Psychological Review
1990, Vol. 97, No. 4, 548-564

Copyright 1990 by the American Psychological Association, Inc.
0033-295X/90/$00.75

Multinomial Processing Models of Source Monitoring

William H. Batchelder
School of Social Sciences
University of California, Irvine

David M. Riefer
California State University, San Bernardino

This article presents a family of processing models for the source-monitoring paradigm in human memory. Source monitoring and the special case of reality monitoring have become very popular as paradigms to assess memory deficits in various subject populations. The paradigm provides categorical data that satisfy product-multinomial constraints, and this lends it nicely to multinomial modeling with processing-tree structures as described in Riefer and Batchelder (1988). The models developed herein are based on ideas from high-threshold signal-detection models, and they involve item-detection parameters, source-identification parameters, and various parameters reflecting guessing biases. The purpose of the models is to provide separate, theoretically based measures of old-item detection and source discrimination. The models may strengthen traditional analyses that are based on ad hoc statistics, as well as avoid flawed interpretations that the traditional analyses may produce. The usefulness of the models is revealed by analyzing published data sets from the areas of reality monitoring and bilingual memory.

An increasingly popular paradigm in memory research is source monitoring (e.g., Eich & Metcalfe, 1989; Hashtroudi, Johnson, & Chrosniak, 1989; Lindsay & Johnson, 1989). In a source-monitoring experiment, subjects study items from two or more different sources. For example, the sources may be different points of origin for the items (e.g., self-produced vs. experimenter produced, or List 1 vs. List 2). Another possibility involves items that come from one point of origin but that differ on some fundamental attribute (e.g., language or modality of input). Once these items have been studied, a recognition memory test is given to measure separately the ability to detect old items from new distracters and the ability to discriminate the source or attribute of detected old items.

Early work using the source-monitoring paradigm concerned the ability of subjects to retain incidental features of items. This research showed that source discrimination was often very accurate for various types of incidental information, such as type font (e.g., Hintzman, Block, & Inskeep, 1972; Light & Berger, 1976), sex of voice (e.g., Craik & Kirsner, 1974; Light, Stansbury, & Rubin, 1973), and input modality (e.g., Bray & Batchelder, 1972; Hintzman et al., 1972). This work was concerned with the issue of which attributes of an item get into long-term memory and the degree of effort required to code these various attributes in memory.

This research was supported by National Science Foundation Grant BNS-8910552.

We are grateful to Xiangen Hu, who provided both conceptual and computer assistance to us at all stages of the project and wrote the computer program that is described in the Conclusion section. We are also thankful for the support of the Irvine Research Unit in Mathematical Behavioral Sciences in carrying out our research.

Correspondence concerning this article and requests for the computer program should be addressed to William H. Batchelder, School of Social Sciences, University of California, Irvine, California 92717.

In the 1980s, reality monitoring (Johnson & Raye, 1981) emerged as a major paradigm in memory research. Reality monitoring is a special case of source monitoring, where subjects attempt to differentiate between memories of real and imagined events. For example, a word list may be presented with half the items accompanied by a picture depicting the word and the other half accompanied by an instruction for the subject to imagine a picture depicting the word (e.g., Johnson, Raye, Foley, & Kim, 1982). Johnson and Raye (1981) have developed an interesting theory of reality monitoring, and they, along with coworkers, have applied it successfully to understand results in a large number of reality-monitoring tasks (e.g., Durso & Johnson, 1980; Foley, Johnson, & Raye, 1983; Johnson, Kahan, & Raye, 1984; Johnson, Raye, Foley, & Foley, 1981).

The reality-monitoring paradigm has become a standard tool in several areas of psychology to show that particular subject populations may have an impairment in their ability to discriminate between perceived and imagined attributes of an item, independent of their ability to detect old from new items. Studies in this area include ones showing that young children may confuse certain sources more than older children may (e.g., Foley & Johnson, 1985), schizophrenic people may show an impairment in source discrimination (e.g., Harvey, 1985), and advanced age and age-related disease may result in more impairment in source discrimination than in item detection (e.g., Hashtroudi et al., 1989; McIntyre & Craik, 1987; Mitchell, Hunt, & Schmitt, 1986).

Other areas of psychology have been informed by source-monitoring experiments. These include the nature of field dependence in personality (Durso, Reardon, & Jolly, 1985), the generation effect in memory (Rabinowitz, 1990; Voss, Vesonder, Post, & Ney, 1987), bilingual language coding in psycholinguistics (Rose, Rose, King, & Perez, 1975; Saegert, Hamayan, & Ahmar, 1975), memory deficits in patients with amnesia or frontal lobe lesions (Janowsky, Shimamura, &

Squire, 1989; Shimamura & Squire, 1987), and the area of eyewitness memory (Lindsay & Johnson, 1989).

It should be evident from the studies just mentioned that source-monitoring experiments attempt to examine two different types of memory: recognition memory for old items and memory for the source of items. However, we have surveyed more than 50 papers that report a source-monitoring paradigm, and we have found that there are no generally accepted measures of these quantities. Instead, a variety of ad hoc statistical approaches have been adapted for disentangling discriminability of source from overall detectability of old items, including the Kruskal-Wallis gamma score (Voss et al., 1987), identification-of-origin scores (e.g., Finke, Johnson, & Shyi, 1988; Johnson, Foley, & Leach, 1988), and hit and false-alarm rates for source identification (e.g., Anderson, 1984; Reardon, Durso, Foley, & McGahan, 1987).

In addition, none of the studies we have surveyed includes any substantive model to analyze its data. The lack of a mathematical model for source monitoring is unfortunate because the separate measurement of item detection from source discrimination, as well as from various guessing biases, is even more complicated than is separate measurement of recognition strength from response bias in standard recognition memory paradigms. This is true because the standard yes–no recognition memory paradigm is a special case of the source-monitoring paradigm, where there is only one source. Standard practice in the memory field is now to analyze yes–no recognition memory data with mathematical models that are based on signal-detection principles (e.g., see Klatzky, 1975), and the source-monitoring area could profit greatly from a similar treatment.

Recently, Riefer and Batchelder (1988) presented a detailed discussion of one class of mathematical models, known as *multinomial models*, that can be used to measure cognitive processes. As we show in the next section, the source-monitoring paradigm is ideally suited for multinomial modeling with processing-tree structures. The goal of a multinomial model of source monitoring is to provide a psychologically interpretable parameterization of the probability structure of the underlying sample space for the paradigm. Then the valid measurement and comparison of various cognitive capacities in source monitoring can be accomplished by estimating the model's parameters and conducting hypotheses tests of the parameters across conditions.

In this article, we develop a family of multinomial processing-tree models for source monitoring in the spirit of Riefer and Batchelder (1988) for separately measuring item detection and source discrimination. Our goal is to show, using several published data sets, that our models can play a useful role in interpreting source-monitoring data. We further show that sometimes the multinomial models support and augment the authors' original analyses with ad hoc statistical measures but at other times, analyses with the models suggest dramatic revisions in the authors' original interpretations of their data.

The article is organized into two main sections: (a) Multinomial Models for Source Monitoring and (b) Empirical Examples. The first main section analyzes the nature of the representation of a source-monitoring experiment and discusses some of the standard treatments of data. Then, a nested family of psychologically interpretable processing models is presented and analyzed in detail, and finally recommendations are given for using the models as measurement tools. We argue on theoretical grounds that source-monitoring experiments should be analyzed in terms of processing models such as the ones we provide. In the second main section of the article, we use the models to analyze and reinterpret data (a) from two reality-monitoring experiments and (b) from two bilingual source-monitoring experiments.

Multinomial Models for Source Monitoring

Data Representation

In a source-monitoring experiment with two sources, A and B, the final memory test consists of a mix of old A and B items along with new distracters, N. The subject is required to classify each item as an A, B, or N. The data from an individual subject can be described completely by the frequency table T given by

$$\mathbf{T} = \begin{array}{c} \\ \text{Source A} \\ \text{Source B} \\ \text{N} \end{array} \begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{N} \end{array} \\ \begin{bmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ Y_{31} & Y_{32} & Y_{33} \end{bmatrix} \end{array} \begin{array}{c} \text{Response} \\ \begin{array}{c} Y_1 \\ Y_2, \\ Y_3 \end{array} \end{array} \tag{1}$$

where Y_{ij} is the frequency of j-type responses to i-type items. The marginal frequency $Y_i = Y_{i1} + Y_{i2} + Y_{i3}$ is the total number of i-type items on the memory test, and usually $Y_1 = Y_2$ and $Y_3 = Y_1 + Y_2$.

We have surveyed many articles that report source-monitoring experiments that fit into the structure of Equation 1; however, only a few of them report data in sufficient detail to reconstruct even a group-level data table in the form of T. This is unfortunate because the appropriate sample space for a source-monitoring experiment consists exactly of events E_{ij} corresponding to the observable frequencies Y_{ij} in Equation 1. Thus, any measure one might want to compute can be obtained from the data representation in Equation 1.

In the next section, we define and analyze a nested family of models for source monitoring, and in the subsequent section, we show how they can be used to analyze source-monitoring experiments. However, before presenting our models, it is instructive to consider some of the potential problems with traditional approaches. By far the most frequently used method of data analysis is to compute three measures for each subject: hits (H), false alarms (F), and identification-of-origin scores (I), defined in terms of Equation 1 as follows:

$$\mathrm{H} = \frac{(Y_{11} + Y_{12}) + (Y_{21} + Y_{22})}{Y_1 + Y_2}, \tag{2}$$

$$\mathrm{F} = \frac{Y_{31} + Y_{32}}{Y_3}, \tag{3}$$

and

$$\mathrm{I} = \frac{Y_{11} + Y_{22}}{(Y_{11} + Y_{12}) + (Y_{21} + Y_{22})}. \tag{4}$$

Typically, separate between-groups statistical tests are conducted on each of these measures.

Several observations need to be made about the approach to data analysis represented by the statistics in Equations 2, 3, and 4. First, the statistics combine data across sources, and this practice may mask differential detectability and source discriminability for each source. Some authors (e.g., Anderson, 1984; Johnson et al., 1988; Voss et al., 1987) calculate identification-of-origin scores separately for each source; that is, they work with two identification-of-origin scores given by

$$I_i = \frac{Y_{ii}}{Y_{i1} + Y_{i2}}, \tag{5}$$

for $i = 1, 2$.

Second, if the source variable is ignored, H and F represent the usual measures of hits and false alarms in a yes–no recognition memory task. It is well documented that separate between-groups analyses of hits and false alarms can lead to a misleading picture of recognition memory data (see Klatzky, 1975, chap. 11), and for several decades standard practice has been to combine hits and false alarms into a statistic that measures item detectability, such as d' from signal-detection theory (e.g., Green & Swets, 1966). In fact, some authors, albeit a minority, combine hits and false alarms in this way for their source-monitoring data (e.g., Reardon et al., 1987).

A third observation is that I (in Equation 4) can be interpreted as an estimate of the conditional probability of a correct source identification given that the item was correctly reported as old. Whereas the statistic I has a certain face validity for measuring source discriminability, we (Batchelder & Riefer, 1986; Riefer & Batchelder, 1988) and others have shown that surface statistics such as I often reflect the combined effects of different cognitive processes. Therefore, these types of statistics may be hard to interpret and in some circumstances may even be misleading. In fact, some authors (e.g., Johnson et al., 1988, Rabinowitz, 1990) have pointed out that I scores that are conditional on correct recognition are dependent on the overall level of recognition in an experiment. This relation between I scores and overall recognition can easily be seen by comparing the numerator and denominator of Equations 2 and 4, respectively. As a consequence, the statistic I may be difficult to interpret when overall recognition memory changes across the experimental conditions in an experiment. Also, as shown at the end of the next section, I scores may be a function of various nuisance factors, such as guessing and response biases, as are separate hit and false-alarm scores in recognition memory. Without a substantive model, there is no rational way to rid the statistic I of these nuisance factors.

The first step in constructing a model for a psychological paradigm is to determine both the sample space and the most general probability model for the paradigm. The data representation in Equation 1 shows that the source-monitoring task lends itself naturally to multinomial modeling as described in Riefer and Batchelder (1988). To see this, note that if responses within each source type are independent and identically distributed over response classes, then the data for each source type are trinomially distributed:

$$\Pr[(Y_{i1}, Y_{i2}, Y_{i3})] = Y_i! \prod_{j=1}^{3} [p_{ij}^{Y_{ij}}]/Y_{ij}! \tag{6}$$

for $i = 1, 2, 3$ and

$$\mathbf{p}_i = (p_{i1}, p_{i2}, p_{i3}) \in \mathcal{G}_3,$$

where $\mathcal{G}_3$ is the three-dimensional probability simplex given by

$$\mathcal{G}_3 = \{(x_1, x_2, x_3) | 0 \le x_i \le 1, \sum_{i=1}^{3} x_i = 1\}. \tag{7}$$

With the further assumption of independence between sources, the general probability model for the entire data table **T** in Equation 1 has the structure of a product multinomial (Read & Cressie, 1988), a special case of a joint multinomial model (Riefer & Batchelder, 1988) given from Equations 6 and 7 as

$$\Pr(\mathbf{T}, \mathbf{P}) = \prod_{i=1}^{3} \Pr[(Y_{i1}, Y_{i2}, Y_{i3})], \tag{8}$$

where

$$\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3) \in \mathcal{G}_3 \times \mathcal{G}_3 \times \mathcal{G}_3 \equiv G. \tag{9}$$

Each **P** in the general model is described by nine probabilities; however, because of the marginal, product-multinomial constraints, there is a total of 6 *dfs* in the characterization of G (see Equation 7). The key to developing a substantive multinomial model is to postulate parameterized processing trees that permit one to derive $\mathbf{P} \in G$ in terms of psychologically meaningful parameters.

Both the general method and some examples of multinomial modeling with processing-tree structures are described in detail in Riefer and Batchelder (1988), and in this article, we conform as much as possible to the notation and sequencing of that article without assuming it as a prerequisite. Unlike the general-purpose data descriptions that log-linear multinomial models provide, multinomial models with processing-tree structures provide a way to measure directly the cognitive capacities that underlie a particular task such as source monitoring.

Processing-Tree Models

As mentioned earlier, a source-monitoring model should provide a way to measure separately the overall detectability of old items from the ability to discriminate the source of items. Models based on auditory signal-detection principles (Green & Swets, 1966) have facilitated the analysis of recognition memory paradigms, so our goal in this section is to extend signal-detection notions to model source monitoring.

Source monitoring is analogous to a standard yes–no signal-detection task, except that the "signal" is one of two sources of items. For example, in auditory detection, the signals might be two frequencies at possibly different energy levels. Then the ability to detect the signal might be driven solely by an energy detector, whereas ability to discriminate the frequency would also involve some type of frequency detector. There are several explicit ways to create signal-detection models with both types of detectors, but in the interest of simplicity, we decided to build our source-monitoring models on high-threshold versions of the corresponding signal-detection models (e.g., see Laming,

1973, chap. 6). High-threshold models for signal detection are probably oversimplified if viewed as a precise theory of auditory signal detection; however, they are known to provide useful measurement tools in recognition memory, as well as in many other nonpsychophysical areas, such as test theory (Lord & Novick, 1968) and cultural consensus analysis (Batchelder & Romney, 1988).

Our joint multinomial model for source monitoring is presented in Figure 1, which displays separate processing-tree models for items from Source A, Source B, and new distracter items. The model assumes that subjects' responses to items in a source-monitoring task are a function of a series of hypothetical cognitive processes: stimulus detection, source discrimination, and various response biases, which are described next.

Stimulus Detection

On test trials, an old item is assumed to be either detected as an old item within memory or to remain undetected. Define D_1 and D_2 to be the probabilities of correct detection for old A and B items, respectively. The values of D_1 and D_2 are allowed to differ for each source because experimental factors may create different detection rates.

Source Discrimination

Conditional on an old item's being correctly detected, the subject either is or is not correctly able to identify the source of that item. Define d_1 and d_2 to be the probabilities of discriminating the source of detected A and B items, respectively. Again, these parameters are allowed to differ over source.

Response Biases

The remaining parameters are nuisance parameters denoting various response biases. Parameter b is the probability of responding "old" to a nondetected item. When an item is detected but the source is not discriminated, with probability $D_i(1 - d_i)$, then parameter a represents the probability of guessing that the item belongs to Source A. Parameter g is also the probability of guessing that an item belongs to Source A, but for undetected items that the subject has guessed are old. In the general case, we allow the possibility that parameters a and g may differ because on psychological grounds one might bias detected and nondetected items differently.

There are two obvious features of the model in Figure 1 that are important to note. First, the individual parameters are conditional process probabilities corresponding to the links in the trees, and they occur in more than one place on the trees. Second, there are often several branches in the trees that combine into the same observable category. In fact, there are 15 branches in the model of Figure 1 that combine into the nine observable events E_{ij}, corresponding to the Y_{ij} frequencies of the data table T in Equation 1. It is easy to write equations for the $\Pr(E_{ij})$ by simply summing branch probabilities over the combined classes, and these equations are, for Source A items,

$$p_{11} = D_1 d_1 + D_1(1 - d_1)a + (1 - D_1)bg, \quad (10a)$$

$$p_{12} = D_1(1 - d_1)(1 - a) + (1 - D_1)b(1 - g), \quad (10b)$$

$$p_{13} = (1 - D_1)(1 - b); \quad (10c)$$

for Source B items,

$$p_{21} = D_2(1 - d_2)a + (1 - D_2)bg, \quad (11a)$$

$$p_{22} = D_2 d_2 + D_2(1 - d_2)(1 - a) + (1 - D_2)b(1 - g), \quad (11b)$$

$$p_{23} = (1 - D_2)(1 - b); \quad (11c)$$

and for new distracters,

$$p_{31} = bg, \quad (12a)$$

$$p_{32} = b(1 - g), \quad (12b)$$

$$p_{33} = (1 - b). \quad (12c)$$

The model in Figure 1 has a total of seven parameters in the parameter space

$$\Omega_7 = \{\langle D_1, D_2, d_1, d_2, b, a, g\rangle\}, \quad (13)$$

where each parameter is a probability that measures some processing capacity. Because the general model of Equations 8 and 9 has only 6 *df*s, the model denoted by Ω_7 is technically nonidentifiable (see Greeno & Steiner, 1964, or Bishop, Fienberg, & Holland, 1975). This means that a given **P** in Equation 9 may be generated, through Equations 10, 11, and 12, by multiple sets

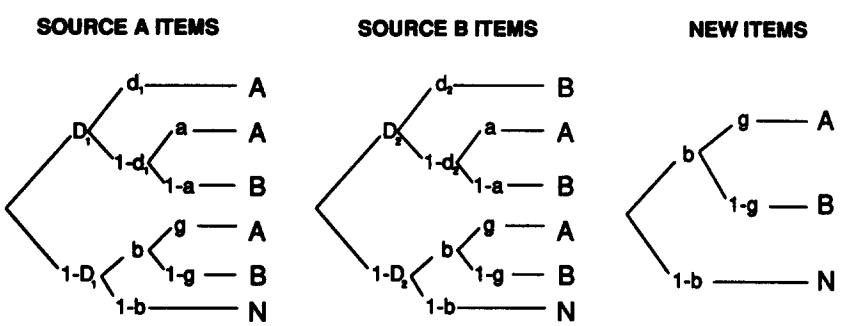

Figure 1. The seven-parameter, joint multinomial model for source monitoring. (D_1 = detectability of the Source A items; D_2 = detectability of the Source B items; d_1 = source discriminability for the Source A items; d_2 = source discriminability for the Source B items; a = guessing that a detected but nondiscriminated item belongs to Source A; b = bias for responding "old" to a nondetected item; g = guessing that a nondetected item belongs to Source A.)

of parameters in Ω_7. This complicates the analysis of the model because classical minimum-distance methods of parameter estimation and hypothesis testing, such as maximum likelihood and minimum X^2 (e.g., Lehmann, 1983; Read & Cressie, 1988), may fail to yield unique estimates of the parameters. The main motivation for creating the substantive model is to provide a tool to measure the parameter values and to make comparisons between groups by testing statistical hypotheses about the parameters, and minimum-distance methods provide a familiar and well-studied approach to such tasks. One way out of the problem of lack of identifiability of Model Ω_7 is to adopt a Bayesian perspective. For example, Chechile and Meyer (1976) showed that nonidentifiable models of the type of Ω_7 can be estimated by Bayesian methods if independent prior uniform [0, 1] distributions are placed on each parameter. A second, and for us preferable, way to handle the nonidentifiability problem is to consider special cases of Ω_7 by imposing restrictions on the parameters. We prefer this approach because the uniform priors create arbitrary advantages to certain parameter values and more important, because a large body of theoretical and practical knowledge exists for the classical, asymptotic analysis of identifiable models.

There are three kinds of restrictions on the current model that are of psychological interest: (a) Assume the detection parameters are equal, $D_1 = D_2$; (b) assume the discrimination parameters are equal, $d_1 = d_2$; and (c) assume the guessing rates are equal, $a = g$. By combining the presence or absence of each restriction independently, this creates eight possible models, which are presented in Figure 2. The model at the top of Figure 2 is the original 7-parameter model that is in Figure 1. The second row of Figure 2 presents three models, each with 6 parameters. Model 6a simplifies the general model by requiring the restriction that $D_1 = D_2$, Model 6b requires $d_1 = d_2$, and Model 6c requires $a = g$. In the next row are three 5-parameter versions of the model, each containing two of the restrictions. Finally, the version of the model at the bottom of Figure 2 is a 4-parameter model that results from all three restrictions, $D_1 = D_2$, $d_1 = d_2$, and $a = g$.

From a statistical viewpoint, Figure 2 represents a nested hierarchy of processing-tree models for source monitoring, each of which corresponds to a joint multinomial model for the data structure in Equation 1. These hierarchical relationships between the models are represented in Figure 2 as directed arrows connecting the models. Thus, for example, Model 5c is a nested subset of Model 6b or 6c, in the sense that any $\mathbf{P} \in G$ that satisfies Model 5c also satisfies Models 6b and 6c.

There are several additional facts and relationships about the models in Figure 2 that have both psychological and data-analytic significance. In order to describe these succinctly, some notational conventions adapted from Riefer and Batchelder (1988) need to be stated. Suppose x stands for one of the models in Figure 2; then the parameter space for model x, as well as the model itself, will be denoted by Ω_x. Also, the subset of $\mathbf{P} \in G$ that can arise from parameter values in Ω_x (through Equations 10, 11, and 12) will be denoted by Ω_x^*. Then, in terms of our notation, the nested hierarchy in Figure 2 satisfies

$$\Omega_x^* \subseteq \Omega_y^* \subseteq G$$

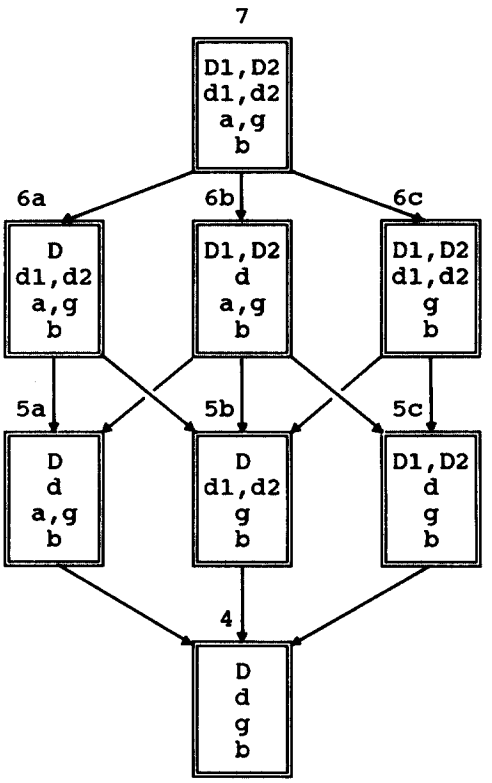

Figure 2. Nested hierarchy for the eight versions of the multinomial model depicted in Figure 1. (Directed arrows indicate subset relations between models. D_1 = detectability of the Source A items; D_2 = detectability of the Source B items; d_1 = source discriminability for the Source A items; d_2 = source discriminability for the Source B items; a = guessing that a detected but nondiscriminated item belongs to Source A; b = bias for responding "old" to a nondetected item; g = guessing that a nondetected item belongs to Source A.)

in case Ω_y has a directed arrow into Ω_x at a lower level (i.e., has more restrictions) in Figure 2. In terms of our notation, Model Ω_{5c} is given by

$$\Omega_{5c} = \{\langle D_1, D_2, d, b, g\rangle\},$$

and Ω_{5c}^* is the subset of G that can be generated by parameters in Ω_{5c}.

Each of the models in Figure 2 imposes some probabilistic restrictions on the nine probabilities characterizing the general model G in Equations 8 and 9. Some of these restrictions are inequality restrictions, and these do not result in a decrease in the number of functionally independent parameters; that is, there is no reduction in the number of degrees of freedom. A familiar example of such a restriction is the fact that the underlying hit probability is not less than the false-alarm probability in most yes–no signal-detection models (Green & Swets, 1966). Other restrictions involve equalities between expressions involving the probabilities in G, and these usually result in a reduction of the number of independent variables characterizing G (a reduction in degrees of freedom). A familiar example of such a restriction is Luce's constant ratio rule for probabilistic choice models (Luce, 1959).

We have undertaken a complete study of the restrictions implied by each model in Figure 2. The results are reported next because they support some interesting psychological restrictions among the models shown in Figure 3 that go beyond those presented in Figure 2. However, the mathematical details behind the restrictions are presented in the Appendix.

First, assume Ω_7 holds. Then the $\mathbf{P} \in G$ of Equation 9 are subject to the inequality constraints

$$p_{33} \geq \frac{p_{i3}p_{3j}}{p_{ij}}, \tag{14}$$

for $i, j = 1, 2$ and

$$\frac{p_{33} - p_{23}}{p_{33} - p_{13}} \leq \frac{p_{33}(1 - p_{21}) - p_{23}(1 - p_{31})}{p_{33}p_{12} - p_{13}p_{32}}. \tag{15}$$

When Ω_{6b} is assumed, no further constraints are added; however, when Ω_{6c} is assumed, the additional inequality constraints

$$p_{ij} \leq \frac{p_{3j}(1 - p_{i3})}{1 - p_{33}}, \tag{16}$$

for $i \neq j$ and $i, j = 1, 2$, must be satisfied. Finally, when Ω_{5c} is assumed, the dimensionally reducing equality constraint

$$\frac{p_{33}p_{12} - p_{13}p_{32}}{p_{33}p_{21} - p_{23}p_{31}} = \frac{p_{32}(p_{33} - p_{13})}{p_{31}(p_{33} - p_{23})} \tag{17}$$

is added to the constraints in Equations 14, 15, and 16.

Models Ω_{6a}, Ω_{5a}, Ω_{5b}, and Ω_4 all involve the assumption of equality of detection rates, $D_1 = D_2$, which leads to the dimensionally reducing equality constraint

$$p_{13} = p_{23}. \tag{18}$$

These four models also satisfy, respectively, the constraints on Models Ω_7, Ω_{6b}, Ω_{6c}, and Ω_5 described in Equations 14 through 17.

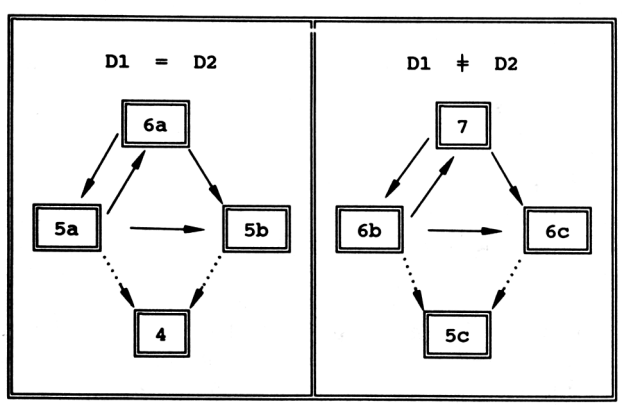

Figure 3. Additional nesting relations between the models for source monitoring. (Double arrows indicate models that are equivalent. Single solid arrows indicate a proper subset relation between models that have the same number of degrees of freedom. Single dotted arrows indicate a nested relation between models that differ by 1 *df*. D_1 = detectability of the Source A items; D_2 = detectability of the Source B items.)

Figure 2 provides a number of subset, or nesting, relationships between the submodels of Ω_7 that follow obviously from the nature of the parameter restrictions. There are several other nesting relationships among the submodels that follow from the preceding analysis. From a structural standpoint, it turns out that the most productive restriction to consider involves whether item-detection rates are the same for both sources ($D_1 = D_2$). Figure 3 shows additional nesting relations among the models that require $D_1 = D_2$ (left panel) and the ones that do not (right panel).

The two panels of Figure 3 have an interesting structural isomorphism. First, for both panels, the topmost model (with the most parameters) is equivalent to the one below and to the left, which has one less parameter; that is, in set notation, $\Omega_7^* = \Omega_{6b}^*$ and $\Omega_{6a}^* = \Omega_{5a}^*$. Both Ω_7 and Ω_{6a} are nonidentifiable; however, their counterparts, namely Ω_{6b} and Ω_{5a}, are identifiable, have unique maximum likelihood estimators (see the Appendix), and have 6 and 5 *df*s, respectively. The reason why Ω_{6a} has only 5 *df*s, despite six parameters, is the dimensionally reducing restriction, $p_{13} = p_{23}$ in Equation 18.

Figure 3 shows that Ω_{6c}^* is a proper subset of Ω_{6b}^*, or in set notation,

$$\Omega_{6c}^* \subset \Omega_{6b}^*. \tag{19}$$

Also:

$$\Omega_{5b}^* \subset \Omega_{5a}^*. \tag{20}$$

Both cases involve comparing the restriction of equal discrimination rates, $d_1 = d_2$, with the restriction of equal guessing biases, $a = g$. Equations 19 and 20 have disappointing psychological implications because they imply that one cannot statistically differentiate these two substantively different restrictions. Although the $d_1 = d_2$ assumption strictly nests the $a = g$ assumption in both panels of Figure 3, the corresponding pairs of models each have the same number of degrees of freedom. Thus, if a data table T is argued to satisfy $a = g$ and $d_1 \neq d_2$, it can equally well be argued to satisfy $d_1 = d_2$ and $a \neq g$.

This is unfortunate because theorists may be interested in determining whether items from different sources differ in their source memory. This would involve testing whether d_1 is significantly different from d_2. An example of this can be found in a study by Voss et al. (1987), who explored whether source memory is better for self-generated than for externally presented items. However, the equivalence relations in Equations 19 and 20 sometimes prevent a complete answer to this question. To test the hypothesis of differential discrimination, one would use Model 5b or 6c, because these models assume that d_1 can differ from d_2. If the null hypothesis that $d_1 = d_2$ is rejected for a set of data, one may be tempted to conclude that source memory significantly differs for the two sources. But the same data set could be analyzed with the equivalent Models 5a and 6b. These models will often fit the data equally well, but they represent a different interpretation of the data—that discriminability is the same for the two sources, but that the subjects' response biases (a and g) differ.

The impasse in data interpretation implied by Equations 19 and 20 is analogous to a phenomenon in Markov learning theory, namely the inability to statistically differentiate the hy-

pothesis of learning only on errors from the assumption of equal learning rates on errors and successes for the general all-or-none Markov learning model (Greeno & Steiner, 1964). As in that case, one has to seek other sorts of empirical observations to separate the hypotheses. One solution to this problem is to conduct a source-monitoring experiment with three sources rather than two. In this case, the data table T in Equation 1 becomes a 4×4 table of frequencies, and the substantive model corresponding to Ω_7 has 12 dfs and 11 parameters. It is a straightforward exercise, in light of the results and methods in the Appendix, to show that this model is identifiable and the hypotheses of equal discrimination rates can be statistically differentiated from the hypotheses of equal guessing biases for detected and undetected items.

The final point for Figure 3 is that when both equal discrimination rates and equal guessing biases are assumed, the resulting models (Ω_4 and Ω_{5c}, respectively) each drop a degree of freedom and can be statistically differentiated from those above them in their respective subhierarchies. This follows from Equation 17 coupled, in the first case, with Equation 18.

Model Analysis of Traditional Measures

The problems with traditional statistical approaches to source-monitoring data discussed earlier can now be shown dramatically in terms of the substantive models in Figure 2. For example, if we assume Model Ω_7 is true, then the identification-of-origin measure I in Equation 4 is given, in terms of Equations 10 through 12, by

$$\begin{aligned} \mathrm{I} = \{&D_1[d_1 + (1 - d_1)a - bg] \\ &+ D_2[d_2 + (1 - d_2)(1 - a) - b(1 - g)] + b\}/ \\ &\{1 - (1 - b)[1 - (D_1 + D_2)/2]\}. \end{aligned} \tag{21}$$

The unwieldy expression in Equation 21 shows that I depends on *all seven parameters* of Ω_7, rather than just d_1 and d_2, which are the source-discrimination parameters of the model.

Even in the extremely simple case that $D_1 = D_2$, $d_1 = d_2$, and $g = a = 1/2$, we obtain an expression for I in all three remaining parameters, namely

$$\mathrm{I} = \frac{D(1 - b + d) + b}{2[D(1 - b) + b]}. \tag{22}$$

Equation 22 shows that I depends systematically on all three of the model's parameters. In fact, simple computation with partial derivatives shows that I uniformly increases with d, as desired; however, it also uniformly increases with D and uniformly decreases with b. The fact that I increases with D is a model-specific justification for the concerns expressed by Johnson et al. (1988) and Rabinowitz (1990), namely that the statistic I depends on the overall recognition level. Thus, the statistic I is not a good measure of source discrimination except in the unlikely case that the groups being compared are equated on *all other* cognitive processing capacities. Otherwise, it could change systematically between groups without any change in source-discrimination capacity, as we demonstrate with actual data in the next section.

Because of these potential problems with the traditional empirical statistics, we recommend to anyone conducting research involving source monitoring that some type of processing-tree model be used to analyze data, at least as a supplement to traditional empirical measures. To put the matter succinctly, there is simply no theory-free way to interpret source-monitoring data, and furthermore, there are good theoretical reasons to doubt the validity of an analysis that is based on the traditional measures in Equations 2 through 5.

Model-Based Data Analysis

In this subsection, we show how source-monitoring data can be analyzed with the processing models in Figure 2. When actually applying the model, one question that researchers need to address is which version of the model should be used to analyze the data for a particular study. Ideally, this should be decided by which version provides the best fit to the data, but this decision may not always be routine. For example, different versions of the model may produce different but quantitatively similar fits. Also, nested versions of the model will, of course, produce somewhat worse fits because they have more restrictions, but this will be balanced by the fact that they have more degrees of freedom.

On the basis of the bifurcation of models in Figure 3, we recommend first deciding whether the detection rates are equal. As shown in Equation 18, the restriction of equal detection rates leads naturally to the following statistical hypotheses:

$$\begin{aligned} \mathrm{H}_0&: p_{13} = p_{23} \\ \mathrm{H}_1&: p_{13} \neq p_{23}. \end{aligned} \tag{23}$$

These hypotheses can easily be decided by the familiar chi-square test of equality of independent proportions defined from T in Equation 1, namely $P_{13} = Y_{13}/Y_1$ and $P_{23} = Y_{23}/Y_2$.

Once the decision about detection rates is made, then the choice among models to use should be among the four in the appropriate panel of Figure 3. Because Models Ω_7 and Ω_{6a} are nonidentifiable, we do not recommend using them to analyze data. Also, because a statistical decision between Ω_{5a} and Ω_{5b} and between Ω_{6b} and Ω_{6c} cannot be justified (because of the equality relations in Equations 19 and 20, discussed earlier), a nonstatistical choice between them must be made. On psychological grounds, the assumption of equality of guessing biases ($a = g$) seems more acceptable than the assumption of equality of discrimination rates ($d_1 = d_2$). This is because the guessing biases are less fundamental cognitive capacities than are the discrimination rates, and there are many source-monitoring paradigms where we would expect the discrimination rates to differ.

The preceding decision leads us to test Model Ω_{5b} versus Ω_4, if H_0 in Equation 23 is accepted, and Model Ω_{6c} versus Ω_{5c}, if H_0 is rejected. These two tests are easily carried out by likelihood ratio methods because they are between nested models that differ by 1 df. Basically, these tests are performed by computing the log-likelihood ratio statistic G^2, which is distributed asymptotically as a chi-square distribution. Thus, if sample sizes are sufficiently large, hypothesis tests between nested models can be conducted with a standard chi-square test. More detail on these techniques, along with statistical references, can be found in Riefer and Batchelder (1988), and the next section illustrates

the techniques with data from several source-monitoring experiments. Researchers following our suggestions to conduct these tests between models should be sensitive to the fact that if they reject either Ω_4 or Ω_{5c} (i.e., rejecting $d_1 = d_2$), there may be an alternative explanation of their data, namely one involving differing guessing biases. As stated earlier, if this issue is particularly important, a new experiment with three sources will permit a decision.

A final question of importance is whether the processing models are close enough to reality to use them in analyzing data. If Ω_4 or Ω_{5c} is accepted, then there is a sound statistical reason to be satisfied with the model chosen because it provides a significant improvement over the models that nest it in Figures 2 and 3. However, if one is led to the saturated models, Ω_{5a} or Ω_{6c}, then one has no direct, statistical way of being confident that the chosen model adequately reflects the underlying cognitive processes. However, Model Ω_7 implies several constraints over the general model that are described in Equations 14 and 15. Thus, one approach would be to "eyeball" these constraints in terms of the data in **T** of Equation 1. As long as there are no obvious violations, we recommend accepting the model and routinely using the sequential decision approach we describe to analyze source-monitoring data. Our recommendation is partly supported because we have applied the approach to many sets of source-monitoring data, and frequently a testable model, such as Ω_4 or Ω_{5c}, has fit the data well. In fact, the next section reports some of these cases. Because the restricted versions often fit well, it seems a reasonable conjecture that the less restricted versions are also viable.

Once a model is selected for use, it is possible to use it to measure the underlying process capacities and test hypotheses about how they compare across experimental groups. These matters are taken up in detail in Riefer and Batchelder (1988), and in the statistical literature, Read and Cressie (1988) is a recent, very readable source.

Empirical Examples

The purpose of this section is to demonstrate that the class of multinomial models just presented does a valid job of measuring cognitive factors in source-monitoring paradigms. To do this, we take two areas within cognitive psychology—reality monitoring and bilingual memory—and show how the multinomial models can be used to shed light on theoretical issues in these areas.

For each of these two areas, we have taken a pair of relevant experiments and have analyzed the experiments using the family of multinomial models in Figure 2. Each of the experiments was chosen because the original article either included a group 3×3 frequency table **T** for the Y_{ij} statistics or presented data in enough detail so that the 3×3 table could be derived. The main focus of each example is to compare the results of the model's analysis with the conclusions of the original investigators, which were based on traditional empirical statistics. If the multinomial model provides a valid measure of source monitoring, then the model's analysis should tend to provide psychologically meaningful interpretations of empirical results and in some cases, we hope, expand on the conclusions reached by the traditional analyses.

Reality Monitoring

Johnson, Foley, and Leach (1988)

A large research area in cognitive psychology deals with people's ability to monitor their own cognitive processes. Perhaps the most well-known example of this is the reality-monitoring paradigm discussed in the introduction, in which people attempt to differentiate between memories of perceived and imagined events. A key element in Johnson and Raye's (1981) theory of reality monitoring is that memories are associated with sensory information and that this type of information is a crucial part of reality-monitoring decisions. In particular, memories produced by imagination are presumed to have fewer sensory characteristics than memories based on actual perceptions, and therefore this information provides an important cue for differentiating between such memories.

Johnson et al. (1988, Experiment 1) explored this issue in a study in which they experimentally manipulated the overlap of sensory characteristics between perceived and imagined memories. They accomplished this by having subjects discriminate between words spoken by another person (Person A) and words that subjects imagined. The imagined words were either in the subject's own voice (S), in the voice of the speaker (A), or the voice of another speaker (B). The main idea behind this manipulation is that words imagined in a subject's own voice have fewer overlapping sensory characteristics with words spoken by another person, and hence source memory should be best under these circumstances. In contrast, source memory should be poorest when subjects both imagine and listen to words in Person A's voice, because the overlap of sensory information should be maximal in this case.

Original analysis. Johnson et al. (1988) measured old–new recognition by determining the number of hits plus the number of correct rejections, divided by the number of test items. It is easy to see that this statistic is a simple extension of Equation 2:

$$\frac{(Y_{11} + Y_{12}) + (Y_{21} + Y_{22}) + Y_{33}}{Y_1 + Y_2 + Y_3}. \tag{24}$$

Johnson et al. (1988) found that there were no significant differences on this measure across the three experimental conditions. Their measure of source memory was the identification-of-origin statistic I from Equation 4. As they hypothesized, imagining in one's own voice produced the best source discrimination, whereas imagining in A's voice produced the worst (with imagining in B's voice in between). After this analysis, Johnson et al. (1988) also examined confusion errors separately for listen and imagine items and found that the significant differences across conditions were due to differences in discriminability of the imagine items but not the listen items.

Model's analysis. The group 3×3 data frequencies from the original Johnson et al. (1988) study are in Table 1, presented separately for the three experimental conditions: (a) listen to A, imagine in subject's voice, or L(a)–I(s); (b) listen to A, imagine in B's voice, or L(a)–I(b); and (c) listen to A, imagine in A's voice, or L(a)–I(a).

The initial step in analyzing data of this type is to determine which version of the multinomial model to use. As outlined in the previous section, this first involves determining whether the

Table 1
Group 3 × 3 Data Tables Constructed From Johnson, Foley, and Leach (1988)

	L(a)–I(s) response			L(a)–I(b) response			L(a)–I(a) response		
Item	L	I	N	L	I	N	L	I	N
Listen	87	8	25	74	16	45	63	13	29
Imagine	14	95	11	23	76	36	46	36	23
New	35	4	201	28	17	225	19	13	178

Note. Data are from "The Consequences for Memory of Imagining in Another Person's Voice" by M. K. Johnson, M. A. Foley, and K. Leach, 1988, *Memory & Cognition, 16,* p. 339. Copyright 1988 by the Psychonomic Society. Adapted by permission. Experimental conditions are as follows: L(a)–I(s) = listen to A, imagine in subject's voice; L(a)–I(b) = listen to A, imagine in B's voice; and L(a)–I(a) = listen to A, imagine in A's voice. L = listen; I = imagine; N = new.

detection rates for the two sources significantly differ (i.e., testing H_0: $p_{13} = p_{23}$ from Equation 23 for the group tables). A statistical test of proportions revealed that the listen items had a significantly higher proportion of misses than the imagine items had across the three conditions, $\chi^2(3) = 8.75$, $p < .05$. This narrows the choice to Model 5c or 6c, because these versions allow D_1 to be different from D_2. To choose between these two remaining models, we performed a goodness-of-fit test on Model 5c. This involves computing the log-likelihood ratio statistic G^2 described earlier. The test revealed that Model 5c did not fit the data well across the three experimental conditions, $G^2(3) = 9.69$, $p < .05$, so we decided to analyze the data using Model 6c. Table 2 presents the parameter estimates from this version of the model. Recognition memory and source memory for the listen items are measured by D_1 and d_1, respectively, and recognition and source memory for the imagine items are represented by D_2 and d_2.

To see whether the model's analysis matches the empirical results of Johnson et al. (1988), we performed likelihood ratio tests on the estimates in Table 2. Like the goodness-of-fit test, these tests involve computing a log-likelihood ratio statistic G^2 (see Riefer & Batchelder, 1988). For recognition memory, the results revealed that recognition of the listen items (D_1) did not significantly differ across conditions, $G^2(2) = 4.98$, $p > .05$. This is not surprising, because these items always involved listening to the same source (Person A) and hence were the same across the three conditions. However, there was a significant effect of conditions on the recognition of the imagine items (D_2), $G^2(2) = 13.92$, $p < .01$, as well as a recognition advantage for imagine items over listen items ($D_1 > D_2$) across the three conditions, $G^2(3) = 8.90$, $p < .05$. Further analysis revealed that both of these significant effects were due to the high recognition of items that subjects imagined in their own voices ($D_2 = .89$).

Concerning source memory, the likelihood ratio test showed that source discrimination for the listen items (d_1) did not differ significantly across conditions, despite the low value of $d_1 = .19$ for the L(a)–I(s) condition, $G^2(2) = 1.16$, $p > .05$. However, source memory for the imagine items (d_2) significantly decreased across conditions, $G^2(2) = 47.52$, $p < .001$. Thus, the effect of the experimental conditions on source memory tends to match that for recognition memory, in that significant differences were found for the imagine items across conditions, but not for the listen items.

Summary. In general, the results of the model seem to do a good job of replicating Johnson et al.'s (1988) empirical results. For example, the model's analysis of source memory, using parameters d_1 and d_2, exactly matches the pattern of results found by Johnson et al. (1988) using confusion errors. Specifically, source discrimination for the imagine items (d_2) steadily decreased as their overlap with the listen items increased, which provides support for Johnson and Raye's (1981) theory. Unfortunately, it is difficult to make a similar comparison on recognition memory, because we examined the detection of listen and imagine items separately using parameters D_1 and D_2, whereas Johnson et al. (1988) used an empirical measure that combined the two types of items. However, the finding of a significant advantage of imagine items over listen items is quite consistent with previous studies that show self-generated information to be more memorable than external information (e.g., Raye & Johnson, 1980; Voss et al., 1987).

Harvey (1985)

Of the many applications of reality monitoring, one of the most important has been in the area of clinical psychology. Johnson (1988) has advocated that the thought disorders associated with schizophrenia and other mental dysfunctions may be a result of reality-monitoring failures. For example, schizophrenic people may have trouble discriminating between their own thoughts and information from external sources, which helps explain the lack of continuity in their speech and behavior.

This hypothesis was tested in a study by Harvey (1985), who compared manic and schizophrenic subjects in standard reality-monitoring tasks. The study included both thought-disordered (TD) and non-thought-disordered (NTD) patients of each group, as well as a normal control group. Two reality-monitoring subtasks were examined, one of which was a say–think task. In this task, subjects were instructed either to say written words out loud or to think of the words to themselves. The other

Table 2
Parameter Estimates for Johnson, Foley, and Leach's (1988) Experiment

Condition	D_1	D_2	d_1	d_2	b	g
L(a)–I(s)	.75	.89	.19	.87	.16	.90
L(a)–I(b)	.60	.68	.59	.68	.17	.62
L(a)–I(a)	.67	.74	.62	.06	.15	.59

Note. D_1 = detectability of the listen items; D_2 = detectability of the imagine items; d_1 = source discriminability for the listen items; d_2 = source discriminability for the imagine items; b = bias for responding "old"; g = guessing that the item was a listen item; L(a)–I(s) = listen to A, imagine in subject's voice; L(a)–I(b) = listen to A, imagine in B's voice; L(a)–I(a) = listen to A, imagine in A's voice.

Table 3
Group 3 × 3 Data Tables Constructed From Harvey (1985)

	Manic subjects						Schizophrenic subjects						Normal subjects		
	NTD			TD			NTD			TD					
Item	S	T	N	S	T	N	S	T	N	S	T	N	S	T	N
Say	22	27	31	43	6	31	13	21	46	44	10	26	23	22	35
Think	7	54	19	20	15	45	4	42	34	32	8	40	9	45	26
New	4	26	50	5	9	66	6	20	54	24	7	49	7	10	63

Note. NTD = non-thought disordered; TD = thought disordered; responses are as follows: S = say; T = think; N = new.

task involved listening to words from two speakers (listen–listen). For the purpose of the current analysis, only data from the say–think task are discussed, because not enough information was provided in the original Harvey article to reproduce a group 3 × 3 table for the listen–listen task.

Original analysis. Harvey's (1985) measure of recognition consisted of correct detections of old items, scored without regard to which source the subject attributed the item. This is basically the same measure of hits as in Equation 2, except that in this study, hits were computed separately for the say and think items:

$$H_i = \frac{Y_{i1} + Y_{i2}}{Y_i}, \qquad (25)$$

for $i = 1, 2$. The major empirical result was a significant interaction between thought disorder (NTD vs. TD) and source (say vs. think). Specifically, TD patients recognized say items better than think items, but NTD patients recognized think items better than say items.

Harvey's (1985) measure of source memory consisted of identification-of-origin scores, also computed separately for the say and think items, which is the measure presented in Equation 5. On this measure, TD manic subjects were not significantly different from NTD manic subjects, but the results for TD schizophrenics were significantly poorer than were the results for NTD schizophrenics. Finally, NTD patients did not significantly differ from the normal subjects on either recognition or source memory.

Model's analysis. Enough information was given in the original Harvey (1985) article to construct group 3 × 3 tables for the say–think task, and these are presented in Table 3 for each of the five groups. A test of equality of proportions showed that there were significant differences across the groups between the detection rates for the say and think items, $\chi^2(5) = 19.90$, $p < .01$. Accordingly, Model 5c, which assumes $D_1 \neq D_2$, was used to analyze the data. We chose Model 5c over Model 6c because a likelihood ratio test comparing these two nested models showed that Model 6c did not significantly improve the fit of Model 5c across the five groups, $G^2(5) = 9.80$, $p > .05$. Table 4 presents the goodness-of-fit tests of Model 5c to the data, as well as the resultant parameter estimates. Parameters D_1 and D_2 represent detection of the say and think items, respectively.

Paralleling Harvey's (1985) findings, the results for TD subjects were significantly higher on item detectability for the say items (D_1), $G^2(2) = 7.90$, $p < .05$, whereas those for NTD subjects were superior for the think items (D_2), $G^2(2) = 8.17$, $p < .05$. The control subjects performed similarly to the NTD patients, with no significant differences between these groups for any of the model's parameters.

For the source memory parameter d, an examination of Table 4 shows that NTD manic subjects were not significantly different than TD manic subjects ($d = .51$ vs. .43, respectively), $G^2(1) = 0.11$, $p > .05$. On the other hand, the difference for the schizophrenic patients was more clearcut. The NTD schizophrenic subjects' results were much better than those of TD schizophrenics on source memory ($d = .87$ vs. .03, respectively). In fact, the d value of .03 for the TD schizophrenics indicates that their performance was nearly at chance level. Interestingly, even though the parameter difference between NTD and TD schizophrenic subjects was quite large, it was only marginally significant, $G^2(1) = 3.07$, $p < .05$ (one-tailed test). This result may reflect a lack of power in the original sample size used by Harvey (1985), who tested only 10 patients in each condition.

Summary. Despite the lack of statistical significance for this last test, the overall pattern of results from the model's analysis almost precisely matches Harvey's (1985) findings based on the empirical statistics. The model's parameters for recognition memory exhibit the same crossover interaction as found in Harvey's analysis. The pattern for the source memory

Table 4
Parameter Estimates and Goodness-of-Fit Tests for Harvey's (1985) Experiment

Group	Parameter estimate					Goodness-of-fit
	D_1	D_2	d	b	g	$G^2(1)$
Manic NTD	.39	.62	.51	.37	.17	0.50
Manic TD	.53	.29	.43	.18	.69	9.94*
Schizophrenic NTD	.11	.36	.87	.34	.21	0.25
Schizophrenic TD	.47	.18	.03	.39	.80	0.18
Normal	.44	.59	.42	.21	.30	1.20

Note. D_1 = detectability of say items, D_2 = detectability of listen items, d = source discriminability; b = bias for responding "old"; g = guessing that the item is a say item; TD = thought disordered; NTD = non-thought disordered.
* $p < .01$.

parameter matches Harvey's results as well, with no systematic differences between NTD and TD manic subjects, and large differences between NTD and TD schizophrenic subjects. This latter result provides support for Johnson's (1988) contention that schizophrenic people have trouble discriminating between items that are experienced externally (listen items) and items that are generated internally (say items).

Bilingual Memory

A key question in the area of bilingual memory concerns how a bilingual person's two languages are represented in cognition. Basically, two opposing theories have been advanced. The dual-code or independence hypothesis (Kolers, 1963) asserts that there are two distinct and separate lexicons for each language. In contrast, the single-code or interdependence hypothesis (McCormack, 1976, 1977) postulates that only one, integrated system exists for both languages. This theory assumes that lexical information is stored in a single, abstract representation that exists on a supralinguistic level, unrelated to the language in which the information occurred.

Many experiments using different cognitive tasks have been conducted on this issue (see Durgunoglu & Roediger, 1987, and Gerard & Scarborough, 1989, for recent reviews). Of particular relevance here, however, are several experiments that have examined people's source memory for bilingual information. In these studies, subjects are presented with information in a mixture of both languages in a source-monitoring task and are then required to remember in which language the information occurred. For the most part, these studies show that source memory for language is highly accurate. Several theorists (e.g., Gerard & Scarborough, 1989; Kolers & Gonzalez, 1980) cite this as support for the dual-code theory, because of its assumption that words are represented in memory directly in the language in which they were stored, which should theoretically facilitate source memory.

However, proponents of the single-code theory assert that this result can be easily explained by a unified lexical system. These theorists assert that the input language is a distinguishing attribute of words, one that can be attached to the abstract code in the form of a tag (see McCormack, 1976). This tag would be comparable with tags that might exist for other nonsemantic attributes of words, such as modality of input, speaker's voice, print style, and so on. As stated in the introduction, prior research has shown that source memory is often very accurate for this type of ancillary information. Thus, if one assumes that people are also very good at using these tags to remember the language of input, then single-code theories are quite capable of predicting the high accuracy of source memory for language.

Saegert, Hamayan, and Ahmar (1975)

If one accepts the idea that source memory for language is based on tags, then the question still remains as to why this memory tends to be so accurate. One possible explanation has been proposed by Saegert et al. (1975). According to them, source memory for language is dependent on the nature of the memory task itself. Most experiments on source memory for language involve a mixed list of unrelated words, with no meaning or semantic structure relating the stimuli. Such an artificial situation diverts attention away from semantic aspects of the stimuli and focuses it more on nonsemantic attributes, such as the language of input. Saegert et al. reasoned that if the memory task required more semantic processing, subjects would have to encode the stimuli on a higher cognitive level, at the expense of the tags.

To test this idea, Saegert et al. (1975, Experiment 2) had trilingual subjects memorize a mixed list of words in both French and English. One group of subjects saw the words presented alone (word group), and another received the words in the context of sentences (sentence group). The idea behind the study was that sentence presentation demands more semantic encoding of the stimuli, which should interfere with the formation of the language tags and hence impede the source memory for language.

Original analysis. Saegert et al.'s (1975) measure of recognition memory was Equation 25, the proportion of items correctly recognized, computed separately for the French and English words. An analysis of variance revealed that recognition was significantly poorer in the sentence group compared with the word group. For their measure of source memory, Saegert et al. used Equation 5, which for this study is the conditional probability of a correct language judgment given correct recognition. Similar to overall recognition memory, this measure of source memory was also significantly poorer in the sentence condition.

Saegert et al. (1975) interpreted these results as being inconsistent with dual-code theories and thus supportive of single-code theories. According to their reasoning, dual-code theories predict that words are stored in memory specifically in the language in which they were encoded, and hence language memory for correctly recognized words should be nearly perfect. The fact that subjects exhibited some forgetting of source memory in the sentence group is argued, then, to be inconsistent with dual-code theories.

However, there is a potential problem with Saegert et al.'s (1975) interpretation of these results. As we stated earlier, traditional measures of source memory, such as the one used by Saegert et al., may reflect the combined effect of different cognitive processes. For example, measures of source memory that are conditional on correct recognition are consequently dependent on the overall level of recognition in an experiment, and these measures may be difficult to interpret when overall recognition is different across experimental conditions. This is precisely what happened in the Saegert et al. study, with recognition and source memory both being poorer in the sentence group. Thus, the change in source memory across conditions may be caused solely by the difference in recognition memory and not by the experimental manipulation of sentences and words. Of course, if this were true, it would force a very different theoretical interpretation of the results than the one proposed by Saegert et al.

Model's analysis. Fortunately, the multinomial model for source monitoring is capable of helping to settle this issue by providing more direct measures of recognition and source memory separately. The group 3×3 data frequencies derived from Experiment 2 in the Saegert et al. (1975) article are pre-

Table 5
Group 3 × 3 Data Tables Constructed From Saegert, Hamayan, and Ahmar (1975)

	Sentence group response			Word group response		
Item	E	F	N	E	F	N
English	184	75	173	152	19	45
French	77	187	168	21	143	52
New	58	75	155	26	19	99

Note. E = English; F = French; N = new.

sented in Table 5 for the word and sentence groups. (There were actually two sentence groups in the study, which we combined into one group for this analysis.)

To determine which version of the model to use to analyze the data, we again started by comparing the detection rates for the French and English items. A statistical test failed to reject H_0: $p_{13} = p_{23}$, $\chi^2(2) = 0.77$, indicating no significant difference in the detection rates. This narrows the choice of models to Model 4 or 5b, because these versions constrain the detection rates to be equal ($D_1 = D_2$). A goodness-of-fit test showed that Model 4 fit the data well, $G^2(2) = 2.37$, $p > .05$, and that Model 5b did not significantly improve on this fit, $G^2(2) = 1.36$, $p > .05$. Thus, Model 4 was used to estimate the parameters. Table 6 presents the goodness-of-fit tests for this model, along with the parameter estimates.

Recognition memory (as measured by D) is significantly poorer for the sentence group compared with the word group, $G^2(1) = 50.41$, $p < .001$ (Table 6). This result matches Saegert et al.'s (1975) conclusion using empirical statistics. However, when source memory (as measured by d) was examined, a very different result was obtained. Specifically, no significant difference between the two groups is evident, $G^2(1) = 0.17$, *ns*. In fact, the estimate of d, although high in both groups, is even slightly higher in the sentence group.

Summary. The multinomial model's interpretation of the data is very different from Saegert et al.'s (1975) interpretation based on empirical statistics. Although Saegert et al. concluded that source memory is inhibited by the semantic context of the sentences, the model reveals that this is actually not the case. Instead, the model shows that source memory is equally high for both groups and that the only effect of the sentences is on overall recognition memory and not on source memory. If Saegert et al.'s earlier reasoning is still valid—that is, that dual-code theories predict high levels of language memory that should be unaffected by context—then the same reasoning would lead one to conclude that their experimental results are actually in support of the dual-code theory, contrary to what Saegert et al. originally concluded.

Rose, Rose, King, and Perez (1975)

In the previous example, the multinomial model's analysis of Saegert et al.'s (1975) data presents a very different picture of the effects of source memory than do the standard empirical statistics. Because of this discrepancy between the modeling approach and the empirical approach, it would be desirable to determine whether this type of result could be replicated in another, independent experiment. Fortunately, a related experiment suitable for this purpose was conducted by Rose et al. (1975, Experiment 1). They wanted to determine whether the phenomenon of accurate source memory for language could be found at more complex cognitive levels. To do this, they presented subjects with a mixed list of English and Spanish sentences (as opposed to words). In one condition, the sentences in each language all related to a common topic, whereas in the other condition, all sentences were unrelated. Rose et al. reasoned that if subjects associate language with a certain type of information, then related sentences will be less discriminable and will therefore make it more difficult to associate the correct language to any particular sentence.

Original analysis. Rose et al. (1975) measured sentence retention using a forced-choice recognition task. Specifically, subjects were tested on original sentences and distracters, and for each, they were instructed to circle one of four response alternatives: (a) a sentence that had been presented, (b) its translation, (c) a distracter, and (d) the distracter's translation. (More detail can be found in the original Rose et al. article.) Sentences were counted as correctly recognized if they had the correct meaning, regardless of the language they were in. Based on this measure, which is essentially the same as in Equation 25, recognition was significantly poorer for the related sentences than for unrelated sentences.

For source memory, Rose et al. (1975) computed Equation 5, the probability of choosing the correct language of the sentence, given that the meaning was correctly recognized. From this, they also computed a standard correction for guessing that consisted of Pr(hits) − Pr(false alarms). These measures exhibited small and nonsignificant differences between conditions; however, the differences were in the hypothesized direction, with source memory poorer for the related sentences.

As can be seen, the same empirical pattern of results occurred in this experiment as was witnessed in the Saegert et al. (1975) study. Source memory was poorer for the related sentences, but this was paralleled by poorer recognition memory for the related sentences as well. Thus, it is again possible that the effect of conditions on source memory, albeit small, was due exclusively to differences in the overall recognition rate.

Model's analysis. Using information from the original Rose et al. (1975) article, we attempted to construct the 3 × 3 tables for their Experiment 1 (see Table 7). This was difficult because of sketchy information given by Rose et al. concerning the false-

Table 6
Parameter Estimates and Goodness-of-Fit Tests for Saegert, Hamayan, and Ahmar's (1975) Experiment

	Parameter estimate				Goodness-of-fit
Condition	D	d	b	g	$G^2(2)$
Sentence group	.27	.95	.46	.48	1.55
Word group	.67	.88	.31	.56	0.82

Note. D = item detectability; d = source discriminability; b = bias for responding "old"; g = guessing that the item was in English.

Table 7
Group 3 × 3 Data Tables Constructed From Rose, Rose, King, and Perez (1975)

	Related sentences response			Unrelated sentences response		
Item	E	S	N	E	S	N
English	164	46	30	181	39	20
Spanish	46	158	36	47	173	20
New	111	107	262	102	85	293

Note. E = English; S = Spanish; N = new.

alarm rates in their experiment. It is possible that our frequencies for the distracter items in Table 7 may be off by a small degree compared with the true frequencies. In any event, this should not greatly change the overall outcome of the model's analysis.

A statistical test of equality of proportions revealed no significant difference between the detection rates for the Spanish and English items, $\chi^2(2) = 0.63$, *ns.* Accordingly, the data were analyzed with Model 4, which assumes $D_1 = D_2$ and which fit the data very well. The resultant parameter estimates derived from this version of the model are in Table 8, along with the results of the goodness-of-fit tests of the model.

Recognition memory is significantly poorer for the related sentences ($D = .75$) than for the unrelated sentences ($D = .86$), $G^2(1) = 9.97$, $p < .01$ (Table 8). However, there is no significant difference for the source-monitoring parameter d, $G^2(1) = 0.01$, *ns.* In fact, the estimate of d is nearly the same in each condition ($d = .64$ vs. .65 for related and unrelated sentences, respectively).

Summary. Once again, the multinomial model reveals that altering semantic context affects the recognition memory of bilingual information but has no effect on source memory for language. This finding was true for both Saegert et al.'s (1975) experiment, in which context was based on words rather than sentences, and Rose et al.'s (1975) study, in which context was based on related versus unrelated sentences.

On a theoretical level, we leave it up to the proponents of the independence and interdependence hypotheses to account for these results by their respective theories. However, it seems to us that the pattern of results revealed by the multinomial model is consistent with recent multiple-code or multilevel theories of bilingual memory (e.g., Durgunoglu & Roediger, 1987; Frenck & Pynte, 1987; Gerard & Scarborough, 1989). These theories state that bilingual information can be encoded in many different forms and at different levels of processing and that the actual pattern of results found in bilingual research depends on the specific task used to measure memory. Frenck and Pynte have pointed out, for example, that results showing independence of the language systems (and thus supporting the dual-code hypothesis) may in fact reflect independence at low levels of processing. It seems reasonable to speculate that memory for language of input may indeed by stored at a low level, along with memory for other nonsemantic attributes of information, such as modality, print style, speaker's voice, and so on. If this is so, then the semantic context of information, which can be assumed to influence higher cognitive levels of processing, may affect overall recognition memory but should not affect memory for language of input. This, of course, is precisely the pattern of results revealed by the multinomial model.

Conclusion

In this article, we have developed and evaluated a class of multinomial models for source monitoring. These models are capable of taking data from source-monitoring paradigms and separately measuring the cognitive capacities that underlie such data. In addition, the hypothesis-testing procedures outlined earlier provide a way of determining whether experimental variables significantly influence these processes.

Perhaps the most important feature of the models is their ability to take source-monitoring data and compute separate measures for old-item detection and source memory. We have argued that models such as the ones presented herein have certain advantages over traditional methods for measuring source memory because empirical statistics are often confounded by different cognitive processes and thus are not a pure measure of any specific one. This point was illustrated with the empirical examples on bilingual memory. Traditional statistical measures of source memory, such as those used by Saegert et al. (1975) and Rose et al. (1975), involve some measure of correct source identification, conditional on the correct detection of items. The problem with such measures is that they are potentially confounded with the overall detectability of the items, as shown precisely in Equations 21 and 22. Thus, the effect of experimental variables on empirical measures of source memory may actually reflect changes in the detectability of items, and not source memory at all. As we found, the model in fact revealed that the differences in source memory in the Saegert et al. and Rose et al. experiments were actually due to differences in detection (D) and not in source memory (d).

Because of the problems with empirical statistics, we recommend to anyone conducting research involving source monitoring that they use parametric multinomial models to analyze data, at least as a supplement to traditional empirical measures. To assist researchers who may wish to use the multinomial models presented in this article, we have available computer programs that perform the parameter estimation, goodness-of-fit, and hypothesis-testing procedures outlined earlier. These programs can be run on an IBM-compatible personal computer and can be obtained by writing to William H. Batchelder.

Table 8
Parameter Estimates and Goodness-of-Fit Tests for Rose, Rose, King, and Perez's (1975) Experiment

	Parameter estimate				Goodness-of-fit
Condition	D	d	b	g	$G^2(2)$
Related sentences	.75	.64	.45	.51	0.64
Unrelated sentences	.86	.65	.39	.55	0.0003

Note. D = item detectability, d = source discriminability, b = bias for responding "old"; g = guessing that the item was in English.

Further work will probably need to be conducted to address certain statistical issues related to our models. For example, one issue concerns the sample size needed to achieve a desired level of power in testing hypotheses about a model's parameters. This point came up in the analysis of Harvey's (1985) experiment, in which small sample sizes prevented some very large differences in parameter values from reaching statistical significance. Another issue, related to sample size, involves statistical inference procedures for these models. As Riefer and Batchelder (1988) have pointed out, parameter estimates and statistical tests for multinomial models are only approximate because they are based on the assumption that the sample size is sufficiently large. As a consequence, small sample sizes not only may lead to reduced power in hypothesis testing but also may introduce systematic bias in the parameter estimators or inflation of Type I error in the goodness-of-fit tests. Of course, these sorts of statistical issues are also important for more traditional statistical models like analysis of variance or log-linear models.

Still another issue concerns the effect of individual differences in the parameter values. The statistical techniques for parameter estimation, goodness of fit, and hypothesis testing are all based on the assumption that data observations are identically distributed, which is violated when there are individual differences in the model's parameters. For example, the four empirical examples in the previous section all involved analysis of *group* 3×3 data tables, that is, data that had been combined across individual subjects. This was necessitated by the fact that only group data were available in the original articles; however, such an analysis is insensitive to possible individual differences between subjects.

One possible method for addressing these issues is to conduct Monte Carlo simulations of the models. Under these simulations, one can systematically vary sample size and introduce individual differences in the parameter values. The results of the computer simulations can then be used to determine the sample size needed for various levels of power, as well as the effect of small sample size and individual differences on parameter estimation and goodness of fit. Riefer and Batchelder (1990) provided a detailed description of how these simulations can be conducted for multinomial models and demonstrated these techniques by conducting an extensive Monte Carlo analysis of their multinomial model for storage and retrieval (Batchelder & Riefer, 1986).

In general, conducting these types of simulations is very straightforward, and some future work in this area for multinomial models of source monitoring may prove beneficial. Our overall hope is that researchers will be able to successfully use the multinomial models for source monitoring to explore theoretical issues in many different areas of cognitive psychology.

References

Anderson, R. E. (1984). Did I do it or did I only imagine doing it? *Journal of Experimental Psychology: General, 113,* 594–613.

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

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika, 53,* 71–92.

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

Bray, N. W., & Batchelder, W. H. (1972). Effects of instructions and retention interval on memory of presentation mode. *Journal of Verbal Learning and Verbal Behavior, 11,* 367–374.

Chechile, R., & Meyer, D. L. (1976). A Bayesian procedure for separately estimating storage and retrieval components of forgetting. *Journal of Mathematical Psychology, 13,* 269–295.

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology, 26,* 274–284.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society Series B,* 1–38.

Durgunoglu, A. Y., & Roediger, H. L., III (1987). Test differences in accessing bilingual memory. *Journal of Memory and Language, 26,* 377–391.

Durso, F. T., & Johnson, M. K. (1980). The effects of orienting tasks on recognition, recall, and modality confusion of pictures and words. *Journal of Verbal Learning and Verbal Behavior, 19,* 416–429.

Durso, F. T., Reardon, R., & Jolly, E. J. (1985). Self-nonself-segregation and reality monitoring. *Journal of Personality and Social Psychology, 48,* 447–455.

Eich, E., & Metcalfe, J. (1989). Mood dependent memory for internal versus external events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 443–455.

Finke, R. A., Johnson, M. K., & Shyi, G. C. (1988). Memory confusions for real and imagined completions of symmetrical visual patterns. *Memory & Cognition, 16,* 133–137.

Foley, M. A., & Johnson, M. K. (1985). Confusions between memories for performed and imagined actions: A developmental comparison. *Child Development, 56,* 1145–1155.

Foley, M. A., Johnson, M. K., & Raye, C. L. (1983). Age-related changes in confusion between memories for thoughts and memories for speech. *Child Development, 54,* 51–60.

Frenck, C., & Pynte, J. (1987). Semantic representation and surface forms: A look at across-language priming in bilinguals. *Journal of Psycholinguistic Research, 16,* 383–396.

Gerard, L. D., & Scarborough, D. L. (1989). Language-specific lexical access of homographs by bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 305–315.

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

Greeno, J. G., & Steiner, T. E. (1964). Markovian processes with identifiable states: General considerations and application to all-or-none learning. *Psychometrika, 29,* 309–333.

Harvey, P. D. (1985). Reality monitoring in mania and schizophrenia. *The Journal of Nervous and Mental Disease, 173,* 67–72.

Hashtroudi, S., Johnson, M. K., & Chrosniak, L. D. (1989). Aging and source monitoring. *Psychology and Aging, 4,* 106–112.

Hintzman, D. L., Block, R. A., & Inskeep, N. R. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior, 11,* 741–749.

Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Source memory impairment in patients with frontal lobe lesions. *Neuropsychologia, 27,* 1043–1056.

Johnson, M. K. (1988). Discriminating the origin of information. In T. F. Ohmanns & B. A. Maher (Eds.), *Delusional beliefs: Interdisciplinary perspectives* (pp. 34–65). New York: Wiley.

Johnson, M. K., Foley, M. A., & Leach, K. (1988). The consequences for memory of imagining in another person's voice. *Memory & Cognition, 16,* 337–342.

Johnson, M. K., Kahan, T. L., & Raye, C. L. (1984). Dreams and reality monitoring. *Journal of Experimental Psychology: General, 113,* 329–344.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88,* 67–85.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal of Psychology, 94,* 37–64.

Johnson, M. K., Raye, C. L., Foley, M. A., & Kim, J. K. (1982). Pictures and images: Spatial and temporal information compared. *Bulletin of the Psychonomic Society, 19,* 23–26.

Klatzky, R. L. (1975). *Human memory.* San Francisco: Freeman.

Kolers, P. A. (1963). Interlingual word association. *Journal of Verbal Learning and Verbal Behavior, 2,* 291–300.

Kolers, P. A., & Gonzalez, E. (1980). Memory for words, synonyms, and translations. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 53–65.

Laming, D. (1973). *Mathematical psychology.* New York: Academic Press.

Lehmann, E. L. (1983). *Theory of point estimation.* New York: Wiley.

Light, L. L., & Berger, D. E. (1976). Are there long-term "literal copies" of visually presented words? *Journal of Experimental Psychology: Human Learning and Memory, 2,* 654–662.

Light, L. L., Stansbury, C., & Rubin, C. (1973). Memory for modality of presentation: Within-modality discrimination. *Memory & Cognition, 1,* 395–400.

Lindsay, S. D., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17,* 349–358.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

McCormack, P. D. (1976). Language as an attribute of memory. *Canadian Journal of Psychology, 30,* 238–248.

McCormack, P. D. (1977). Bilingual linguistic memory: The independence–interdependence issue revisited. In P. A. Hornby (Ed.), *Bilingualism: Social and educational implication* (pp. 57–66). New York: Academic Press.

McIntyre, J. S., & Craik, F. I. M. (1987). Age differences in memory for item and source information. *Canadian Journal of Psychology, 41,* 175–192.

Mitchell, D. B., Hunt, R. R., & Schmitt, F. A. (1986). The generation effect and reality monitoring: Evidence from dementia and normal aging. *Journal of Gerontology, 41,* 79–84.

Rabinowitz, J. C. (1990). Effects of repetition of mental operations on memory for occurrence of origin. *Memory & Cognition, 18,* 72–82.

Raye, C. L., & Johnson, M. K. (1980). Reality monitoring vs. discriminating between external sources of memories. *Bulletin of the Psychonomic Society, 15,* 405–408.

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

Reardon, R., Durso, F. T., Foley, M. A., & McGahan, J. R. (1987). Expertise and the generation effect. *Social Cognition, 5,* 336–348.

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

Riefer, D. M., & Batchelder, W. H. (1990). *Statistical inference for multinomial processing tree models* (Tech. Rep. No. MBS 90-05). Irvine: University of California, Mathematical Behavioral Sciences.

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

Saegert, J., Hamayan, E., & Ahmar, H. (1975). Memory for language of input in polyglots. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 607–613.

Shimamura, A. P., & Squire, L. R. (1987). A neuropsychological study of fact memory and source amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 464–473.

Voss, J. F., Vesonder, G. T., Post, T. A., & Ney, L. G. (1987). Was the item recalled and if so, by whom? *Journal of Memory and Language, 26,* 466–479.

Appendix

Model Constraints

We suppose that the underlying probabilities p_{ij} in **P** of the general product-multinomial model in Equations 8 and 9 are restricted to satisfy the equations $p_{ij}(\theta) = p_{ij}$, for $i, j = 1, 2, 3$. In this expression, θ is the vector of parameters in the parameter space of some model Ω_x in Figure 3. First, note that all models have the parameters g and b, which are given in terms of p_{ij} from Equation 12 by

$$b = p_{31} + p_{32} \tag{A1}$$

and

$$g = \frac{p_{31}}{p_{31} + p_{32}}. \tag{A2}$$

Next, assume Model Ω_7, and note from Equations 10c and 11c that the detection parameters must satisfy the equations

$$D_1 = \frac{p_{33} - p_{13}}{p_{33}} \tag{A3}$$

and

$$D_2 = \frac{p_{33} - p_{23}}{p_{33}}. \tag{A4}$$

Finally, from Equations 10b and 11b, it is easy to use Equations A1 through A4 to derive the relationships

$$(1 - d_1)(1 - a) = \frac{p_{33}p_{12} - p_{13}p_{32}}{p_{33} - p_{13}} \equiv A_1 \tag{A5}$$

and

$$(1 - d_2)a = \frac{p_{33}p_{21} - p_{23}p_{31}}{p_{33} - p_{23}} \equiv A_2. \tag{A6}$$

Because $0 \le A_1, A_2 \le 1$, Equations A5 and A6 imply the inequality constraints

$$p_{33} - p_{i3} \ge p_{33}p_{ij} - p_{i3}p_{3j} \ge 0, \tag{A7}$$

for $i, j = 1, 2$, and $i \ne j$. If Equation A7 is satisfied, it follows that $0 \le D_i \le 1$ in Equations A3 and A4, as required. Furthermore, algebraic manipulations show that Equation A7 holds if and only if the inequality restrictions in Equation 14 hold. Equation 15 follows from the fact that $A_1 + A_2 \le 1$.

If Model Ω_{6b} is assumed, Equations A1 through A4 remain in effect, and Equations A5 and A6 imply

$$d = 1 - (A_1 + A_2) \tag{A8}$$

and

$$a = \frac{A_2}{A_1 + A_2}. \tag{A9}$$

Because Equations A8 and A9 impose no additional constraints on p_{ij} in **P**, we have $\Omega_7^* = \Omega_{6b}^*$.

Next, assume Model Ω_{6c} holds; that is, $a = g$. In this case, Equations A5 and A6 impose additional constraints, because g is determined from Equation A2. These restrictions are that

$$0 \le A_1/(1 - g), \qquad A_2/g \le 1,$$

and they result in the constraints

$$0 \le \frac{(p_{33}p_{ij} - p_{i3}p_{3j})(p_{3i} + p_{3j})}{(p_{33} - p_{i3})p_{3j}} \le 1, \tag{A10}$$

for $i, j = 1, 2$, and $i \ne j$. Equation 16 in the main text follows from Equation A10 by algebraic manipulation.

Finally, assume Model Ω_{5c}. In this case $d_1 = d_2 = d$ and $a = g$, so Equations A5 and A6 require $(1 - d)(1 - g) = A_1$ and $(1 - d)g = A_2$. Because g is determined by Equation A2, d is overdetermined, and this leads to the dimension-reducing equality restriction in Equation 17.

In the preceding paragraphs, we have discussed necessary and sufficient conditions on p_{ij} in **P** that characterize the four models in the right panel of Figure 3. To derive Equation 19, note that the numerical assignment $p_{11} = p_{13} = p_{21} = p_{23} = 0$, $p_{12} = p_{21} = 1$, $p_{31} = p_{32} = 1/4$, and $p_{33} = 1/2$ satisfies the constraints on Ω_{6b} but fails to satisfy the constraints on Ω_{6c}, so $\Omega_{6c}^* \subset \Omega_{6b}^*$.

All the models in the left panel of Figure 3 require equality of detection rates, that is, $D_1 = D_2$, and Equations A3 and A4 show that this leads to the dimension-reducing constraint given in Equation 18. The rest of the analysis of Models Ω_{6a}, Ω_{5a}, Ω_{5b}, and Ω_4 exactly parallels the analysis of their counterparts, Ω_7, Ω_{6b}, Ω_{6c}, and Ω_{5c}, respectively. Of course, some of the constraints may be simplified by adding the constraint $p_{13} = p_{23}$.

Parameter Estimation

In this article, as well as in Riefer and Batchelder (1988), we have used the maximum likelihood approach to parameter estimation and hypothesis testing. To obtain maximum likelihood estimators (MLEs) for any of the identifiable models Ω_x in Figure 3 (all but Ω_7 and Ω_{6a}), one first computes the likelihood function, $L(\mathbf{T}, \theta)$, where **T** is the matrix of obtained data frequencies in Equation 1 and $\theta \in \Omega_x$ is the appropriate parameter vector. The result is

$$L(\mathbf{T}, \theta) = \prod_{i=1}^{3} Y_i! \prod_{j=1}^{3} [p_{ij}(\theta)^{Y_{ij}}]/Y_{ij}!, \tag{A11}$$

where the $p_{ij}(\theta)$ are given by Equations 10 through 12 with the simplifications in parameters dictated by particular model Ω_x. Next, one must find the numerical vector $\theta \in \Omega_x$ that maximizes Equation A11 for fixed **T**. There are many gradient search techniques for maximizing Equation A11 mentioned in Riefer and Batchelder (1988), and the computer program we offer to readers in the Conclusion section is based on the EM (expectation maximization)-algorithm discussed in Dempster, Laird, and Rubin (1977). Also, Riefer and Batchelder (1988) discussed methods for obtaining confidence intervals for parameters.

In the special case of Ω_{6c}, if the data satisfy the constraints of Equations 14 through 16, then Equations A1 through A4 yield MLEs $\hat{g}$, $\hat{b}$, $\hat{D}_1$, and $\hat{D}_2$, respectively, providing the data proportions $P_{ij} = Y_{ij}/Y_i$

replace the underlying probabilities p_{ij}. Also, from Equations A5 and A6, we see that

$$\hat{d}_1 = \frac{1 - \hat{g} - \hat{A}_1}{1 - \hat{g}} \tag{A12}$$

and

$$\hat{d}_2 = \frac{\hat{g} - \hat{A}_2}{\hat{g}}, \tag{A13}$$

where $\hat{A}_1$ and $\hat{A}_2$ result by replacing the probabilities p_{ij} by the empirical proportions P_{ij} as before.

If the data do not satisfy the constraints of Ω_{6c} and one still wants to obtain MLEs, the iterative search methods mentioned earlier are required.

The hypotheses reported in the Empirical Examples section involve likelihood ratio tests based on the G^2 statistic described in Riefer and Batchelder (1988). Maximum likelihood parameter estimates must be obtained under both the null and alternative hypotheses, and this involves iterative search for the models reported in this article. These are incorporated into the computer program offered to readers in the Conclusion section.

Received October 23, 1989
Revision received March 21, 1990
Accepted April 2, 1990 ■

Butcher, Geen, Hulse, and Salthouse Appointed New Editors, 1992–1997

The Publications and Communications Board of the American Psychological Association announces the appointments of James N. Butcher, University of Minnesota; Russell G. Geen, University of Missouri; Stewart H. Hulse, Johns Hopkins University; and Timothy Salthouse, Georgia Institute of Technology as editors of *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, the Personality Processes and Individual Differences section of the *Journal of Personality and Social Psychology*, the *Journal of Experimental Psychology: Animal Behavior Processes*, and *Psychology and Aging*, respectively. As of January 1, 1991, manuscripts should be directed as follows:

- For *Psychological Assessment* send manuscripts to James N. Butcher, Department of Psychology, Elliott Hall, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455.
- For *JPSP: Personality* send manuscripts to Russell G. Geen, Department of Psychology, University of Missouri, Columbia, Missouri 65211.
- For *JEP: Animal* send manuscripts to Stewart H. Hulse, Johns Hopkins University, Department of Psychology, Ames Hall, Baltimore, Maryland 21218.
- For *Psychology and Aging* send manuscripts to Timothy Salthouse, Georgia Institute of Technology, School of Psychology, Atlanta, Georgia 30332.

Manuscript submission patterns make the precise date of completion of 1991 volumes uncertain. Current editors will receive and consider manuscripts through December 1990. Should any 1991 volume be completed before that date, manuscripts will be redirected to the newly appointed editor-elect for consideration in the 1992 volume.