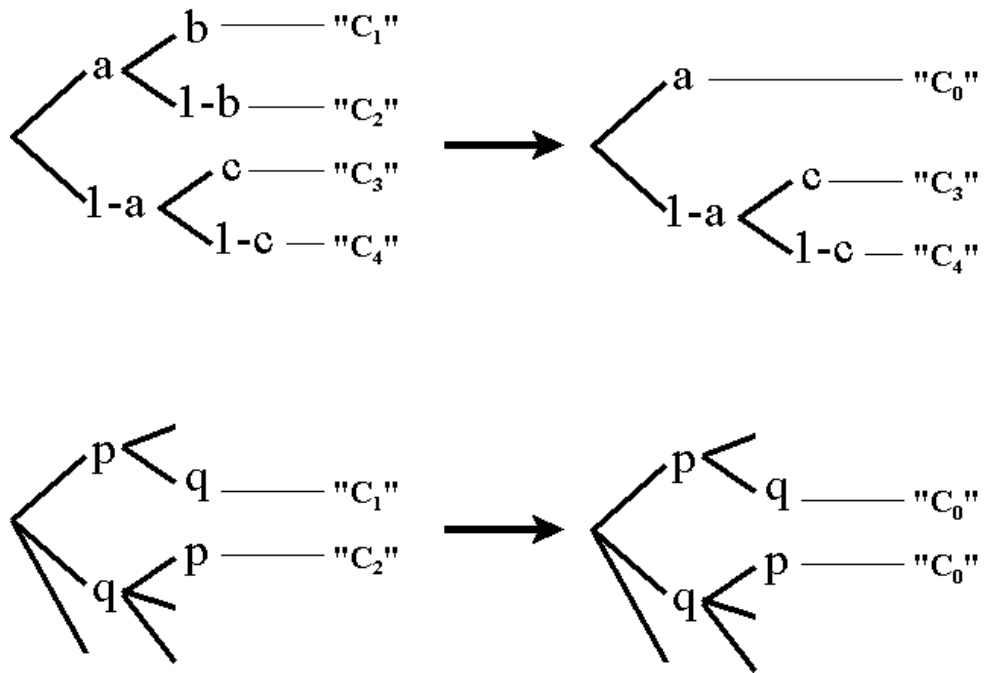


Figure 3. Examples of models created by combining leaves.

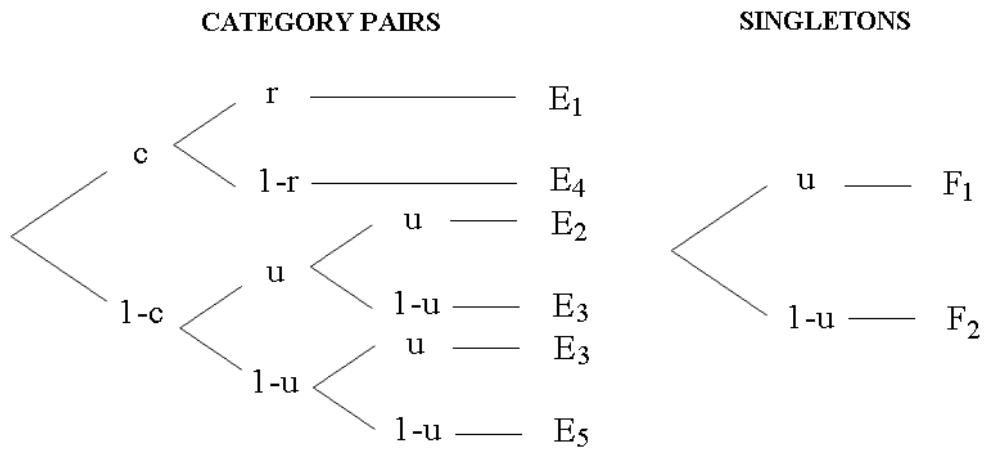


Figure 4. Batchelder and Riefer's (1999) multinomial processing tree model of pair-clustering.

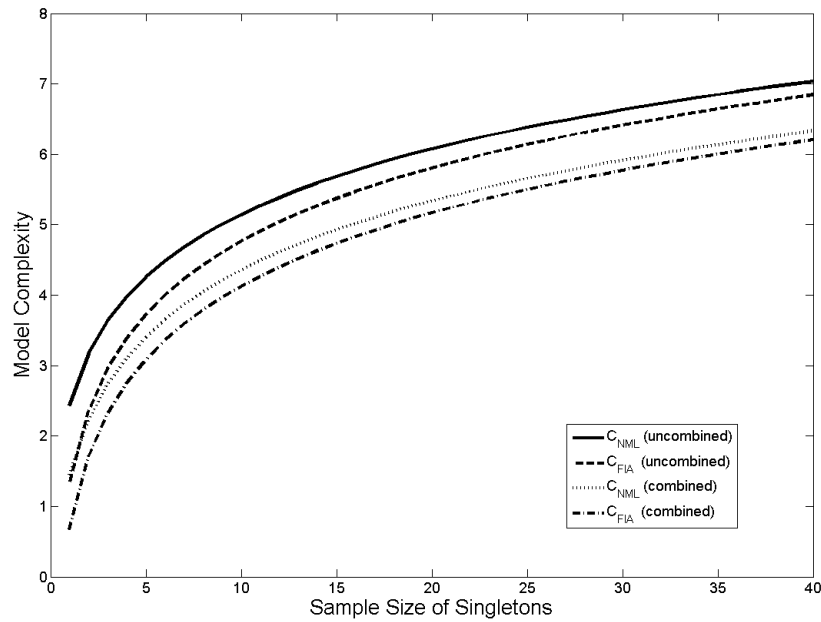


Figure 5.  $C_{NML}$  and  $C_{FIA}$  complexity curves for the model in Figure 4, plotted as a function of the sample size of singletons. The sample size of category pairs are set to be twice that of singletons. The uncombined version of the model assumes five distinct response categories  $E_1 - E_5$  for paired items. In the combined version of the model, the two response categories  $E_4$  and  $E_5$  are combined into one category. Note that all four models have the same number of parameters (3).

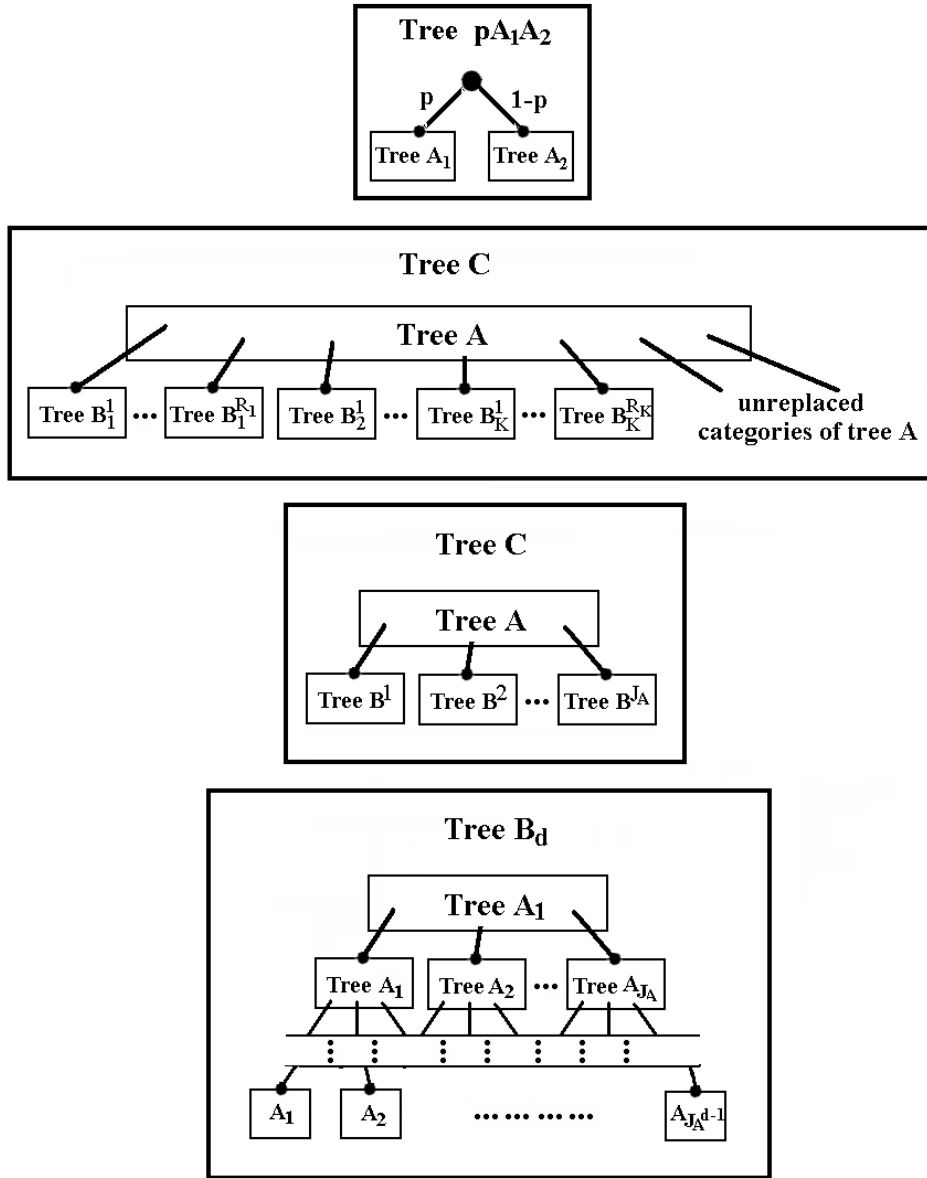


Figure 6. Graphical illustrations of model constructions. The top panel shows the new model constructed in Axiom 2 and Proposition 4 by joining two trees  $\mathcal{A}_1$  and  $\mathcal{A}_2$ . The second panel portrays the situation described in Lemma 3 and Proposition 5 and 6, in which some of the categories of  $\mathcal{A}$  are replaced by  $\mathcal{B}_k^r$ . In the third panel, a new tree  $\mathcal{C}$  is created by replacing every category of tree  $\mathcal{A}$  by  $\mathcal{B}^j$ ,  $j = 1, 2, \dots, J_{\mathcal{A}}$ , as in Proposition 7. The bottom panel demonstrates how tree  $\mathcal{B}_{d-1}$  in Proposition 8 is created by adding the same tree structure to every category of itself recursively.

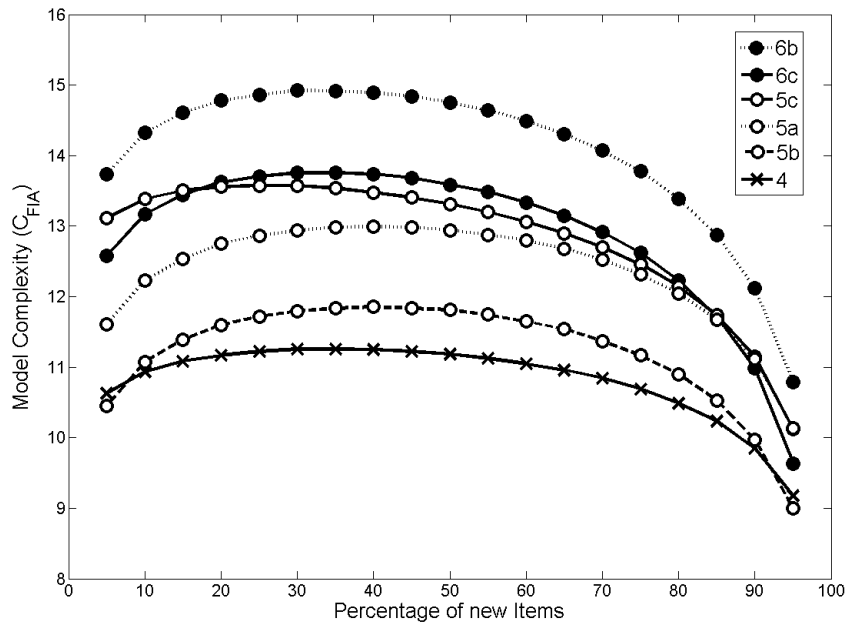


Figure 7.  $C_{FIA}$  complexity curves for six source-monitoring models in Figure 1. The total sample size is  $N = 1000$ .