

1  
2  
3  
4  
5  
6  
7  
8  
9 Minimum Description Length Model Selection of Multinomial  
10 Processing Tree Models  
11  
12

13  
14  
15  
16 Hao Wu and Jay I. Myung  
17

18 The Ohio State University  
19

20  
21 William H. Batchelder  
22

23 University of California, Irvine  
24  
25  
26  
27  
28

29 Word count of text and appendices: 7700  
30

31 Running head: Multinomial Processing Tree Models  
32  
33

34 Corresponding author and address:  
35

36 Hao Wu  
37

38 Department of Psychology  
39

40 Ohio State University  
41

42 1835 Neil Avenue  
43  
44

45 Columbus, Ohio 43210-1351  
46  
47

48 E-mail: wu.498@osu.edu  
49

50 Tel: 614-292-5510  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## MULTINOMIAL PROCESSING TREE MODELS

2

## Abstract

Multinomial processing tree (MPT) modeling has been widely and successfully applied as a statistical methodology for measuring hypothesized latent cognitive processes in selected experimental paradigms. This paper concerns the problem of selecting the “best” MPT model from a set of scientifically plausible MPT models given observed data. The likelihood ratio test is often employed in model selection for MPT models, but it is a null hypothesis significance test that assesses the descriptive adequacy of a given null model, and as such, does not necessarily help identify the best approximating model to the truth, which is the hallmark of model selection. Model selection methods such as the Akaike Information Criterion and the Bayesian Information Criterion do not fully take into account all relevant dimensions of model complexity, such as the number of parameters, model structure, and parametric inequality constraints, the latter two of which are of particular importance for MPT models. In this paper, we introduce a minimum description length (MDL) based model selection approach that overcomes the limitations of the aforementioned methods and therefore is well suited for model selection of MPT models. To help ease the computational burden of implementing MDL, we provide a computer program in *MatLab* that performs MDL-based model selection for any MPT model, with or without inequality constraints. Finally, we discuss applications of the the MDL approach to well-studied MPT models with real data sets collected in two different experimental paradigms: source monitoring and pair-clustering. The aforementioned *MatLab* program may be downloaded from [www.psychonomic.org/archive](http://www.psychonomic.org/archive).

## Introduction

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

Since its introduction in the 1980s, MPT models have been successfully applied to modeling performance in a wide range of cognitive tasks including associative recall, source monitoring, eyewitness memory, hindsight bias, object perception, speech perception, propositional reasoning, social networks, and cultural consensus. Batchelder and Riefer (1999) lists over 80 applications of MPT models in various areas of cognitive and social psychology. MPT models have also been applied to estimate cognitive deficits in special populations (see Batchelder & Riefer, 2007; Chechile, 2007, for a review of such applications). The use of MPT models to assess special populations is often referred to as *cognitive psychometrics* representing the fact that theoretically motivated models are employed as measurement tools of cognitive functioning (Batchelder, in press; Batchelder & Riefer, 2007; Riefer et al., 2002). In all these applications, MPT models intended to offer researchers more instructive and informative interpretations of data than those based on the traditional data analytic approaches such as the analysis of variance (ANOVA).

## MULTINOMIAL PROCESSING TREE MODELS

4

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

In the present study, we are concerned with the logic of selecting the “best” MPT model from a set of scientifically plausible MPT models that are available to account for a given data set. A researcher may entertain multiple scientific hypotheses about the underlying processes, each formulated as a distinct MPT model<sup>1</sup>, and may wish to determine which one of these models “best” describes the observed data in some defined sense; this is the problem of model selection (Myung & Pitt, 1997). By selecting among theoretically motivated models, the researcher is able to identify from alternative theories the one best supported by empirical observations. To illustrate, consider the question of how different languages of bilingual people are cognitively represented. Several theories addressing this issue differ as to whether or not information presented in a particular language retains a language specific tag. Source monitoring experiments were conducted to differentiate these theories (e.g. Saegert, Hamayan & Ahmar, 1975; Rose et al., 1975), and these theories, as represented by their corresponding source monitoring MPT models (to be elaborated in the next section) that assume different treatment effects on their parameters, can be compared by model selection (Batchelder & Riefer, 1990). Similarly, other theoretical issues in cognitive psychology such as sequential vs. non-sequential processes and automatic vs. control processes can also be addressed by comparing MPT models with different tree structures (e.g., Schweickert, 1993; Bishara & Payne, 2008).

In addition to evaluating multiple scientific theories behind different MPT models, model selection can also be employed as a tool for examining the validity of an MPT model.

---

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

1  
2  
3  
4  
5 1 The validity of an MPT model concerns whether or not it is warranted to interpret a  
6  
7 2 parameter in the model as representing the underlying cognitive process that it is explicitly  
8  
9  
10 3 postulated to represent (see, e.g., Batchelder & Riefer, 1999; Riefer et al., 2002; Schweickert  
11  
12 4 & Chen, 2008). To establish validity, it is necessary to apply experimental treatments  
13  
14 5 that have predictable selective influence on the parameters. For example, if a model has  
15  
16 6 a parameter  $\theta$  that is postulated to measure the ability to retrieve items from memory,  
17  
18 7 then experimental manipulations that should affect levels of retrievability should result in  
19  
20 8 predictable changes in  $\theta$  but no change in parameters postulated to measure other things.  
21  
22 9 To determine whether the desired selected influence is present for a particular MPT model,  
23  
24 10 it is necessary to select among different versions of the model assuming different patterns  
25  
26 11 of treatment effects.

27  
28  
29  
30  
31 12 Because of its importance in evaluating scientific theories and establishing validity of  
32  
33 13 MPT models, model selection is of particular interest in MPT modeling. To perform model  
34  
35 14 selection, one must account for the effect of model complexity. This is because model  
36  
37 15 complexity can affect the predictive capacity or accuracy of a model, which is the hallmark  
38  
39 16 of model selection (Myung, 2000; Myung & Pitt, 1997). In the case of MPT models, it  
40  
41 17 has been shown that they can vary greatly in complexity due to not only the number of  
42  
43 18 parameters but also importantly, functional form of the models such as tree structure and  
44  
45 19 parameter constraints (Wu, Myung & Batchelder, submitted). However, as will be discussed  
46  
47 20 later in this paper, the likelihood ratio test (LRT: Read & Cressie, 1988), currently in wide  
48  
49 21 use for MPT modeling, does not select models based on their predictive accuracy. Other  
50  
51 22 popular selection methods such as Akaike information criterion (AIC: Akaike, 1973) and  
52  
53 23 Bayesian information criterion (BIC: Schwartz, 1978) do not fully account for all dimensions  
54  
55  
56  
57  
58  
59  
60

1 of model complexity. Given these limitations, a model selection method that fully accounts  
2 for model complexity is called for.

3 In the present paper we introduce such a method for MPT modeling. This is minimum  
4 description length (MDL) model selection. MDL has been successfully applied to addressing  
5 various model selection problems in cognitive modeling (e.g., Lee, 2001; Pitt, Myung &  
6 Zhang, 2002; Navarro & Lee, 2004; Lee & Pope, 2001; Myung, Pitt & Navarro, 2007) but  
7 is entirely absent in MPT modeling, with the exception of our own work (Wu, Myung &  
8 Batchelder, submitted). To help researchers not familiar with numerical computing, in this  
9 paper we make available a general purpose computer program that implements MDL based  
10 model selection for virtually all types of MPT models.

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

## 17 Multinomial Processing Tree Models

18 Multinomial processing tree (MPT) models assume that the observed categorical re-  
19 sponses in an experiment follow from a series of latent cognitive events. These events  
20 are represented by a tree structure, in which non-terminating nodes represent the events,  
21 branches that follow from a node represent all possible outcomes of the event, with the  
22 probabilities of these outcomes being either parameters in the model or known constants,

1 and leaves (terminating nodes) of the tree structure represent the observed responses from  
2 subjects. Because different sequences of events may lead to the same response, a response  
3 category may include more than one leaf in the tree.

4 To illustrate how an MPT model works, consider the one-high-threshold model  
5 (1HTM) for source monitoring experiments as depicted in Figure 1. In a source moni-  
6 toring experiment, participants first study a list of items from two sources, A and B, and  
7 then are asked to judge the source of test items as either from A, from B, or new (N; i.e.  
8 a new item from neither source). The 1HTM for such experiments consists of three dis-  
9 tinct trees (Batchelder & Riefer, 1990), each modeling hypothetical processes assumed to  
10 be involved in responding to a given type of items. A distinguishing feature of this model  
11 is that it assumes that old items can be correctly detected with probabilities  $D_1$  and  $D_2$   
12 for items from sources A and B, respectively. If an old item is correctly detected as old, a  
13 discriminating decision on its source is made, with success probabilities  $d_1$  and  $d_2$  for the  
14 two sources, respectively. If any of the two processes fails, guessing processes follow. For  
15 new items, however, the model assumes no detection process and instead response selection  
16 is governed by guessing processes only. The model postulates three types of guessing pro-  
17 cesses represented by parameters  $b$ ,  $g$  and  $a$  (see Figure 1 for details). By putting various  
18 constraints on the model parameters, a hierarchy of sub-models can be derived from the  
19 model, which is shown in Figure 2. For instance, the equality constraints of  $D_1 = D_2$   
20 and  $d_1 = d_2$ , which amount to saying that the detection and discrimination probabilities  
21 both stay the same across items from different sources, results in 1HTM-5a. On the other  
22 hand, if we assume that only the source discrimination probabilities are the same for both  
23 sources ( $d_1 = d_2$ ) but not the detection probabilities ( $D_1 \neq D_2$ ), then 1HTM-6b, which

## MULTINOMIAL PROCESSING TREE MODELS

8

1 nests 1HTM-5a, is obtained instead.

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

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

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

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

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

18 Now let us assume that several participants make categorical responses to the same  
 19 set of items and that their responses are independently and identically distributed into the  
 20  $J$  categories of a model. Let  $n_j$  be the number of these responses that fall into category  $C_j$ ,

1  
2  
3  
4  
5  $\mathbf{n} = (n_1, n_2, \dots, n_J)'$  and  $N = \sum_j n_j$ . Then  $\mathbf{n}$  is distributed as a multinomial probability  
6  
7 distribution given by

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

8  
9  
10  
11  
12 where the multinomial probabilities  $p_j$  follows the computational rules in equations (2) and  
13  
14  
15 (1).

16  
17 The above mathematical description of BMPT models with constants  $a_{ijs}$ ,  $b_{ijs}$  and  
18  
19  $c_{ij}$ , though uniquely and sufficiently specifying the distribution of the data, can be cum-  
20  
21 bersome as an input to computer programs. For this purpose, Purdy and Batchelder (in  
22  
23 press) has devised a much more concise and elegant representation of BMPT models. Their  
24  
25 string representation scheme exploits the recursive properties of the tree structure and in-  
26  
27 cludes only branching probabilities and categories in the model. To illustrate, the string  
28  
29 representation of a coin flipping Bernouli model is given by  $pHT$ , where H and T are out-  
30  
31 comes of the process and  $p$  is the probability of obtaining the outcome of H. To obtain  
32  
33 the string representation for a more complex BMPT model, one begins with representa-  
34  
35 tion of the decision process at the root node, and then replaces the two outcomes with the  
36  
37 representations of the decision processes that follow those outcomes. To illustrate, take  
38  
39 the tree of source A items in the 1HTM in Figure 1. We first represent the item detec-  
40  
41 tion process with  $D_1$ (“detected”)(“undetected”). We then replace the outcome “detected”  
42  
43 with the representation of the discrimination process  $d_1A$ (“source unidentified”) and the  
44  
45 outcome “undetected” with that of the guessing process  $b$ (“guess as old”)N. Now we get  
46  
47  $D_1d_1A$ (“source unidentified”)N. We continue the replacement until the  
48  
49 string contains only branching probabilities and response categories. The string represen-  
50  
51 tation of the tree is  $D_1d_1AaABbgABN$ . This representation makes the input to computer  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5 programs much easier and will be exploited in our *MatLab* program described later in the  
6  
7 present article.  
8  
9

### 10 Methods of Model Selection

11  
12 As mentioned in the Introduction, model selection is a necessary and crucial step in  
13  
14 the application of MPT models. Various model selection methods have been proposed in  
15  
16 the past for this purpose. In the following we review some of these methods including the  
17  
18 minimum description length method.  
19  
20  
21

#### 22 *Likelihood Ratio Test*

23  
24  
25  
26 The  $G^2$ -based likelihood ratio test (LRT) is the most commonly used method of  
27  
28 inference in MPT modeling (e.g., Riefer & Batchelder, 1988; Hu & Batchelder, 1994; Hu  
29  
30 & Phillips, 1999). At the center of this approach is the likelihood ratio based test statistic  
31  
32 by which the adequacy of a model is evaluated in the null hypothesis significance testing  
33  
34 framework. Specifically, LRT requires the setting of two models, full and reduced, such that  
35  
36 the reduced model is nested in the full model with a reduction in the number of parameters.  
37  
38 The  $G^2$  test statistic is then defined as  $G^2 = -2 \ln LR$ , where  $LR$  is the ratio of the maximum  
39  
40 likelihood of the reduced model to that of the full model and  $\ln$  is the natural logarithm.  
41  
42  
43 Under the null hypothesis that the reduced model is correct, when the sample size  $N$  is  
44  
45 large enough, the sampling distribution of  $G^2$  is shown to follow a  $\chi^2$ -distribution with the  
46  
47 degrees of freedom equal to the difference in numbers of parameters between the models,  
48  
49 provided that certain regularity conditions are satisfied (e.g. Read & Cressie, 1988). If the  
50  
51 value of  $G^2$  is large enough to fall in the rejection region of the sampling distribution, then  
52  
53  
54  
55  
56  
57  
58  
59  
60 the null hypothesis is rejected and the full model is chosen. Otherwise, the reduced model

1  
2  
3  
4  
5 is chosen over the full model.  
6

7  
8       The  $G^2$ -based LRT is generally a useful method of model evaluation, but has several  
9  
10 limitations in its use as a model selection method for MPT models. First, the method can  
11  
12 only be used for comparing *pairs* of *nested* models, one pair at a time. This effectively  
13  
14 excludes its application to the situation in which multiple models with or without nesting  
15  
16 relationships are being compared. Second, the regularity conditions of the test require that  
17  
18 the maximum likelihood estimate (MLE) under either model should not be on the boundary  
19  
20 of the parameter space (see Shapiro, 1988, for an alternative procedure). This implies that  
21  
22 LRT is not able to take into account inequality constraints in the models. To see this,  
23  
24 because the inclusion of inequality constraints does not change the degrees of freedom of the  
25  
26 LRT, it changes the result of the test only when parameters of either model are estimated on  
27  
28 the boundary defined by those constraints, but that would violate the regularity conditions  
29  
30 mentioned above and render the test invalid. For the same reason, LRT cannot be employed  
31  
32 to compare two nested models with the same number of parameters but different functional  
33  
34 forms, such as 1HTM-6a and 6b in Figure 2.  
35  
36  
37  
38

39  
40       Besides the above issues, it is important to note that the goal of LRT is to assess the  
41  
42 *descriptive adequacy* of a given null model in the null hypothesis significance test framework,  
43  
44 but not to choose among a set of candidate models the one that best captures the regularities  
45  
46 underlying the data (Myung & Pitt, 1997). As such, LRT does not necessarily help identify  
47  
48 the best approximating model to the truth, which is what model selection is about. This  
49  
50 latter criterion is known as *generalizability* in statistics (e.g., Myung, 2000; Myung & Pitt,  
51  
52 1997). In the rest of this section we discuss various model selection criteria proposed as  
53  
54 generalizability measures and the importance of model complexity in determining a model's  
55  
56  
57  
58  
59  
60

1 generalizability.

## 2 *Generalizability and Model Complexity*

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

9 Model complexity or flexibility has to do with a model's intrinsic capability to fit a  
10 wide range of data patterns. Generally speaking, a model with many parameters is more  
11 complex than a model with fewer parameters. Further, models with the same number  
12 of parameters but different equation forms can also differ in complexity. This is called  
13 the "functional form" dimension of model complexity (Myung & Pitt, 1997). To give an  
14 example, two psychophysics models,  $y = ax^b + \varepsilon$  and  $y = a \log(x + b) + \varepsilon$  with  $\varepsilon \sim N(0, \sigma^2)$ ,  
15 may differ in complexity, despite the fact that they both have two parameters. Because of  
16 its flexibility, complex models tend to overfit the current data, thereby generalizing poorly  
17 to future observations and should therefore be penalized in model selection.

18 In the case of MPT models, differences in complexity due to functional form can arise  
19 in a variety of ways: from different tree structures, different parametric constraints, and/or  
20 different category assignments to the leaves of a tree. For example, consider 1HTM-5a,  
21 5b and 5c shown in Figure 2, each of which imposes a distinct set of equality constraints  
22 on the parameters of the largest model 1HTM-7. Although all three models have five

1 parameters, their complexity may be quite different from one another. This is in fact what  
 2 Wu, Myung and Batchelder (submitted) found. Their results showed that the difference  
 3 in complexity between 1HTM-5a and 5b is larger than that between 1HTM-5b and 4. An  
 4 implication is that the complexity difference due to functional form of MPT models can be  
 5 even greater than that due to the number of parameters. The finding such as this points to  
 6 the importance of accounting for the functional form dimension of complexity, in particular,  
 7 in model selection with MPT models.

#### 8 *Minimum Description Length*

9 Various model selection methods that estimates a model's generalizability have been  
 10 proposed in statistics.<sup>2</sup> Among them, Akaike information criterion (AIC: Akaike, 1973) and  
 11 Bayesian information criterion (BIC: Schwartz, 1978) have been used in MPT modeling,  
 12 though minimum description length (MDL: Rissanen, 1996, 2001; Grünwald, 2007) has not,  
 13 to our knowledge. In what follows, we briefly review AIC and BIC before turning our  
 14 discussion to MDL.

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

$$AIC = -2 LML + 2 S \quad (4)$$

$$BIC = -2 LML + S \ln N \quad (5)$$

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

1 where  $LML(= \ln f(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})))$  denotes the natural logarithm of the maximized likelihood  
2 (ML),  $\mathbf{x}$  is the current data,  $S$  is the number of parameters and  $N$  is the sample size.  
3 For both model selection criteria, among a set of competing models, the model with the  
4 smallest criterion value is judged to best generalize and is thus preferred. We can see that  
5 in both equations the first term is related to the fit of the model, while the second term  
6 represents a complexity penalty, thereby formalizing the Occam's razor (Myung & Pitt,  
7 1997). However, both AIC and BIC penalize complex models only by their number of  
8 parameters, neglecting their functional form complexity. Consequently, both criteria are  
9 not appropriate for selecting among models with inequality constraints.

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

14 The principle of minimum description length (MDL) originates from algorithmic cod-  
15 ing theory in computer science. According to this principle, statistical modeling is viewed  
16 as data compression, and the best model is the one that compresses the data as tightly as  
17 possible. A model's ability to compress the data is measured by the shortest code length  
18 with which the data can be coded with the help of the model. The resulting code length is  
19 related to generalizability such that the shorter the code length, the better the model gener-  
20 alizes (Grünwald, 2007; Grünwald, Myung & Pitt, 2005; Grünwald, 2000; Myung, Navarro  
21 & Pitt, 2006).

22 The Fisher Information Approximation (FIA: Rissanen, 1996) represents a formal  
23 implementation of the MDL principle for model selection. It gives the shortest code length

1 with which a model can code the data.<sup>3</sup> This criterion is defined as

$$FIA = -LML + C_{FIA} \quad (6)$$

2 with

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

3 where  $\mathbf{I}(\boldsymbol{\theta})$  is the Fisher information matrix (e.g. Casela & Berger, 2001) of sample size one  
 4 with its elements given by  $I(\boldsymbol{\theta})_{ij} = -E \left[ \frac{\partial^2 \ln f(x_1|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right]$ .<sup>4</sup> A smaller criterion value indicates  
 5 better generalization, and thus, the model that minimizes the criterion should be chosen.

6 There are several observations one can make about FIA. First, in this selection crite-  
 7 rion, generalizability is measured as a trade-off between goodness of fit (LML) and complex-  
 8 ity ( $C_{FIA}$ ). Second, regarding the complexity measure of FIA,  $C_{FIA}$ , its first term captures  
 9 the effects of the number of parameters ( $S$ ) and its second term captures the functional form  
 10 effects through the Fisher information matrix ( $\mathbf{I}(\boldsymbol{\theta})$ ). Especially, note that the functional  
 11 form complexity in  $C_{FIA}$  is expressed as an integral over the parameter space. As such,  
 12 it would therefore be straightforward to represent inequality constraints on parameters in  
 13  $C_{FIA}$  as the constraints simply reduce the size of the parameter space. Third, FIA is re-  
 14 lated to BIC in that mathematically both criteria can be viewed as approximations to minus

<sup>3</sup>More precisely, it is a large sample approximation to normalized maximum likelihood (NML: Myung, Navarro & Pitt, 2006; Rissanen, 2001), which can be considered the shortest code length a model can achieve in coding the current data in the worst case of its true distribution. NML can be computationally intensive and will not be discussed in this paper.

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

1 twice the log marginal likelihood in Bayesian statistics when Jeffreys' prior  $\pi(\boldsymbol{\theta}) \propto \sqrt{|\mathbf{I}(\boldsymbol{\theta})|}$   
2 is assumed, but FIA gives a better approximation than BIC (Grünwald, 2007, Chapter 8).

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

#### 10 MDL Model Selection of MPT Models

##### 11 *A Computer Program for $C_{\text{FIA}}$ Computation*

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

18 The general purpose *MatLab* program for computing the quantity  $C_{\text{FIA}}$  for BMPT  
19 models will be available for download from the Psychonomic journal archive web site  
20 (<http://www.psychonomi.prg/archive>; see Archived Materials). The program evaluates the  
21 integral using a Monte Carlo algorithm. The technical details of this numerical integra-  
22 tion algorithm are described in Wu, Myung and Batchelder (submitted). The scope of the

1  
2  
3  
4  
5 1 program is general enough to compute the complexity of any BMPT model. Given that  
6  
7 2 every MPT model can be reparameterized into an equivalent BMPT model, the program is  
8  
9  
10 3 applicable to all MPT models. The program can also incorporate inequality constraints on  
11  
12 4 parameters in so far as the constrains are of the form  $\theta_1 < \theta_2 < \dots < \theta_k$  or its combinations.<sup>5</sup>  
13

14  
15 The program assumes that the BMPT model has a single tree structure. When  
16  
17 6  $K$  trees are present in one model, these trees should be combined into one single tree  
18  
19 7 with multinomial probabilities  $p_k = N_k/N$ , where  $N_k$  is the sample size for tree  $k$  in the  
20  
21 8 experimental design and  $N$  is the total sample size. To illustrate how this is done, consider  
22  
23  
24 9 model 1HTM shown in Figure 1 and suppose that the sample sizes for items from sources A,  
25  
26 10 B and N are 250, 250 and 500, respectively. The three trees should then be joined together  
27  
28  
29 11 to form a single tree with numerical probabilities  $p_A = p_B = 0.25$  and  $p_N = 0.5$ . Although  
30  
31 12 all three trees in 1HTM are BMPT models, the new tree we obtained by joining them is  
32  
33 13 not because there are three branches from the root node. It needs to be turned into a  
34  
35 14 BMPT model through reparameterization. To reparameterize, one first joins the two trees  
36  
37 15 for sources A and B to a single node with branching probabilities 0.5, and then joins the  
38  
39 16 resulting binary tree with the tree for new items with branching probabilities 0.5. This is  
40  
41 17 shown in Figure 3.  
42  
43  
44

45 18 The *MatLab* program involves a function `BMPTFIA` with six input argu-  
46  
47 19 ments and six output arguments: `function [CFIA,CI,lnInt,CI1,lnconst,CI2] =`  
48  
49 20 `BMPTFIA(s,parameters,ineq0,category,N,Sample)`. The input and output arguments  
50  
51

52  
53 <sup>5</sup>Knapp and Batchelder (2004) has shown that BMPT models with such inequality constraints can be  
54  
55 reparameterized into BMPT models without inequality constraints. Our program computes the integral of  
56  
57 the original model directly without invoking such reparameterization.  
58  
59  
60

## MULTINOMIAL PROCESSING TREE MODELS

18

1 are described below.

2       The first input argument **s** is related to the string representation of the BMPT model  
 3 as discussed earlier. It can be obtained by replacing all categories in the string by the capital  
 4 letter “C” and all branching probabilities, including parameters and fixed constants, by the  
 5 lower case letter “p”. For example, for model 1HTM shown in Figure 1, this argument  
 6 should be **s**=“ppppCpCCppCCCpCpCCppCCCpCCC”.

7       The second input argument **parameters** is a row vector that assigns parameters or  
 8 constants to the p’s in the string **s**. Its length should be the same as the number of p’s in **s**,  
 9 and its elements correspond to the p’s according to their order in **s**. Positive integer elements  
 10 in **parameters** assign parameters to the corresponding p’s, with the same integer denoting  
 11 the same parameter. Constants are assigned to the p’s using the *negation* of their values. For  
 12 model 5c with multinomial probabilities .25, .25 and .5 for sources A, B and N, respectively,  
 13 this input argument should be **parameters**= [-0.5, -0.5, 1, 3, 5, 4, 5, 2, 3, 5, 4, 5, 4, 5]. In  
 14 this vector, the five parameters ( $D_1$ ,  $D_2$ ,  $d$ ,  $b$ ,  $g$ ) are coded using integers 1 through 5,  
 15 respectively, and the first two elements of the vector (-0.5’s) are the probability constants  
 16 we used to join the three trees into a single tree.

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

1           `ineq0= []`.

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

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

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

15           We now provide an example of the application of the computer program. Consider  
16 again 1HTM-5c in Figure 2 with an inequality constraint of  $D_1 > D_2$  and sample sizes 250,  
17 250 and 500 for the three kinds of stimuli. In a Monte Carlo run with `Sample = 200,000`,  
18 we obtained `CFIA= 12.6182`, `CI= [12.6113, 12.6251]`.

19           The *MatLab* program described above gives only the complexity term,  $C_{FIA}$ . As  
20 shown in (6), to obtain the value of FIA for a given MPT model, one must also com-  
21 pute the goodness of fit term,  $-LML$ . This term can be obtained from a user-friendly  
22 program called *GPT.EXE* (Hu & Phillips, 1999) that is available for free download from  
23 <http://www.xiangenhu.info/>. This program performs maximum likelihood estimation and

1  
2  
3  
4  
5 1 outputs best-fit parameter values and the value of  $-LML$  for any MPT model with and  
6  
7 2 without inequality constraints.  
8  
9

10 3 In what follows, we demonstrate the use of the *MatLab* program for model selec-  
11  
12 4 tion of MPT models in two different experimental paradigms: source monitoring and pair  
13  
14 5 clustering.  
15  
16

### 17 18 6 *Modeling Source Monitoring Data* 19

20  
21 7 Our first example concerns the source monitoring experiment of Rose et al. (1975,  
22  
23 8 Experiment 1). The purpose of their study was to examine whether accurate source memory  
24  
25 9 of language could occur at the semantic level of language processing. In their experiment,  
26  
27 10 subjects studied a mixed list of English and Spanish sentences before being tested on recog-  
28  
29 11 nition and source memory performance. Contextual relationships between the sentences  
30  
31 12 were manipulated in the experiment such that in one condition, the sentences were se-  
32  
33 13 mantically related to a common topic, whereas in the other condition, all sentences were  
34  
35 14 semantically unrelated. Rose et al. (1975) reasoned that if language source information is  
36  
37 15 available at the semantic level of processing, because the contextual relationship among the  
38  
39 16 sentences makes them semantically less distinguishable, language source memory would be  
40  
41 17 poorer for related sentences than for unrelated sentences.  
42  
43  
44  
45

46  
47 18 Based on an analysis of variance of the data, Rose et al. (1975) concluded that there  
48  
49 19 was no significant difference between the two treatment conditions. Batchelder and Riefer  
50  
51 20 (1990, Tables 7 and 8) reanalyzed the same data with model 1HTM-4, as shown in Figure 2,  
52  
53 21 that can distinguish the effect of item detection from that of source discrimination. Their  
54  
55 22 LRT results suggested that recognition memory ( $D$ ) was significantly poorer for related than  
56  
57  
58  
59  
60

1  
2  
3  
4  
5 for unrelated sentences, but there was no significant difference in source monitoring ( $d$ ).  
6  
7 Based on this finding, Batchelder and Riefer (1990) concluded that contextual relationship  
8  
9  
10 is detrimental to item detection but not to source discrimination, indicating that language  
11  
12 source information is not available at the semantic level.  
13

14 We re-analyze the data as reproduced in Batchelder and Riefer (1990, Table 7) in a  
15  
16 model selection framework using MDL, as well as AIC and BIC. The results are shown in  
17  
18 Table 1. The first model  $M_0$  has 8 parameters, four ( $D, d, b, g$ ) for each treatment condition,  
19  
20 without any equality constraints. A total of 15 MPT models were further created by impos-  
21  
22 ing various equality constraints on the parameters. For the rest of the models, a subscript  
23  
24 notation is used to indicate how the parameters are constrained to be equal across the two  
25  
26 treatment conditions, related and unrelated. For example, in  $M_{Dd}$ , both  $D$  and  $d$  are set to  
27  
28 equal across the two conditions whereas  $b$  and  $g$  are allowed to vary across the conditions. In  
29  
30 addition to such equality constraints, inequality constraints are also considered for models  
31  
32 where either of the two parameters,  $D$  and  $d$ , is allowed to differ between the two conditions  
33  
34 because theories underlying those models predict a particular direction of difference: both  
35  
36 parameters are expected to be smaller for related sentences than for unrelated sentences.  
37  
38 No inequality constraints on the guessing parameters,  $b$  and  $g$ , are imposed even if they  
39  
40 are allowed to differ across the two conditions. Because parameter estimates do not vio-  
41  
42 late these inequality constraints in the data, incorporating inequality constraints does not  
43  
44 change the values of LML, AIC and BIC. In contrast, inequality constraints do change the  
45  
46 value of  $C_{FIA}$  and thus the value of FIA. In Table 1, the latter two values obtained under  
47  
48 inequality constraints are denoted by  $C'_{FIA}$  and  $FIA'$ .  
49  
50  
51  
52  
53  
54  
55

56 From the table we can observe that as the number of parameters decreases,  $-LML$   
57  
58  
59  
60

## MULTINOMIAL PROCESSING TREE MODELS

22

1 increases while  $C_{\text{FIA}}$  decreases, as expected, exhibiting a tradeoff between goodness of fit  
2 and complexity. Of particular interest is the observation that among models with the same  
3 number of parameters,  $C_{\text{FIA}}$  varies greatly, indicating the effects of functional form on  
4 complexity. The difference in complexity due to functional form between two models can  
5 sometimes be even greater than the difference in goodness of fit. The case in point is the  
6 comparison between  $M_d$  and  $M_g$ . In terms of  $-LML$ ,  $M_d$  fits better the data than  $M_g$   
7 (36.18 vs 36.61) but is more complex ( $C_{\text{FIA}} = 20.7 > C_{\text{FIA}} = 19.7$ ), thus yielding an overall  
8 larger FIA value than  $M_g$  (56.9 vs. 56.3). As a result,  $M_g$  is preferred to  $M_d$  under FIA,  
9 though the latter would be selected if we were to treat the two models equally complex  
10 as in AIC and BIC. On the same token, inequality constraints can have similar effects on  
11 model complexity. For example,  $M_{Ddbg}$  ( $FIA = 55.2$ ) is preferred to  $M_{bg}$  ( $FIA = 55.6$ ) if  
12 no inequality constraints on  $b$  and  $g$  in  $M_{bg}$  are considered, but the preference is reversed if  
13 the constraints are considered ( $FIA' = 54.2$  for  $M_{bg}$ ).

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

1  
2  
3  
4  
5 1 The model selection analysis discussed so far is conducted for models obtained by  
6  
7 2 considering all possible combinations of constraints, equality and inequality, on the four  
8  
9 3 parameters ( $D, b, g, b$ ). In addressing the theoretical issue raised in Rose et al. (1975) and  
10  
11 4 Batchelder and Riefer (1990), however, one only needs to consider the two parameters  $D$   
12  
13 5 and  $d$  of main interest. Under this more focused scope, there are four relevant models to  
14  
15 6 compare:  $M_0$ ,  $M_D$ ,  $M_d$  and  $M_{Dd}$ , along with their inequality constraints. Among these four,  
16  
17 7  $M_d$  is favored under all three criteria, AIC, BIC and FIA. According to this best generalizing  
18  
19 8 model, the two treatment conditions with related and unrelated sentences, differ in item  
20  
21 9 recognition ( $D$ ) but not in source monitoring ( $d$ ). An implication of this conclusion is that  
22  
23 10 semantic information is useful in the recognition task but does not include any information  
24  
25 11 about language source.

26  
27  
28  
29  
30  
31 12 Essentially the same conclusion as ours was reached by both Rose et al. (1975) and  
32  
33 13 Batchelder and Riefer (1990). Although this particular example may be somewhat dis-  
34  
35 14 appointing, one should note that generally speaking, FIA-based model selection analysis  
36  
37 15 allows one to entertain and evaluate all types of theoretical hypotheses of interest, and that  
38  
39 16 the effort is in turn likely to generate much richer and deeper insights into the underlying  
40  
41 17 cognitive processes than analyses based on traditional methods such as LRT and analysis  
42  
43 18 of variance.

#### 44 45 46 47 48 19 *Modeling Pair-clustering Data*

49  
50  
51  
52 20 Our second example demonstrating the application of FIA based model selection is  
53  
54 21 related to the pair clustering experiment of Batchelder and Riefer (1980, experiment 1a).  
55  
56 22 The purpose of this experiment was to examine the effects of within-category spacing on  
57  
58  
59  
60

## MULTINOMIAL PROCESSING TREE MODELS

24

1 recall performance. They hypothesized that a small lag between a pair of categorically  
2 related words facilitates the formation and storage of a pair-cluster whereas a large lag  
3 facilitates the retrieval process. In the experiment, participants first studied a word list  
4 that consisted of both paired words and singletons and were then tested in a free recall  
5 task. Paired words were two words that were categorically related. In the word list, each  
6 pair of words occupied positions that were separated by  $J = 0, 4, 12, 24$  words unrelated to  
7 the pair. Five trials of this study-recall cycle were repeated. The data set we reanalyze in  
8 this paper was from Batchelder and Riefer (1986, Table 1).

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

1 for the entire data set.

2 Results from separate LRTs for the five trials performed by Batchelder and Riefer  
 3 (1986) showed that parameter  $c$  was significantly different across lag conditions for data  
 4 from all five trials. However,  $r$  was significant only for trials 1 and 3, but not for the other  
 5 trials. Another LRT with data from all five trials combined revealed that  $c$  was significantly  
 6 different across lag conditions, but  $r$  was only marginally significant. These results, taken  
 7 together, supported the hypothesis that small lag between category pairs facilitates the  
 8 formation and storage of pair clusters, but offered only marginal support for the hypothesis  
 9 that long lag facilitates the retrieval of pairs. Further and importantly, given that boundary  
 10 maximum likelihood estimates have been obtained due to the inequality constraints when  
 11 fitting the model, the LRTs results would be uninterpretable.

12 Now we re-analyze the data by FIA based model selection. Table 2 summarizes the  
 13 results. There are eight models to be compared.  $M_0$  is the model with 65 parameters  
 14 described above.  $M_u$  with 45 parameters assumes a single  $u$  for each trial as in typical pair  
 15 clustering models. From this model, various equality constraints on  $c$  and  $r$  are applied to  
 16 form the rest of models. Models  $M_{ur}$  ( $M_{uc}$ ) assumes the same  $r$  ( $c$ ) across the four conditions.  
 17 In  $M_{ucr}$ , both  $c$  and  $r$  are assumed to be the same throughout the lag conditions. The three  
 18 “primed” models,  $M'_u$ ,  $M'_{ur}$  and  $M'_{uc}$ , differ from the un-primed ones in that in the former,  
 19 additional inequality constraints are imposed on the relevant parameters across the four  
 20 lag conditions, such as  $c_{J=0} > c_{J=4} > c_{J=12} > c_{J=24}$  or  $r_{J=0} < r_{J=4} < r_{J=12} < r_{J=24}$ , or  
 21 both, to embody the theoretical hypotheses concerning the order of those parameters. Such  
 22 constraints do not change the number of parameters in the model but they may lead to  
 23 a larger  $-LML$  value as the maximum likelihood is searched over the restricted and thus

1 smaller parameter space. This is indeed observed in Table 2 for all three pairs of models.  
 2 For the same reason, inequality constraints reduce model complexity.

3 From Table 2, one can observe the trade-off between goodness of fit and model com-  
 4 plexity; the smallest  $-LML$  value of 141.3 is achieved by the most complex model  $M_0$  with  
 5  $C_{FIA} = 137.0$ . At the other end of the complexity spectrum,  $M_{ucr}$  with the fewest number  
 6 of parameters (15) gives the largest  $-LML$  value (206.2) and the smallest  $C_{FIA}$  value (43.9).  
 7 We also note that models with the same number of parameters can differ greatly in their  
 8 complexity. For example, the four models,  $M_{uc}$ ,  $M'_{uc}$ ,  $M_{ur}$  and  $M'_{ur}$ , all have 30 param-  
 9 eters yet their  $C_{FIA}$  complexity varies from the lowest 59.9 to the largest 79.5, due to the  
 10 combination of tree structure and inequality constraints on parameters. Such complexity  
 11 difference between two models with the same number of parameters can be larger than the  
 12 difference in LML, thus affecting model selection results. Such a pattern of result is indeed  
 13 observed in the table. That is, among the same four models with 30 parameters, both AIC  
 14 and BIC select  $M_{ur}$  whereas FIA picks  $M'_{ur}$ . This is because although  $M_{ur}$  fits the data  
 15 better than  $M'_{ur}$  ( $-LML = 168.0$  vs  $169.5$ ), the model is more complex ( $C_{FIA} = 79.5$  vs  
 16  $63.6$ ) and thus yields a larger FIA than its counterpart ( $FIA = 247.5$  vs  $233.1$ ).

17 Turning the discussion to model selection, among the eight models compared, AIC  
 18 favors the 30-parameter  $M_{ur}$  and BIC favors the most restrictive model  $M_{ucr}$  whereas FIA  
 19 selects  $M'_u$ . Note that the model  $M'_u$  imposes inequality constraints on both  $c$  and  $r$  in the  
 20 directions consistent with the experimental hypotheses of Batchelder and Riefer (1980). In  
 21 other words, our FIA-based reanalysis of the data supports the hypotheses in their ordered  
 22 form. As discussed earlier, the LRT results by Batchelder and Riefer (1986) indicated  
 23 that the hypothesized within-pair spacing effect on parameter  $r$  was inconclusive while the

1  
2  
3  
4  
5 hypothesized effect on parameter  $c$  was supported. The specific order relationships among  
6  
7 the hypotheses were not, however, examined. Model  $M'_{ur}$  embodies this interpretation of  
8  
9 the data and interestingly, turns out to be the second best model after  $M'_u$ . Especially, if  
10  
11 the three models with inequality constraints were not among the competing models, FIA  
12  
13 would choose  $M_{ur}$ , leading to a different conclusion. This shows that the hypothesized  
14  
15 order relationship of parameters, which may restrict the parameter space and reduce the  
16  
17 complexity of the model, can lead to different model selection conclusions, and as such,  
18  
19 should not be neglected.  
20  
21  
22  
23

24 To summarize, we demonstrated the application of the FIA-based model selection  
25  
26 approach for selecting among MPT models of pair-clustering for the Batchelder and Riefer  
27  
28 (1980) data set. The flexibility of the approach allowed us to construct and test a variety of  
29  
30 MPT models including models with inequality constraints on parameters. We compared the  
31  
32 results from our analysis to those from the LRT-based analysis of the same data reported  
33  
34 in Batchelder and Riefer (1986). By and large, the same scientific conclusions were drawn  
35  
36 from either analysis, though our FIA-based analysis provides more definitive support for  
37  
38 the hypotheses in their ordered form originally formulated and tested in Batchelder and  
39  
40 Riefer (1980,1989).  
41  
42  
43  
44  
45

## 46 Conclusion

47  
48  
49 Multinomial processing tree modeling represents a theoretically motivated and sta-  
50  
51 tistical justified methodology for evaluating cognitive capacities for various experimental  
52  
53 paradigms. Selecting among different MPT models is especially important both in ad-  
54  
55 dressing theoretical issues and in validating an MPT model in a particular experimental  
56  
57  
58  
59  
60

1 paradigm. In this paper we have introduced the MDL based model selection method to the  
2 practitioners of MPT modeling. Especially, to facilitate the use of this new methodology, we  
3 provide a general purpose computer program in *MatLab* that can be exploited to compute  
4 FIA for any MPT model. Two example applications of MDL model selection with real data  
5 sets selected from different experimental paradigms are also discussed.

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

18 As it is usually the case with any new methodology, MDL as presented in this pa-  
19 per is not without shortcomings. For example, it does not address the issues of model  
20 misspecification and individual differences. To allow for model misspecification, exact  
21 equality constraints should be replaced by fuzzy equality constraints as done in an ele-  
22 gant sampling-based method known as Population-Parameter Mapping (PPM) proposed  
23 by Chechile (1998). Regarding individual differences, one way of incorporating this im-

1  
2  
3  
4  
5 1 portant factor is through hierarchical modeling in which parameter values corresponding  
6  
7 2 different individuals are assumed to be sampled from a common distribution (see, e.g.,  
8  
9  
10 3 Klauer, 2006; Smith & Batchelder, in press). Although it is possible in theory to address  
11  
12 4 these issues within the MDL framework, it is beyond the scope of the present work.  
13

14  
15 In conclusion, model selection lies at the core of the scientific inference process. Ac-  
16  
17 6 cordingly, a theoretically well-justified and widely applicable methodology can help advance  
18  
19 7 science. We believe that MDL represents such a methodology that provides versatile yet  
20  
21 8 powerful tools for assessing the validity of MPT models in a way that goes beyond the  
22  
23 9 shortcomings of the current methods such as LRT, AIC, and BIC.  
24

25  
26 One final note. It is important to note that statistical model selection techniques  
27  
28 11 alone, however sophisticated, are not a panacea for all inference problems. Other non-  
29  
30 12 statistical means of model evaluation such as plausibility, interpretability, and explanatory  
31  
32 13 adequacy are equally, if not more, important. Instead of automatic tools implemented in  
33  
34 14 softwares, statistical model selection methods can be most useful if it is combined with the  
35  
36 15 judicious use of sound subjective but scientific judgement.  
37  
38  
39  
40

#### 41 Archived Materials

42

43  
44  
45 17 The following materials [and links] associated with this article may be ac-  
46  
47 18 cessed through the Psychonomic Societys Norms, Stimuli, and Data archive,  
48  
49 19 [www.psychonomic.org/archive](http://www.psychonomic.org/archive).  
50

51  
52 20 To access these files [or links], search the archive for this article using the journal  
53  
54 21 name (Psychonomic Bulletin & Review), the first authors name (Wu), and the publication  
55  
56 22 year (to be filled in if accepted).  
57  
58  
59  
60

## MULTINOMIAL PROCESSING TREE MODELS

30

1  
2  
3  
4  
5 1 FILE: ????.zip  
6

7 2 DESCRIPTION: The compressed archive file contains one file, BMPTFIA.m, a *Mat-*  
8

9  
10 3 *Lab* m-file that computes the FIA complexity  $C_{\text{FIA}}$  for general BMPT models.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Review Only

## References

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
- 1 Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In  
2 Kotz and Johnson (1992): *Breakthroughs in Statistics*. NY: Springer Verlag.
- 3  
4 Batchelder, W.H. (in press). Cognitive psychometrics: Using multinomial processing tree models  
5 as measurement tools. In S. Embretson and J. Roberts (Eds.). *New directions in psychological*  
6 *measurement with model based approaches*. APA Books.
- 7 Batchelder, W. H. & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall  
8 of clusterable pairs. *Psychological Review*, *87*, 375-397
- 9 Batchelder, W. H. & Riefer, D. M. (1986). The statistical analysis of a model for storage and  
10 retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*,  
11 *39*, 129-149.
- 12 Batchelder, W. H. & Riefer, D. M. (1990). Multinomial processing models of source monitoring.  
13 *Psychological Review*, *97*, 548-564.
- 14 Batchelder, W. H. & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process  
15 tree modeling. *Psychonomic Bulletin and Review*, *6*, 57-86.
- 16 Batchelder, W.H., & Riefer, D.M. (2007). Using multinomial processing tree models to measure  
17 cognitive deficits in clinical populations. In R.W. J. Neufeld (Ed.). *Advances in Clinical cognitive*  
18 *science: Formal modeling of processes and symptoms*. (pp. 19-50). Washington, D.C.: American  
19 Psychological Association Books.
- 20 Bishara, A. J. & Payne, B. K. (2008) Multinomial process tree models of control and automaticity  
21 in weapon misidentification. *Journal of Experimental Social Psychology*, *45*, 3, 524-534
- 22 Casela, G. & Berger, R. L. (2001) *Statistical Inference*. 2nd edition Duxbury Press;

## MULTINOMIAL PROCESSING TREE MODELS

32

- 1  
2  
3  
4  
5 1 Chechile, R.A. (1998). A new method for estimating model parameters for multinomial data. *Journal*  
6  
7 *of Mathematical Psychology, 42*, 432-471.  
8  
9  
10 3 Chechile, R.A. (2004). New models for the Chechile-Meyer task. *Journal of Mathematical Psychol-*  
11  
12 *ogy, 48*, 364-384.  
13  
14 5 Chechile, R.A. (2007). A model-based storage-retrieval analysis of developmental dyslexia. In R.W.J.  
15  
16 Neufeld (Ed.), *Advances in clinical cognitive sciences: Formal modeling of processes and symp-*  
17  
18 *toms.* (pp. 51-79), Washington, D.C.: American Psychological Association Books.  
19  
20  
21 8 Grünwald, P. (2000). Model selection based on Minimum Description Length. *Journal of Mathemat-*  
22  
23 *ical Psychology, 44*, 133-152.  
24  
25  
26 10 Grünwald, P. (2007). *The Minimum Description Length Principle.* MIT Press.  
27  
28  
29 11 Grünwald, P., Myung, I. J. & Pitt, M. A.(2005). *Advances in Minimum Description Length: Theory*  
30  
31 *and Applications.* MIT Press.  
32  
33  
34 13 Hu, X. & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with  
35  
36 the EM algorithm. *Psychometrika, 59*, 21-47.  
37  
38  
39 15 Hu, X. & Phillips, G.A. (1999). GPT.EXE: A powerful tool for visualizing and analysis of general  
40  
41 processing tree models. *Behavior Research Methods, Instruments, & Computers, 31*, 220-234.  
42  
43  
44 17 Klauer, K. C. (in press). Hierarchical multinomial processing tree models: a latent trait approach.  
45  
46 *Psychometrika*  
47  
48 19 Knapp, B., & Batchelder, W.H. (2004). Representing parametric order constraints in multi-trial  
49  
50 applications of multinomial processing tree models. *Journal of Mathematical Psychology, 2004*,  
51  
52 215-229.  
53  
54  
55 22 Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psy-*  
56  
57 *chology, 45*, 131-148.  
58  
59  
60

- 1  
2  
3  
4  
5 1 Lee, M. D. & Pope, K. J. (2006). Model selection for the rate problem: A comparison of significance  
6 testing, Bayesian, and minimum description length statistical inference. *Journal of Mathematical*  
7 *Psychology, 50*, 193-202.  
8  
9  
10  
11  
12 4 Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical*  
13 *Psychology, 44*, 190-204.  
14  
15  
16  
17 6 Myung, I. J., Forster, M. R. & Browne, M. W. (2000). Guest editors' introduction, special issue on  
18  
19 7 model selection. *Journal of Mathematical Psychology, 44*, 1-2.  
20  
21  
22 8 Myung, J. I., Navarro, D. J. & Pitt, M. A. (2006). Model selection by normalized maximum likeli-  
23 hood. *Journal of Mathematical Psychology, 50*, 167-179.  
24  
25  
26  
27 10 Myung, I. J. & Pitt, M. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach.  
28  
29 11 *Psychonomic Bulletin & Review, 4(1)*, 79-95.  
30  
31  
32 12 Myung, I. J., Pitt, M. & Navarro, D. J. (2007). Does response scaling cause the generalized context  
33 model to mimic a prototype model? *Psychonomic Bulletin & Review, 14(6)*, 1043-1050.  
34  
35  
36  
37 14 Navarro, D. J. & Lee, M. D. (2004). Common and distinctive features in stimulus representation: A  
38  
39 15 modified version of the contrast model. *Psychonomic Bulletin & Review, 11*, 961-974.  
40  
41  
42 16 Pitt, M. A., Myung, I. J. & Zhang, S (2002). Toward a method of selecting among computational  
43  
44 17 models of cognition. *Psychological Review, 109*, 472-491.  
45  
46  
47 18 Purdy, B.P. & Batchelder, W.H. (in press). A context free language for binary multinomial processing  
48  
49 19 tree models. *Journal of Mathematical Psychology*  
50  
51  
52 20 Raftery, A. E. (1999). Bayes factor and BIC. *Sociological Methods & Research, 27*, 411-427.  
53  
54  
55 21 Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New  
56  
57 22 York: Springer-Verlag.  
58  
59  
60

## MULTINOMIAL PROCESSING TREE MODELS

34

- 1  
2  
3  
4  
5 1 Riefer, D. M. & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive  
6  
7 2 processes. *Psychological Review*, 95, 318-339.  
8  
9  
10 3 Riefer, D.M. & Batchelder, W.H. (1991). Statistical inference for multinomial processing tree mod-  
11  
12 4 els. In Jean-Paul Doignon and Jean-Claude Falmagne (Eds.). *Mathematical psychology: Current*  
13  
14 5 *developments*. New York: Springer-Verlag, 313-335.  
15  
16 6 Riefer, D.M., & Batchelder, W.H. (1995) A multinomial modeling analysis of the recognition-failure  
17  
18 7 paradigm. *Memory & Cognition*, 23, 611-630.  
19  
20  
21 8 Riefer, D.M., Hu, X. & Batchelder, W.H. (1994) Response strategies in source monitoring. *Journal*  
22  
23 9 *of Experimental Psychology: Learning, Memory, and Cognition*, 1994, 20, 680-693.  
24  
25  
26 10 Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D. & Manifold, V. (2002). Cognitive  
27  
28 11 psychometrics: Assessing storage and retrieval deficits in special populations with multinomial  
29  
30 12 processing tree models. *Psychological Assessment*, 14, 184-201.  
31  
32  
33 13 Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Infor-*  
34  
35 14 *mation Theory*, 42, 40-47.  
36  
37  
38 15 Rissanen, J. J. (2001). Strong optimality of the normalized ML models as universal codes and  
39  
40 16 information in data. *IEEE Transactions on Information Theory*, 47, 1712-1717.  
41  
42  
43 17 Rose, R. G., Rose, P. R., King, N. & Perez, A. (1975). Bilingual memory for related and unrelated  
44  
45 18 sentences. *Journal of Experimental Psychology: Human Learning and Memory*, 1, 599-606.  
46  
47  
48 19 Saegert, J., Hamayan, E. & Ahmar, H. (1975). Memory for language of input in polyglots. *Journal*  
49  
50 20 *of Experimental Psychology: Human Learning and Memory*, 5, 607-613.  
51  
52  
53 21 Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.  
54  
55  
56 22 Schweickert, R. (1993). A multinomial processing tree model for degradation and reintegration in  
57  
58 23 immediate recall. *Memory and Cognition*, 21, 168-175  
59  
60

- 1  
2  
3  
4  
5 1 Schweickert, R. & Chen, S. (2008). Tree inference with factors influencing processes in a processing  
6  
7 2 tree. *Journal of Mathematical Psychology*, 52, 158-183  
8  
9  
10 3 Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate anal-  
11  
12 4 ysis. *International Statistical Review*, 56, 49-62.  
13  
14  
15 5 Smith, J. B. & Batchelder, W. H. (in press). Beta-MPT: Multinomial processing tree models for  
16  
17 6 addressing individual differences. *Journal of Mathematical Psychology*  
18  
19  
20 7 Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic*  
21  
22 8 *Bulletin & Review*, 14, 779-804.  
23  
24  
25 9 Wagenmakers, E. -J. & Waldorp, L. (2006). Editors' introduction. *Journal of Mathematical Psychol-*  
26  
27 10 *ogy*, 50, 99-100.  
28  
29  
30 11 Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Socio-*  
31  
32 12 *logical Methods & Research*, 27, 359-397  
33  
34  
35 13 Wu, H., Myung, J. I. & Batchelder, W. H. (submitted). On the minimum description length com-  
36  
37 14 plexity of multinomial processing tree models. *Journal of Mathematical Psychology*  
38  
39  
40  
41

#### 15 Acknowledgements

16 This paper is based on Hao Wu's Master of the Arts thesis submitted to the Ohio State  
17 University in July 2006. It is supported in part by National Institute of Health Grant R01-  
18 MH57472 to JIM. Work on this paper by the third author was supported in part from a grant  
19 from the National Science Foundation: SES-0616657 to X. Hu and W. H. Batchelder (Co-  
20 PIs). We wish to thank Mark A. Pitt and Michael W. Browne for valuable feedbacks  
21 provided for the project. Correspondence concerning this article should be addressed to  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

*MULTINOMIAL PROCESSING TREE MODELS*

36

- 1 Hao Wu, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus,
- 2 OH 43210. E-mail: [wu.498@osu.edu](mailto:wu.498@osu.edu). Tel: 614-292-5510

For Review Only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## MULTINOMIAL PROCESSING TREE MODELS

Table 1: Summary of model selection results for source monitoring data from Rose et al. (1975, experiment 1). The data can be found in Batchelder and Riefer (1990, Table 7). See the main text for the description of the models. It should be noted that inequality constraints apply only to parameters  $D$  and  $d$  (but not to  $b$  or  $g$ ) when the corresponding equality constraints are not present. The row  $FIA'$  shows FIA values if inequality constraints are taken into account, while the  $FIA$  gives FIA values if those constraints are neglected. The total sample size is  $N = 1920$ .

Models	$M_0$	$M_g$	$M_b$	$M_d$	$M_D$	$M_{bg}$	$M_{dg}$	$M_{Dg}$
S	8	7	7	7	7	6	6	6
$-LML$	36.17	36.61	38.23	36.18	41.16	38.66	36.61	41.60
$C_{AIC}$	8	7	7	7	7	6	6	6
$AIC$	44.17	43.61	45.23	43.18	48.16	44.66	<b>42.61</b>	47.60
$C_{BIC}$	30.24	26.46	26.46	26.46	26.46	22.68	22.68	22.68
$BIC$	66.41	63.07	64.69	62.64	67.62	61.34	59.29	64.28
$C_{FIA}$	22.2	19.7	19.4	20.7	20.4	16.9	18.2	17.8
$FIA$	58.4	56.3	57.6	56.9	61.6	55.6	54.8	59.4
$C'_{FIA}$	20.9	18.4	18.0	20.0	19.7	15.5	17.5	17.1
$FIA'$	57.1	55.0	56.2	56.2	60.9	54.2	54.1	58.7

Models	$M_{db}$	$M_{Db}$	$M_{Dd}$	$M_{dbg}$	$M_{Dbg}$	$M_{Ddg}$	$M_{Ddb}$	$M_{Ddbg}$
S	6	6	6	5	5	5	5	4
$-LML$	38.30	41.84	41.55	<b>38.73</b>	42.28	41.98	42.31	42.74
$C_{AIC}$	6	6	6	5	5	5	5	4
$AIC$	44.30	47.84	47.55	43.73	47.28	46.98	47.31	46.74
$C_{BIC}$	22.68	22.68	22.68	18.90	18.90	18.90	18.90	15.12
$BIC$	60.98	64.52	64.23	<b>57.63</b>	61.18	60.88	61.21	57.86
$C_{FIA}$	17.8	17.3	18.5	15.1	14.6	15.8	15.3	12.5
$FIA$	56.1	59.1	60.1	53.8	56.9	57.8	57.6	55.2
$C'_{FIA}$	17.1	16.6	18.5	14.4	13.9	15.8	15.3	12.5
$FIA'$	55.4	58.4	60.1	<b>53.1</b>	56.2	57.8	57.6	55.2

## MULTINOMIAL PROCESSING TREE MODELS

38

Table 2: Summary of model selection results for pair-clustering data from Batchelder and Riefer (1980, experiment 1a). See the main text for the description of the models. The total sample size is  $N = 3220$ .

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

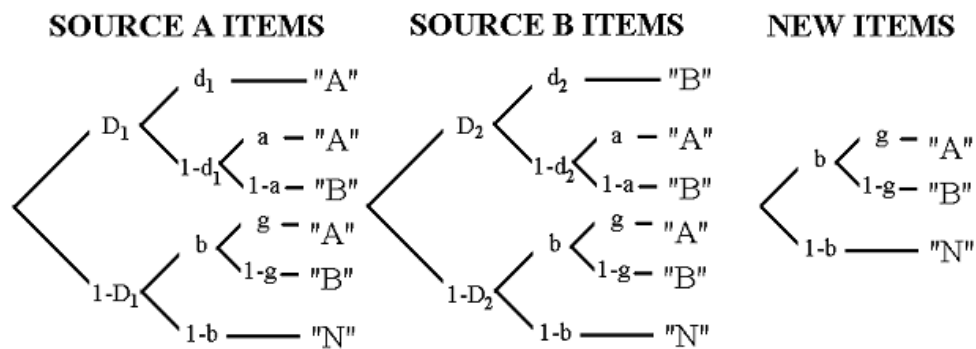


Figure 1. The one-high-threshold (1HT) multinomial processing tree model of source monitoring. The parameters are defined as follows:  $D_1$  (detectability of source A items);  $D_2$  (detectability of source B items);  $d_1$  (source discriminability of source A items);  $d_2$  (source discriminability of source B items);  $a$  (guessing that a detected but nondiscriminated item belongs to source A);  $b$  (guessing "old" to a nondetected item);  $g$  (guessing that a nondetected item biased as old belongs to source A category). Adapted from Batchelder and Riefer (1990).

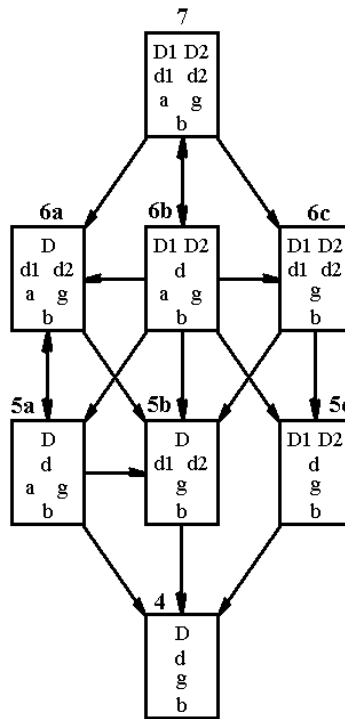


Figure 2. The nested hierarchy of eight versions of the 1HTM model in Figure 1, created by imposing successive constraints on the parameters. In the figure, the model parameters for each model are listed and a directed arrow from one model to another means that the second model is nested in the first.

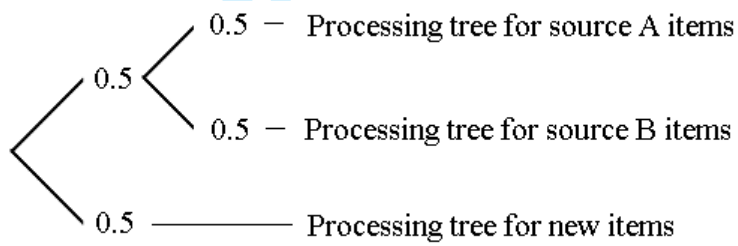


Figure 3. Example of combining the three processing trees in the 1HTM (shown in Figure 1) into one BMPT model. The sample sizes are assumed to be 250, 250 and 500 for source A items, source B items and new items, respectively.

MULTINOMIAL PROCESSING TREE MODELS

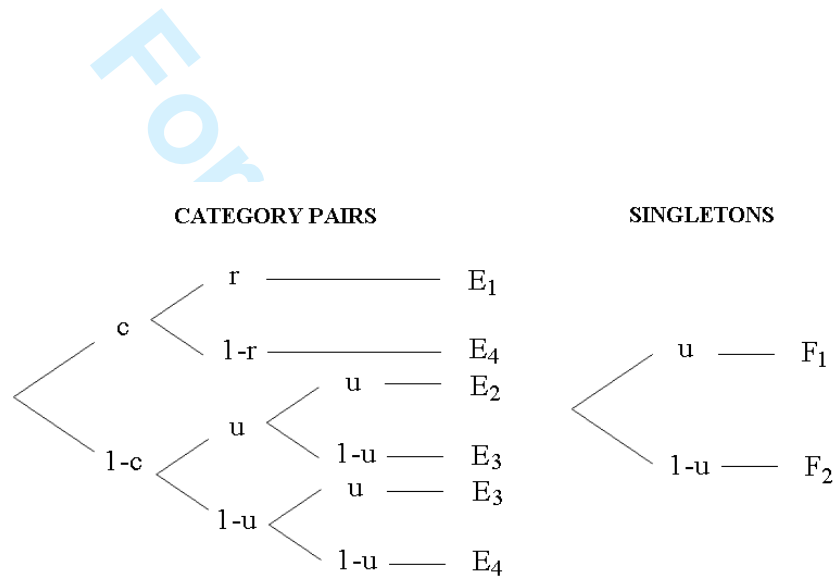


Figure 4. Batchelder and Riefer's (1999) multinomial processing tree model of pair-clustering.