Psychonomic Bulletin & Review
2008, 15 (4), 713-731
doi: 10.3758/PBR.15.4.713

Assessing individual differences in categorical data

Jared B. Smith and William H. Batchelder
University of California, Irvine, California

In cognitive modeling, data are often categorical observations taken over participants and items. Usually subsets of these observations are pooled and analyzed by a cognitive model assuming the category counts come from a multinomial distribution with the same model parameters underlying all observations. It is well known that if there are individual differences in participants and/or items, a model analysis of the pooled data may be quite misleading, and in such cases it may be appropriate to augment the cognitive model with parametric random effects assumptions. On the other hand, if random effects are incorporated into a cognitive model that is not needed, the resulting model may be more flexible than the multinomial model that assumes no heterogeneity, and this may lead to overfitting. This article presents Monte Carlo statistical tests for directly detecting individual participant and/or item heterogeneity that depend only on the data structure itself. These tests are based on the fact that heterogeneity in participants and/or items results in overdispersion of certain category count statistics. It is argued that the methods developed in the article should be applied to any set of participant × item categorical data prior to cognitive model-based analyses.

This article develops statistical methods to detect individual differences in cognitive data in the form of categorical observations from a set of participants, each of whom responds to the same set of item events—for example, memory items, item serial positions, or repeated choice trials. The purpose of these methods is to determine whether there is heterogeneity in either participants or item events that might make it inappropriate to aggregate the data, respectively, over participants or item events. These determinations are especially important in the case where the data will be analyzed with a cognitive model. If heterogeneity in participants or item events is detected, several recent developments in computational statistics make it feasible to augment a parametric model with a hierarchical level that models random effects on the parameters. On the other hand, a model that includes random effects assumptions on the parameters is clearly more complex (e.g., Myung & Pitt, 1997) and capable of accounting for more data than the same model is, assuming that all effects are equal; so the decision of whether or not to include random effects in a model analysis is a crucial one. Our methods for informing this decision, because they are based on the sampling assumptions of the basic data structure itself, are completely agnostic as to the appropriate cognitive model underlying the data. Before developing the methods, it is useful to review the way that most cognitive models have been employed in the past and how researchers have tried to deal with parametric heterogeneity.

Parametric stochastic models of cognition are usually assessed with data taken from one or more experimental conditions consisting of a group of participants, each of whom provides responses to the same set of item events (hereafter, *items*). Within a given experimental condition, subsets of these observations are usually pooled (aggregated) and treated as if they are a *sample* from the cognitive model. In the terminology of mathematical statistics, the assumption that a sequence of observations constitutes a sample from a parametric model means that the observations arise from a sequence of *independent and identically distributed* (i.i.d.) random variables, whose common distribution is determined through the model equations by some fixed but unknown set of model parameters (e.g., Hogg, McKean, & Craig, 2005; Lehmann & Romano, 2005). One particular form of this assumption is when data on a particular item are aggregated over participants and analyzed on the assumption of no individual differences—that is, the assumption of *participant homogeneity*. Another form of the assumption is when the data from each particular participant are treated as a sample over items. It is not uncommon in list-memory experiments (e.g., the middle items of free recall, old or new items in recognition memory) to aggregate the data over both participants and items, thereby treating all the observations in some experimental condition as a sample from the model.

Most examples in the cognitive modeling literature that discuss violations of i.i.d. have focused only on the validity of the assumption that the observations are identically distributed, and the assumption of independence, or its counterpart *exchangeability* for Bayesian analyses (e.g., Gelman, Carlin, Stern, & Rubin, 2003), is assumed to be true. There have been several efforts to model sequential effects that arise from violations of the independence assumption. For example, sequential effects have been studied recently in the area of categorization (e.g., Jones, Love,

W. H. Batchelder, whbatche@uci.edu

& Maddox, 2006), absolute identification (e.g., Karpiuk, Lacouture, & Marley, 1997), and choice response time (RT; e.g., Gilden, 2001; Thornton & Gilden, 2005; Wagenmakers, Farrell, & Ratcliff, 2004). In this article, we will study the effects of violations of the identically distributed assumption while assuming independence; however, in the General Discussion we will briefly consider the consequences of violations of independence, as well.

Many studies have shown, sometimes dramatically, that if the aggregated data do not satisfy the identically distributed assumption, analysis of a cognitive model on the aggregated data may lead to misleading conclusions (e.g., Ashby, Maddox, & Lee, 1994; Clark, 1973; Curran & Hintzman, 1995; Estes, 1956; Hintzman, 1980, 1993). In particular, the statistical properties of aggregated data that do not satisfy this assumption may not correspond to the properties of a sample from the model with any fixed values of the model parameters. Furthermore, this lack of correspondence may occur even if the aggregated observations arise independently from the model, with heterogeneity in parameter values over observations. Analyzing inappropriately aggregated data may lead to a false rejection of a valid model; furthermore, if two or more cognitive models are competing in their ability to account for a given set of data, a false assumption of parameter homogeneity can have major effects on the conclusions (e.g., Batchelder, 1975; Haider & Frensch, 2002; Heathcote, Brown, & Mewhort, 2000).

The deleterious effects of aggregation are particularly troublesome with models that involve nonlinear relationships between model parameters and data statistics, because the aggregated data may lead to biased estimates of the parameters. However, linear models are not immune to the troubles that can arise from parameter heterogeneity. In particular, even though parameter estimation for linear models will not be biased and the model is not likely to be falsely rejected due to the presence of parameter heterogeneity, the standard errors of the parameter estimates will be underestimated in most cases. As a result, even in the case of linear models, ignoring parameter homogeneity can present significant issues for those modelers who wish to perform inference or construct confidence intervals on their parameter estimates.

In psychometric modeling, unlike cognitive modeling, researchers have a long history of developing models that permit parameter heterogeneity over participants and items (e.g., de Boeck & Wilson, 2004; Embretson & Reise, 2000; Lord & Novick, 1968). Further, in the last several decades, statisticians have developed a number of classical (e.g., Agresti, 2002; Titterington, Smith, & Makov, 1985) and Bayesian (e.g., Congdon, 2005; Gelman et al., 2003; Gill, 2002) approaches to modeling parameter variability across observations. Recently, a number of these statistical approaches have begun to appear in the cognitive modeling literature to handle cases where there is participant and/or item parameter heterogeneity. These approaches include finite mixture models (DeCarlo, 2002; Klauer, 2006; Lee & Webb, 2005), Bayesian hierarchical models (e.g., Batchelder & Riefer, 2007; Karabatsos & Batchelder, 2003; Rouder & Lu, 2005; Rouder, Sun, Speckman, Lu, & Zhou, 2003), and Dirichlet process modeling (Navarro, Griffiths, Steyvers, & Lee, 2006). What these methods have in common is the assumption that the cognitive model accurately explains the distribution underlying each participant's (item's) data, except that the parameters of the cognitive model that govern these data vary in a statistically specified way over participants (items). In essence, the cognitive model is posed at two levels: a base level and a hierarchical level. On the base level, the model specifies a family of cognitively interpretable probability distributions on the observation space that are indexed by the parameters of the model; and, on the hierarchical level, the model postulates a parametric family of distributions, indexed by hyperparameters, that is designed to describe heterogeneity in parameters across observations. This approach follows standard statistical methods for creating hierarchical (random effects) models (e.g., Gelman et al., 2003; Gill, 2002).

Many of the methods for analyzing hierarchical models are computationally intensive, because in addition to the hyperparameters designed to explain the parameter variability, separate model parameters are drawn from the hyperdistribution for each participant's (or item's) data. All of these parameters enter into the posterior distribution of the parameters given the data. Moreover, some cognitive models have complex, structural specifications that would not allow them to be modified hierarchically in the standard ways described above. In addition, if these approaches are used to augment models that are applied to data sets that closely approximate parametric homogeneity, it is quite possible that incorrect interpretations of the data will result due to overfitting noise in the data (e.g., Myung & Pitt, 1997). Therefore, what are needed are statistical methods in the cognitive modeling literature that can signal violations of the identically distributed sampling assumption. Such methods could be developed separately for each cognitive model; however, it would be more valuable to develop tests of the assumption of homogeneity at the level of the *basic data structure* itself, irrespective of the cognitive model under consideration.

Statistical tests for participant or item heterogeneity based on the sampling assumptions behind the data structure would constitute an important and arguably necessary preliminary step toward deciding how to analyze the data with any particular cognitive model. For example, if our tests fail to detect violations of homogeneity, it would seem unwise to risk overfitting of the data by endowing a cognitive model with a mechanism for parameter heterogeneity. On the other hand, if the tests reveal that there is participant heterogeneity but fail to signal item heterogeneity, one could decide to aggregate the data over items but not over participants. Unfortunately, in some cases of participant heterogeneity, individual participant data may be too sparse to analyze separately with any degree of reliability, and it is in these cases that researchers need a version of their model that allows for random effects over participants in the parameters. In fact, if parameter heterogeneity is properly modeled, statistical inference is often facilitated by a hierarchical model, even if there is sufficient data to analyze each participant separately. Fi-

nally, if the tests indicate that there is heterogeneity in both participants and items, it will be impossible to analyze each datum separately with a parametric model; in order to proceed with data analysis, it will therefore be necessary to add hierarchical assumptions to the model.

We will restrict our analyses to a standard case in cognitive experiments; where in any particular experimental condition each of a set of participants makes a response that falls into one of a small finite set of categories to each of a set of items. This case of categorical data covers a large corpus of list-memory paradigms, including, among others, the experimental setting for free recall, process dissociation, recognition memory, serial list learning, and source monitoring. In addition, most categorization, classification, concept learning, psychometric threshold determination, and prediction experiments involve categorical data. Any of the many paradigms that are covered in the multinomial process tree (MPT) modeling literature (e.g., Batchelder & Riefer, 1999, 2007) involve categorical data over participants and items. Of course, there are familiar paradigms that do not fit into this structure—for example, where RTs or other continuous performance scores are collected. In the General Discussion, we will briefly discuss some ideas for extending our work to paradigms that involve data structures with continuous type data.

Our tests are based on the fact that parameter heterogeneity in most statistical models results in *overdispersion* of some standard statistics of the participants × items data array. Overdispersion of a statistic means that it has a variance larger than the model can explain with a single set of parameter values behind all the observations. In the case of categorical data obtained over participants and items, the natural statistics to examine for overdispersion are ones that relate to the variability of category frequencies across participants and/or items. We will present statistical tests to detect overdispersion, and we will briefly describe some natural hierarchical sampling models that one could employ to help understand the nature of the overdispersion when it occurs.

The outline of this article is as follows. First, we take up the simplest situation where the observations are dichotomous (e.g., yes/no; recalled/not recalled; Category A/Category B) in order to tie our approach to familiar statistical concepts. In this case, the statistics to examine for overdispersion are the participant and item total scores, and we will see that violations of the binomial theorem play an important role in detecting overdispersion in participants or items.

Following the dichotomous case, we consider the more general case of data with more than two categories; this is the main focus of the article. In this case, the statistics to check for overdispersion are ones that quantify the variability of category counts observed by aggregating over participants or items. We provide statistical tests designed to detect overdispersion of these statistics. Following this section, we briefly discuss some standard hierarchical models at the level of the data structure itself that may be used to explore the nature of the overdispersion detected by the tests presented in the previous two sections. In the final section, we will discuss the implications for our analyses of a violation of the independence assumption; we will say something about the case of continuous data, and we will summarize our general recommendations for analyzing categorical data prior to any cognitive modeling analyses. Throughout the article, we will illustrate our tests with data from cognitive experiments.

THE CASE OF DICHOTOMOUS DATA

Data Example 1: Data From a Free Recall Experiment

Before presenting a general framework for developing statistical tests for identifying participant and/or item heterogeneity in categorical data, we will consider a simple free recall memory experiment to illustrate the main ideas behind our approach. In the experiment, we examined whether a standard population of students enrolled in introductory psychology at the University of California–Irvine would show individual differences on a standard multitrial free recall memory task. Our goal was to conduct an experiment that was as similar as possible to the dozens of other free recall experiments in the literature where data were aggregated over participants and items (excluding primacy and recency item buffers). We used the participant pool from introductory psychology classes, which is the usual population of participants in free recall experiments, and we used norms to restrict the variability in the recallability of the items.

Method

Participants. One hundred nine University of California–Irvine undergraduates enrolled in introductory psychology courses received course credit for participation in the experiment. All participants were tested individually in the same four-trial free recall task.

Materials and Procedure. Each participant received a random ordering of the same 40-noun list. Words were chosen through a search of the UWA Psychology: MRC psycholinguistic database with restrictions on familiarity (450–650), concreteness (450–675), imaginability (450–650), length (4–7 letters), and frequency (35–125). These ranges were chosen in order to minimize item heterogeneity of the list words. The 40 nouns were drawn from the resulting database list, such that obvious relations between the words were eliminated. The experiment was programmed to run on an IBM-compatible personal computer using Borland's Visual C++ Builder.

Participants were briefed by the experimenter on the nature of the experiment, then led to individual computer stations. From that point on, all instructions were visually presented on a computer screen. Participants were instructed that they would be shown the same list of words four times in identical order, and after each presentation they would receive a test of their memory for the list. The free recall procedure was described along with the method for response entry. Before the presentation of the list words, participants were reminded of their task to remember as many words as possible for a later free recall test. During the study phase, words were presented one at a time in the center of the computer screen. Each word was presented for 2 sec, with a 1-sec interval between presentations. After list presentation, participants were given a reminder of the response procedure for the recall test. During the test phase, participants typed the recalled words and pressed "enter" in order to register each response. As they typed a response, the word appeared in a small box on the right side of the screen. Each entered word appeared in a list on the right side of the screen so that participants could see all their responses for a given trial. Each testing phase lasted 5 min. The learning and testing phases were then repeated three more times.

Results

We will return to the data from Data Example 1 several times in the article. Figure 1 presents the serial position curves for the experiment for each of the four repeated trials. Each curve corresponds to a trial, and the proportion of correct recalls aggregated over participants is plotted against the serial position of the item in the studied list. All but the first trial exhibit the typical pattern of such curves, namely elevated recall for the items presented early in the study list (primacy items) and the items presented at the end of the list (recency items) and relatively flat recall for the items in the middle of the list. In addition, the serial position curves show that performance at all serial positions tended to improve with repeated trials. On the first trial, the expected recency effect was not observed, probably because some participants had to learn to adapt to the method of computerized data input.

For the first analysis, we dropped the first 6 and last 6 items in the list to reduce the primacy and recency portions of the recall data, resulting in a total of 28 items. This created a case where, in our terms, item homogeneity of the remaining items was likely. Thus, the data on any trial can be viewed as a 109 × 28, participant × item array, $\mathbf{D} = (x_{ij})$, with entry

$$x_{ij} = \begin{cases} 1 & \text{if participant } i \text{ recalled item } j, \\ 0 & \text{otherwise,} \end{cases}$$

where $i = 1, 2, \ldots, 109$ and $j = 1, 2, \ldots, 28$.

If we assume that the observations in **D** arise from the i.i.d. assumption, it follows that the observations on each trial *are* a sample from a Bernoulli process, with some fixed but unknown parameter $p \in (0, 1)$, common to all participants and items. The parameter p represents the probability that any particular entry in **D** is coded as 1. From this assumption, it is easy to see that the row sums or total recall scores,

$$r_i = \sum_{j=1}^{28} x_{ij},$$

for each participant will be binomially distributed. The binomial distribution with parameters M and p is given by

$$\Pr\left(R_i = r_i\right) = \binom{M}{r_i} p^{r_i} (1-p)^{M-r_i}, \tag{1}$$

for $r_i = 0, 1, \ldots, M$, and it has mean equal to Mp and variance $Mp(1 - p)$—in this case $M = 28$—and so the mean and variance of the row sums will be $28p$ and $28p(1 - p)$, respectively. On the other hand, if there is participant heterogeneity, the N row sums are likely to exhibit overdispersion by having a sample variance larger than that predicted by the binomial distribution. A rough gauge of the amount of overdispersion is given by comparing the observed variance of the row sums and the variance derived from the binomial distribution using an estimate of the recall parameter p. The usual estimator of p from a sample of Bernoulli trials is the number of successes (recalls in this case) divided by the total number of observations, and this estimator is also a maximum likelihood estimator, or MLE (e.g., Hogg et al., 2005). In terms of the data array **D**, this yields the estimate

$$\hat{p} = \sum_{i=1}^{109} \sum_{j=1}^{28} x_{ij} / 3{,}052.$$

Table 1 presents a comparison between the actual distribution of the row sums and the best-fitting binomial distribution for all four trials. A large amount of overdispersion is seen in the comparison of the predicted binomial variance

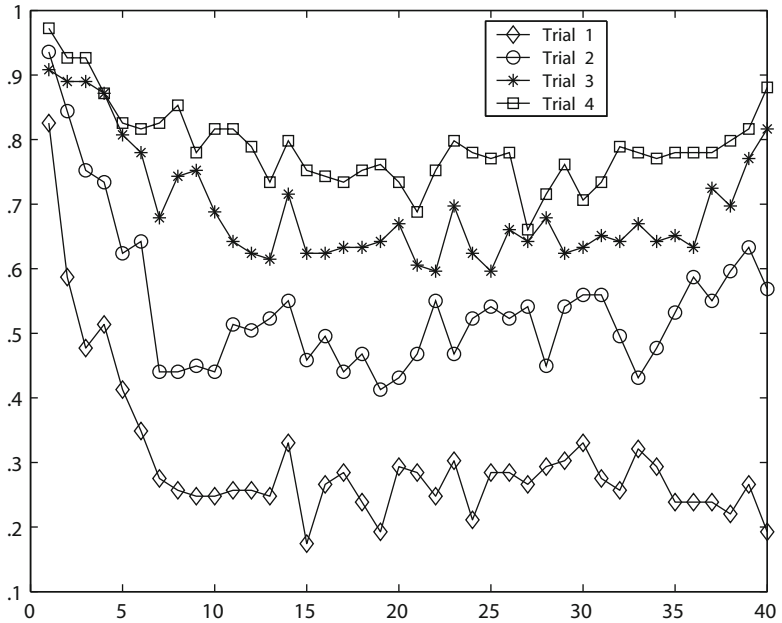

Figure 1. Serial position curves for free recall experiment from Data Example 1. Proportion recalled over item position is presented for each of the four test trials.

Table 1
Variance of Participants' Row Sums (R_i)

	Estimated Variance for Binomial	Observed Variance in Data	χ^2 Test
Trial 1	5.53	12.67	247.6
Trial 2	7.00	18.16	280.2
Trial 3	6.31	24.74	423.2
Trial 4	5.03	17.84	382.8

Note—Analysis used Items 7–34 of each list. Variances for binomial model are estimated using the observed mean. All chi-square values are greater than 142, the critical value for $p = .01$, $df = 108$.

with the actual sample variance of the row sums for all four trials. The amount of overdispersion due to participant heterogeneity is quite striking, because the experiment was designed to be like many other such experiments, in which data are typically aggregated over both participants and items. We think it is reasonable to suspect that large amounts of participant heterogeneity may well be present in most of the published free recall data sets that have been presented in experimental studies of memory.

Certain phenomena in free recall, such as the primacy and recency effect, will not be challenged by the presence of participant heterogeneity, because estimates of group recall proportions are linear combinations of individual participant proportions. However, comparing the amounts of these effects across groups may be affected by heterogeneity because of confidence interval expansion due to heterogeneity. More importantly, most cognitive models of multitrial free recall (e.g., Batchelder, Chosak-Reiter, Shankle, & Dick, 1997) and models of actual recall order (e.g., Howard & Kahana, 2002) are nonlinear, and conclusions based on applying them to aggregated data may be strongly affected by participant heterogeneity.

The same logic that was illustrated with a fragment of the data from the experiment can be used to examine overdispersion in the items, but we will see that the methods to detect heterogeneity in both participants and items simultaneously are not at all straightforward. What we need at this point is a general formal framework for examining overdispersion in both participants and items for data that fall into a finite set of categories. There are two reasons for exploring the dichotomous case in more detail before considering the general case that allows more than two response categories. First, the special case of dichotomous data reveals a number of aspects of our approach to the general case in a familiar setting involving the binomial distribution. Second, cognitive modelers may plan to model more detailed aspects of the dichotomous data, and a discovery of overdispersion at the level of the dichotomous data may signal the need to incorporate a hierarchical level to the cognitive model used to analyze the detailed data. For example, in categorical data one may group several related categories and dichotomize observations as to whether or not they fall into the set of grouped categories.

A Framework for Detecting Overdispersion in Dichotomous Data

Data representation. Assume that data consist of N participants, each responding to M items, where each response is scored into one of two categories—for example, corresponding to "correct," scored as "1," and "incorrect," scored as "0." Then the data structure consists of an $N \times M$ matrix $\mathbf{D} = (x_{ij})_{N\times M}$. In order to understand the data structure statistically, we define dichotomous, 1–0 random variables X_{ij} corresponding to each combination of participant and item, and array them in a corresponding random matrix

$$\mathbf{X} = \left(X_{ij}\right)_{N\times M}. \tag{2}$$

In terms of statistical theory, any data matrix $\mathbf{D}$ is a realization of the matrix random variable $\mathbf{X}$. It is useful to define random variables for the row and column sums of $\mathbf{X}$ in Equation 2, namely

$$R_i = \sum_{j=1}^{M} X_{ij} \quad \text{and} \quad C_j = \sum_{i=1}^{N} X_{ij},$$

with corresponding realizations (data observations)

$$r_i = \sum_{j=1}^{M} x_{ij} \quad \text{and} \quad c_j = \sum_{i=1}^{N} x_{ij},$$

respectively. It is assumed that the participant–item random variables are mutually stochastically independent; however, as mentioned in the introduction, we will discuss the possible effects of violating the independence assumption in the general discussion section. The independence assumption is represented by the probability distribution of $\mathbf{X}$ given by

$$\Pr[\mathbf{X} = \mathbf{D}] = \prod_{i=1}^{N}\prod_{j=1}^{M} \Pr\left(X_{ij} = x_{ij}\right), \tag{3}$$

for all possible realizations $\mathbf{D} = (x_{ij})$. Let the individual participant–item probabilities of a correct response be denoted by $p_{ij} = \Pr(X_{ij} = 1) \in (0, 1)$. Then we can rewrite Equation 3 by

$$\Pr\left[\mathbf{X} = \left(x_{ij}\right)\right] = \prod_{i=1}^{N}\prod_{j=1}^{M} p_{ij}^{x_{ij}} \left(1 - p_{ij}\right)^{\left(1-x_{ij}\right)}. \tag{4}$$

Hypothesis tests for a single source. Equation 4 is a baseline against which we can formulate more restrictive statistical hypotheses about the data structure that reflect the assumption of homogeneity in participants and/or items. These hypotheses reflect restrictions on the individual participant–item probabilities denoted by $\mathbf{P} = (p_{ij})_{N\times M}$. The most stringent hypothesis about $\mathbf{P}$ is the assumption that all the random variables in $\mathbf{X}$ are identically distributed. This assumption amounts to the belief that both participants and items are homogeneous and the X_{ij} form a sequence of Bernoulli trials. This hypothesis is represented by

$$\mathbf{H_1}: p_{ij} = p, \text{ for some } p \in (0, 1), \tag{5}$$

where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, M$. The hypothesis is the same one as was used in the example from the free recall experiment; however, it was not analyzed completely in that application. In the dichotomous case, $\mathbf{H_1}$ corresponds to the hypothesis behind pooling data over both participants and items, so it is desirable to explore its consequences and then develop a statistical test of the hypothesis.

If $\mathbf{H_1}$ holds, there are three immediate consequences on the distributions of the row and column sums of $\mathbf{X}$. The first is that the R_i are i.i.d. binomially distributed, with parameters M and p [denoted by $R_i \rightarrow \text{Bin}(M, p)$; see Equation 1]. The second consequence is that the C_j are i.i.d. $C_j \rightarrow \text{Bin}(N, p)$. Finally, the covariance between any row and any column sum is given by $\text{Cov}(R_i, C_j) = p(1 - p)$, where the positive covariance occurs because R_i and C_j share a single term—namely, X_{ij}.

The next hypothesis is designed to reflect the belief that while there is item homogeneity there may be participant heterogeneity. That hypothesis can be represented by

$$\mathbf{H_2}: p_{ij} = p_i, \text{ for } p_i \in (0, 1), \tag{6}$$

where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, M$. There is a well-known statistical test that can be used to test $\mathbf{H_1}$ versus $\mathbf{H_2}$ by testing independence in an $N \times 2$ contingency table, $\mathbf{C}_{N\times 2}$, displayed in Table 2. Each of the N rows of the table consists of the number of corrects, r_i, and number of errors, $M - r_i$, made by a participant i on the M items. The test is the familiar chi-square test for independence in a contingency table, described in any introductory psychological statistics textbook (e.g., Hays, 1988; Moore & McCabe, 2006). The logic of the test rests on the fact that individual participant differences will be reflected in heterogeneity in the row sums of $\mathbf{D}$. This heterogeneity will lead to a violation of independence between participant and performance scores.

The test statistic that arises from the representation in Table 2 is given in the usual way for tests of independence in a contingency table. Let

$$R_e = \frac{\sum_{i=1}^{N} R_i}{N} \text{ and } R_e^* = M - R_e$$

be the expected number of corrects and errors, respectively, for any participant under the independence assumption for Equation 3. Then the test statistic χ^2 for the chi-square test of independence is given by

$$\chi^2 = \sum_{i=1}^{N}\left[\frac{\left(R_i - R_e\right)^2}{R_e} + \frac{\left(M - R_i - R_e^*\right)^2}{R_e^*}\right]. \tag{7}$$

As long as there is a sufficient number of observations in the cells of the $N \times 2$ contingency table, χ^2 has an approximate chi-square distribution with $(N - 1)$ degrees of freedom (df). In any application one replaces the random variables R_i in Equation 7 with data values r_i and the resulting value of χ^2 can be compared with tabled values of the chi-square with $N - 1$ df. The rule of thumb usually quoted for the adequacy of the chi-square approximation is to have expected counts of at least five per cell (e.g., Hays, 1988). If this condition is not satisfied and M is small, an exact test may be applied (e.g., Agresti, 2002). For larger M with insufficient expected counts, a permutation test developed later in this article provides another reasonable alternative to the statistic in Equation 7.

Table 2
Structure of the $N \times 2$ Contingency Table $\mathbf{C}_{N\times 2}$

	Corrects	Errors	Totals
Participant 1	r_1	$M - r_1$	M
Participant 2	r_2	$M - r_2$	M
...	...	...	M
Participant N	r_N	$M - r_N$	M
Totals	$\sum_{i=1}^{N} r_i$	$N \cdot M - \sum_{i=1}^{N} r_i$	

The third column of Table 1 gives the value of χ^2 for all four of the trials of the free recall experiment. The data satisfied the rule of thumb, so we used the chi-square approximation. With 108 df, we can reject $\mathbf{H_1}$, the assumption of participant homogeneity, at the $p = .01$ level, if $\chi^2 > 142$. As can be seen in the table, participant homogeneity is strongly rejected at this level for all four trials, indicating that there is a significant amount of participant heterogeneity on each trial.

In the free recall example, we suspected that there might be participant heterogeneity, and this led to the strategy of applying the test in Equation 7 to participant totals rather than item totals. In other experimental settings, one might have prior reasons to be more concerned about the possibility of item heterogeneity rather than participant heterogeneity. In that case, if one were willing to assume participant homogeneity, one could use the same approach based on a test of independence on the item totals—namely, the c_j. The only difference is that the appropriate data table would be an $M \times 2$ table with entries c_j and $N - c_j$ for each item, $j = 1, 2, \ldots, M$.

Use of the chi-square test to compare hypotheses $\mathbf{H_1}$ and $\mathbf{H_2}$ is designed to detect heterogeneity in only one source, and from Equations 5 and 6 it requires the assumption of homogeneity in the other source. Because the previous test indicated strong evidence for participant heterogeneity, it is not a good idea to proceed with the chi-square test for item homogeneity. In the next subsection, we provide a test that can be used to test homogeneity in one of the sources even if heterogeneity is present in the other.

Hypothesis tests for heterogeneity in both participants and items. The situation becomes more complicated than the preceding test for homogeneity if one wants to test for homogeneity for both participants and items. The difficulty arises from the fact that heterogeneity from one source may mask or shrink measures of overdispersion of the other source. For example, in the recall experiment, the most extreme case of participant heterogeneity would be where each participant either recalled all M items or failed to recall all M items. In that case all the column sums, the c_j, would be identical and there would be no variability in the item total scores. More generally, next we show that under our independence assumption increasing variability in the row (column) sums results in decreasing variability in the column (row) sums.

Suppose for $\mathbf{X} = (X_{ij})$ in Equation 2, the R_i are independent, and $R_i \rightarrow \text{Bin}(M, p_i)$ for each i. Let

$$\mu_p = \sum_{i=1}^{N} p_i / N$$

denote the mean of the participant success probabilities. Then for any $j = 1, 2, \ldots, M$, independence of the variables in **X** implies that

$$\mathrm{Var}\left(C_j\right) = \sum_{i=1}^{N} \mathrm{Var}\left(X_{ij}\right) = \sum_{i=1}^{N} p_i\left(1 - p_i\right).$$

Carrying out the sum in this equation and substituting from the formula for the variance of the participant probabilities,

$$\sigma_p^2 = \left(\sum p_i^2 / N\right) - \mu_p^2,$$

yields

$$\mathrm{Var}\left(C_j\right) = N\mu_p\left(1 - \mu_p\right) - N\sigma_p^2. \quad (8)$$

Equation 8 is maximized when $\sigma_p^2 = 0$, namely for $p_i = \mu_p$; $i = 1, \ldots, N$; and $\mathrm{Var}(C_j)$ decreases monotonically as σ_p^2 increases. Thus, we have shown that given a particular mean, μ_p, inhomogeneity in the participant success probabilities as measured by σ_p^2 shrinks rather than increases column sum variability; and, of course, the same follows for the effect of column sum variability on row sum variability. This shrinkage means that the chi-square test in Equation 7 might not signal inhomogeneity in a source, even if, in the presence of inhomogeneity, it is in the other source, as well.

If variability of both items and participants is of concern, a nonparametric Monte Carlo permutation test can be constructed, instead of the chi-square test, to check for inhomogeneity in either items or participants. In general, permutation tests condition on some aspect of the data, then sample many possible arrangements of data, given these conditions. For each data arrangement, a particular statistic of interest is calculated along with the probability of the arrangement (the probability calculations are omitted when all permutations are equally likely given the relevant conditions). This process results in an empirical distribution of this statistic that can be used as the reference distribution for a test on the data of interest. Early applications of the permutation tests were limited by the large number of permutations required to fully realize the empirical distribution for anything but very small data sets. This difficulty has been resolved in the last several decades with the advent of more-powerful computers and the introduction of Monte Carlo sampling schemes (e.g., Agresti, 1992; Efron & Tibshirani, 1993). The example below develops the permutation test to check for item heterogeneity in the face of possible participant inhomogeneity, although with obvious notational change it can be adjusted to test for participant variability as well.

The first step of the test is to calculate the variability of the item scores in **D** using some measure of variability such as $\mathrm{Var}(C_j)$ or the χ^2 value from Equation 7 when it is adapted to item variability rather than to participant variability. The permutation test developed here is flexible and can be used with almost any measure of column sum variability. The logic of the test is to compare the observed item variability score with the distribution of item variability scores obtained by examining a suitable number of appropriate permutations of the data matrix **D**. The appropriate permutations are ones consistent with the idea that within each participant all items are equally likely to receive a score of 1. This consideration amounts to the assumption that any permutation of a participant's row is equally likely. Ideally, the test looks at a frequency distribution of item variability scores based on every possible combination of permutations of the entries in each row of **D**. However, in practice the number of combinations of permutations (namely $M!^N$) may be exceedingly large. So instead of an exhaustive search of all possible combinations of permutations, one can simulate a random sample of a large number by suitable Monte Carlo draws. The analysis can be done using basic commands in statistical programming languages such as R (R Development Core Team, 2005). (For details on programming the permutation test, see the Appendix.) If a straightforward variability measure (e.g., variance) is used, the simulation is computationally simple; for example, a permutation test of 10,000 draws for a data set of 50 participants, and 100 items took less than a minute to run on a modern desktop computer.

The resulting variability scores provide a distribution under the null hypothesis of item homogeneity, and these scores are not based on the assumption of homogeneity of the other variable (in this case, participants) as required for the chi-square test described earlier. The percentile rank of the observed variability score within the distribution of simulated variances can then be used to derive a p value for the test of the hypothesis of no item variability. In particular, if the observed value of the item variability lies at the upper $100(1 - p)\%$ of the simulated distribution, one can reject the hypothesis of item homogeneity at the corresponding value of p. However, if the observed variability score is very small relative to the distribution, say at the 1% or 5% value, it is not valid to conclude that item homogeneity holds. In fact, an observed very small value in the distribution may indeed cast doubt on the original independence assumption in Equation 3, namely that the X_{ij} are independent. This difficulty is not unique to permutation tests of homogeneity, and it will be further elaborated in the discussion section.

We applied this permutation test to the free recall data from Data Example 1 for the middle 28 items (serial positions). We used the variance of the column totals as the measure of item variability. A sample of 10,000 combinations of permutations was simulated to make up the reference distributions. The test results provided no evidence to reject the belief that the middle items were homogeneous. In particular, Table 3 reports the observed column variances for all four trials, and they are well within the middle 95% range of the variances obtained from the permuted data sets. We also applied this permutation test to examine

Table 3
Permutation Test on Items 7–34 for Middle Serial Positions

	2.5th Percentile	97.5th Percentile	Actual Variance	p
Trial 1	11.22	32.78	18.70	.20
Trial 2	13.89	41.07	35.15	.90
Trial 3	11.81	34.92	25.88	.76
Trial 4	9.65	28.40	22.54	.84

Note—To derive the distribution of the variance of the column sums, 10,000 permutations were used.

Table 4
Permutation Test on Items for All Serial Positions

	2.5th Percentile	97.5th Percentile	Actual Variance	p
Trial 1	13.51	32.99	166.17	<.0001
Trial 2	15.65	38.47	150.47	<.0001
Trial 3	12.92	31.74	87.84	<.0001
Trial 4	10.28	24.89	47.76	<.0001

Note—To derive the distribution of the variance of the column sums, 10,000 permutations were used.

variability across all 40 serial positions. Since we included both the primacy and the recency items, we suspected that there would be inhomogeneity in items (serial positions). As shown in Table 4, the observed variances of the column totals fall outside the middle 95% range of the variances calculated from the randomly permuted data sets.

In general, we recommend that researchers employ the permutation test rather than the more familiar chi-square test of independence, because it does not require special conditions on the cell counts; more importantly, it is a valid test of homogeneity in one of the sources, whether or not there is variability in the nontested source. A possible concern with using the permutation test is that it might not have the same power characteristics as a suitable parametric test; in the Appendix, we report one study of the power of the permutation test that leads us to believe that the test has good power characteristics.

THE CASE OF CATEGORICAL DATA

A Framework for Categorical Data

Data representation. Dichotomous data are a special case of the more general categorical data framework. We can derive tests for participant and/or item homogeneity in categorical data by expanding the logic used for the dichotomous case. Categorical data consist of observations that each fall into one of K categories $\mathbf{C} = \{c_1, c_2, \ldots, c_K\}$. The statistical analysis of categorical data requires a straightforward modification of the notation surrounding Equation 2. The data on participants and items can be represented in a three-way matrix $\mathbf{D} = (x_{ij,k})_{N\times M\times K}$, where $x_{ij,k} = 1$ if the response of participant i to item j is in category c_k, otherwise it is zero. In this formulation, each participant × item provides a vector of length K over the categories with a single entry "1" for the category of the response and the rest of the entries are zeros. Any particular $\mathbf{D}$ can be constructed from the observations of a matrix random variable $\mathbf{X} = (X_{ij,k})_{N\times M\times K}$, where

$$X_{ij,k} = \begin{cases} 1 \text{ if participant } i \text{ responds to item } j \\ \quad \text{in category } c_k, \\ 0 \text{ otherwise.} \end{cases} \tag{9}$$

The random variables defined above are subject to a structural constraint that for every combination of participant and item, (ij),

$$\sum_{k=1}^{K} X_{ij,k} = 1,$$

so they are obviously not independent. However, in this case the independence assumption postulates that the $N \times M$ random vectors $\mathbf{X}_{ij} = (X_{ij,1}, \ldots, X_{ij,K})$ are mutually stochastically independent; that is,

$$\Pr\left[\left(\mathbf{X}_{ij}\right)_{N\times M} = \left(\mathbf{x}_{ij}\right)_{N\times M}\right] = \prod_{i=1}^{N}\prod_{j=1}^{M} \Pr\left(\mathbf{X}_{ij} = \mathbf{x}_{ij}\right) \tag{10}$$

for every possible pattern of realizations $(\mathbf{x}_{ij})_{N\times M}$.

Corresponding to each combination of participant and item, there is a marginal probability distribution over the K categories, $\mathbf{P}_{ij} = (p_{ij,k})_{1\times K}$, where $p_{ij,k} = \Pr(X_{ij,k} = 1)$, $0 < p_{ij,k} < 1$, and

$$\sum_{k=1}^{K} p_{ij,k} = 1.$$

These marginal probability distributions, the $\mathbf{P}_{ij}$, are analogous to the p_{ij} of the dichotomous case and constitute the parameters of a general participant × item categorical data structure.

Hypothesis tests for a single source. If we assume mutual stochastic independence of the $\mathbf{X}_{ij}$, then it is possible to formulate a number of hypotheses concerning the $\mathbf{P}_{ij}$. The simplest case is to test to see whether the participants are homogeneous under the assumption that there is item homogeneity. This hypothesis is represented by

$$\begin{aligned} &\mathbf{H_1}: p_{ij,k} = p_k, \\ &\quad \text{for all } i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, M, \\ &\quad 0 < p_k < 1, \text{ and } \sum_{k=1}^{K} p_k = 1, \end{aligned} \tag{11}$$

and the alternative hypothesis that there may be participant heterogeneity is represented by

$$\begin{aligned} &\mathbf{H_2}: p_{ij,k} = p_{i,k}, \\ &\quad \text{for all } j = 1, 2, \ldots, M, \\ &\quad 0 < p_{i,k} < 1, \text{ and } \sum_{k=1}^{K} p_{i,k} = 1. \end{aligned}$$

Equation 11 is an expansion of the hypothesis in Equation 5 to cover the case where $K \geq 2$. Equation 5 stated the conditions that led to the chi-square test in Equation 7. The hypothesis in Equation 11 states that all participants and items, not just two categories in the dichotomous case, are assumed to share the same distribution on the K categories. Let

$$F_{i,k} = \sum_{j=1}^{M} X_{ij,k}$$

be a random variable denoting the number of items that fall into category k for participant i, $k = 1, \ldots, K$; $i = 1, \ldots, N$. Then, under hypothesis $\mathbf{H_1}$, the frequency counts $\mathbf{F}_i = <F_{i,k}>_{k=1}^{K}$ over categories are governed by the multinomial distribution, with $\mathbf{P} = <p_k>_{k=1}^{K}$ [denoted by Mult(M, $\mathbf{P}$)], given by

$$\Pr\left(\mathbf{F}_i = <f_{i,k}>_{k=1}^{K}\right) = M! \prod_{k=1}^{K} \frac{p_k^{f_{ik}}}{f_{ik}!}, \tag{12}$$

where $<f_{i,k}>_{k=1}^{K}$ is any vector of nonnegative integers whose sum is M. For the multinomial distribution, the mean counts in any category c_k are given by $\mu_k = -Mp_k$, the variance is given by $\sigma_k^2 = Mp_k(1 - p_k)$, and for any two categories, k and k', $\mathrm{Cov}(f_k, f_{k'}) = -Mp_kp_{k'}$ (Evans, Hastings, & Peacock, 2000). If $\mathbf{H_1}$ is false (under our independence assumption), there should be overdispersion in the category counts across participants. There are several ways to assess this overdispersion under the assumption of item homogeneity assumed as a part of $\mathbf{H_1}$. One possible measure of dispersion is the random variable

$$SS = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(F_{i,k} - \bar{F}_k\right)^2, \tag{13}$$

where

$$\bar{F}_k = \left(\sum_{i=1}^{N} F_{i,k}/N\right)$$

is the mean over participants of the counts in category c_k. The value of *SS* represents the average of the squared discrepancies of each participant's counts from the average counts in a category, so it is analogous to the variance of the row scores used for assessing participant homogeneity in the dichotomous case. However, unlike the case of row score variance, which has the binomial variance as the prediction for homogeneity, *SS* does not have a convenient reference distribution. The method below overcomes this problem by expanding the chi-square test that was employed in the dichotomous case.

It is easy to expand the chi-square test in Equation 7 to cover the case of categorical data. The first step is to aggregate the data over items to create an $N \times K$ contingency table $\mathbf{C}_{N\times K} = (F_{i,k})_{N\times K}$, analogous to Table 2 with K columns. Since these frequencies are aggregated over M items,

$$\sum_{k=1}^{K} F_{i,k} = M.$$

Hypothesis $\mathbf{H_1}$ implies that there will be independence in $\mathbf{C}_{N\times K}$, and from that assumption the expected counts $F_{i,k}^*$ in the ikth cell can be calculated in the usual way (e.g., Hays, 1988). The result is

$$F_{i,k}^* = \frac{M \cdot F_k}{N \cdot M} = \frac{F_k}{N},$$

where

$$F_k = \sum_{i=1}^{N} F_{i,k}.$$

Then the value of χ^2 for the test is given by

$$\chi^2 = \sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\left(F_{i,k} - F_{i,k}^*\right)^2}{F_{i,k}^*}. \tag{14}$$

If there are sufficient counts in the contingency table to justify the chi-square approximation, χ^2 will have an approximate chi-square distribution with $(N - 1) \times (K - 1)$ *df*. Both the rule of thumb for sufficient counts and the option using a Fischer exact test or permutation test, if there are not sufficient counts, are the same as discussed in the last section for the chi-square test in the case of dichotomous data. There are also tests related to the chi-square test that can be used if the number of items each participant responds to are not identical (e.g., Kim & Margolin, 1992).

Data Example 2: Recall Clustering of Schizophrenics

It is possible to apply the chi-square test to an experiment reported by Riefer, Knapp, Batchelder, Bamber, and Manifold (2002). In that article, several groups of patients in a VA hospital were run in a multitrial free recall experiment involving free recall of clusterable pairs. The data were analyzed with a multinomial model of pair clustering developed by Batchelder and Riefer (1980, 1986). Participants in this experiment were given lists of words to study in which each word had a category associate (e.g., *car*, *taxi*; *apple*, *pear*) widely separated within the study list. In this case, items corresponded to pairs of words related by category membership. Response categories were delineated by the recall performance on a category pair during test. The recall of any give pair (item) was coded into four categories: c_1, both words recalled adjacently; c_2, both words recalled but not adjacently; c_3, exactly one word in a pair recalled; and c_4, both words not recalled.

In order to illustrate the chi-square test, we selected a group of $N = 29$ participants classified as schizophrenic, each of whom provided recall data on $M = 20$ pairs of clusterable items on each of six study–test trials. We applied the χ^2 statistic in Equation 14 separately for each of the six trials of the schizophrenic group to test the hypothesis of participant homogeneity in Equation 11. With 20 items per participant and four categories, the average count per category was 5, so we used the chi-square approximation. There were 29 participants, so the test has $(29 - 1) \times (4 - 1) = 84$ *df*. The values of χ^2 from Equation 14 are reported in Table 5 for each trial. Because the critical value for a chi-square at the $p = .01$ level is 117.06, we can comfortably reject the hypothesis of participant homogeneity, and in fact the evidence for participant heterogeneity is seen to increase over the first few trials. This increase is likely due to the different learning rates of the participants, as well as to differences in their storage and retrieval capacities.

The data in Table 5 suggest that the step of pooling over participants in analyzing the data with the pair-clustering model is questionable. Riefer et al. (2002) did aggregate the data over both participants and items in their application of the pair-clustering model. However, they also provided separate analyses with the model that assumed participant heterogeneity, and the values of the estimated parameters were strongly affected by pooling the data over

Table 5
Chi-Square Statistic for Participant Homogeneity in the Schizophrenia Data

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
χ^2	123.1	172.5	259.4	307.2	306.6	310.7

Note—All chi-square values are greater than 117.06, the critical value for $p < .01$ with $df = 84$.

participants. Despite the effect on parameter estimates, the authors showed in this study that none of the main conclusions they drew from the pooled data were changed when participant heterogeneity was considered, and this rather fortuitous result was important for their work. However, high levels of heterogeneity are likely to be present in experimental groups defined by special populations such as schizophrenics, and we expect that cognitive modeling studies with such groups will inevitably require the addition of random effects assumptions to a model. Batchelder and Riefer (2007) is an extended discussion of the issue of individual differences in clinical studies, including the one discussed above in Data Example 2.

In this section, we have developed a chi-square test of participant homogeneity, under the assumption that item homogeneity holds. There may be situations in which one is concerned with the possibility of item homogeneity in a case where participant homogeneity can be assumed. In such situations, the chi-square test in Equation 14 can be used on the frequency counts for each item aggregated over participants. Basically, the roles of N and M are reversed in the relevant equations. However, when heterogeneity in both participants and items is possible, it is necessary to provide different tests. The next section provides these.

Variability from two sources. For the case of dichotomous data, we showed in Equation 8 that an increase in variance in one source leads to a decrease in observed variance in the other. We cannot show, as we did in the dichotomous case, the result that in all cases increases in a variability measure on one source decreases variability in the other source; but there is reason to be concerned. For one, any partition of the category system into two categories will manifest this effect by the logic shown for dichotomous data. In addition, we have done simulation studies in which we modeled both participant and item inhomogeneity, using a categorical version of the well-known Rasch (1960) model from psychometric test theory. The results showed that as participant inhomogeneity increased, variability in the frequency counts across items measured by χ^2 in Equation 14 (reversing the roles of participants and items) decreased. This shrinkage is analogous to that found for the column sum variance in the dichotomous case.

The issue of item variability in the face of possible participant inhomogeneity can be resolved in the multinomial case through a permutation test similar to that used for dichotomous data. Recall that in the multinomial case the response matrix, **D**, has an added dimension compared with the binomial case. For each participant–item, this dimension has K elements with $K - 1$ elements set to "0" and one cell set to "1" to indicate the response category. The general structure of the permutation test remains the same for the multinomial case except that we permute over items the row of K element vectors for each participant rather than the row of 1–0 responses of each participant's item responses. For each permutation, we calculate the $N \times K$ contingency table, $\mathbf{C}_{N\times K}$. Given this representation, one can use any relevant measure of variability such as SS from Equation 13, or χ^2 from Equation 14. As in the dichotomous case, the permutation test is flexible and other measures of variability may be used. Outside of these minor differences, the permutation test for the multinomial case follows the same steps as for the dichotomous test. The multinomial permutation test can be implemented through the same programs as the dichotomous permutation test (see the Appendix for sample code in MATLAB and R).

Table 6
Song Data Results From Permutation Test With χ^2 As the Statistic

	2.5th Percentile	97.5th Percentile	χ^2 Critical Value	Actual Variance
Participant test	108.75	170.10	182.40	550.40
Item test	42.82	82.30	93.86	316.13

Note—To derive the distribution of the χ^2 statistic, 10,000 permutations were used. The χ^2 critical value is the 97.5th percentile of the chi-square distribution, with df of 147 for participants and 69 for items.

Data Example 3: Song Data

We applied the permutation test to a 24 participant × 50 item, tip-of-the-tongue experiment from Riefer, Kevari, and Kramer (1995). Participants were presented with the theme songs to a series of 50 television shows and were asked to name the title of the corresponding show. Responses were separated into 4 categories: c_1, correct response; c_2, a partial response (e.g., able to name actors from the show but unable to recall the show's name); c_3, an incorrect response; and c_4, "don't know." Since each participant was exposed to the same 50 television shows, the data structure fits the framework in Equation 9 for categorical data. Thus, we can perform one permutation test for participant homogeneity and a separate one for item homogeneity without having to worry about the possible effect of inhomogeneity from the other source. The results of these permutation tests using the χ^2 statistic for both participants and items are shown in Table 6. The χ^2 statistic for the observed data falls well above the simulated 95% confidence interval for both participants and items.

The occurrence of significant item heterogeneity in this study is not that surprising because, unlike most experimental studies of memory, there was no effort to select homogeneous items, and in fact the items were chosen from many different types such as dramas, comedies, and westerns. The reader may note that in this study, one would have come to the same conclusion by using the chi-square test of the statistic χ^2 in Equation 14; however, as indicated in Table 6, the 95% range resulting from the permutation test is moderately smaller than would be expected under a suitable chi-square distribution. These results show the effect of shrinkage in the variability score of one source produced by variability in the other source.

MODELING INHOMOGENEITIES IN THE DATA

The Case of Heterogeneity in a Single Source

If heterogeneity in participants and/or items is detected, the next step in applying a cognitive model is to decide how best to analyze the data with the model. If there is

only one source of inhomogeneity, one solution if possible is to analyze the data separately for each level of that source. This practice is standard in the area of psychophysics, where it is understood that each participant has his or her own energy transduction characteristics, and it may be a necessary step to take in certain cognitive tasks in which there is not a single cognitive model flexible enough to describe the behavior of all of the participants (items). However, in many cognitive experiments, there is not enough data to produce reliable estimates for individual participants (items). It is in these cases that it is especially attractive to augment a cognitive model with a hierarchical level that specifies the nature of parametric random effects.

Even if there is enough data to analyze participants (items) separately, it may be desirable to employ a hierarchical modeling analysis. For example, a correctly specified hierarchical model will in general provide better estimates of participants' parameters than individual participant analyses will; however, as the amount of data per participant increases, the amount of improvement provided by hierarchical modeling decreases. As a result, the choice between hierarchical modeling and individual participant (item) analysis depends on trade-offs in the researcher's needs concerning accuracy of parameter recovery, the amount of data available per unit, and the accuracy with which parameter variability is specified in the model.

If one decides to add random effects to a cognitive model to handle heterogeneity in participants (items), there are two general ways that one might proceed. The first way would be to specify random effects assumptions into one's cognitive model without further data exploration; the second way would be to continue to explore the heterogeneity in the data with a number of standard hierarchical models only on the basis of the sampling assumptions of the data. The first way requires one to select a reasonable specification for the random effects on the basis of psychological considerations and incorporate it directly into the cognitive model. Since most models involve several parameters, such a specification will involve either some smooth multivariate hyperdistribution on the parameters or a discrete distribution leading to a finite mixture model. The introduction provides references for both approaches to specifying parametric random effects, but it may not be clear which approach is most productive in any particular case. For example, the decision to go with a finite mixture model should be based on the belief that there are several types of participants (items), within each of which there is reason to be comfortable with the i.i.d. assumption. This approach would be natural, if the participants were drawn from special populations where substantially different degrees of a clinical condition or cognitive deficit are likely to be present. On the other hand, if the participants have been drawn from introductory psychology classes, it is arguable that the marginal distribution of any particular parameter probably would be a smooth unimodal distribution over some latent cognitive skill, and this would be a reason to avoid the mixture model approach. Of course, there is a possibility of specifying a mixture model with heterogeneity in observations within a component of the mixture. This approach is discussed in psychometric test theory (e.g., von Davier & Carstensen, 2007), but to our knowledge it has yet to be employed in cognitive modeling.

There is a potential problem with proceeding directly from the detection of heterogeneity in a single source to specifying a hierarchical version of one's cognitive model. If the hyperdistribution for a cognitive model is misspecified, the analysis of the data may be misleading, even if the base cognitive model is correctly specified. Fortunately, practice has shown that misspecification at the hierarchical level is not as severe as misspecification at the base level (e.g., Agresti, Caffo, & Ohman-Strickland, 2004; Gelman et al., 2003; Gill, 2002). Nevertheless, this direct approach to handling heterogeneity may not be desirable in some cases. For example, when several models are in close competition, misspecification at the hierarchical level may lead to an incorrect choice among models. Also, if it is not clear to the researcher whether to employ a mixture model or a smooth hyperdistribution, an incorrect choice may lead to problems of interpretation. For example, in some of our work we have found that fitting mixture models with small numbers of components to data simulated from smooth hyperdistributions has led to mixture components that handle the data well but invite misleading conclusions about the distribution of participant parameters. We believe that the issues concerning a choice between mixture models and smooth hyperdistributions in cognitive modeling can be an important topic for further research.

The second way to deal with data where heterogeneities are detected in a single source is to study the data with some natural random effects versions of the underlying data sampling model, such as the binomial distribution in Equation 1, or the multinomial distribution in Equation 12. These models are natural ones, based on the data structure itself, not on any theoretical cognitive assumptions; so they are agnostic as to the underlying cognitive model behind the data. Although rare today in cognitive modeling, random effects modeling of the variability in data is common within biostatistics and psychometrics, including the use of a number of statistically sophisticated methods. We believe that a researcher may be able to use the parameter distributions estimated from one of these augmented models to better understand the nature of the underlying participant or item variability (e.g., Albert, 1999; Albert & Chib, 1997). In turn, this exploratory knowledge may point the way to how best to incorporate random effect assumptions into a particular cognitive model.

One such model commonly used for explaining variability in one source for categorical data is the Dirichlet-multinomial model, which has as a special case the beta-binomial model for the case of two categories (e.g., Evans et al. 2000). We will illustrate this approach by applying the beta-binomial model to the free recall data discussed earlier. Another general approach for handling parameter variability in the sampling model is finite mixture modeling (e.g., McLachlan & Peel, 2000; Titterington et al., 1985). In the case of categorical data, a finite mixture

multinomial model assumes that each vector of category frequencies in Equation 12 is generated by the multinomial distribution with one of a small, finite set of possible probability distributions, instead of the single distribution in Equation 12. For example, in the case of participant variability, each participant is thought to fall into one of several latent classes, each equipped with its own probability distribution over the categories. Next, we illustrate the beta-binomial model for dichotomous data as one example of how data exhibiting heterogeneity in a single source can be further explored with models motivated by the data sampling assumptions.

The beta-binomial model for dichotomous data. Consider the case of dichotomous data in Equation 2, in which tests have revealed participant variability but in which it is reasonable to assume item homogeneity. In this case, each participant's responses over items follow a Bernoulli process, whereby the probability parameter p varies from participant to participant. We can postulate a hyperdistribution, $g(p)$, $0 < p < 1$, of the individual participant probabilities, along with the assumption that the participant probabilities are drawn i.i.d. from this distribution. A natural and flexible distribution family for the p_i is the beta distribution [denoted by Beta(α, β)] given by

$$g(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1}, \qquad (15)$$

where $\Gamma(.)$ is the gamma function and the parameters satisfy $0 < \alpha, \beta$ (e.g., Evans et al., 2000). The beta distribution has mean

$$E(p) = \frac{\alpha}{\alpha+\beta} = \mu_p$$

and variance

$$\mathrm{Var}(p) = \frac{\mu_p\left(1-\mu_p\right)}{\alpha+\beta+1} = \sigma_p^2.$$

Figure 3, discussed later, depicts the shape of four beta distributions that were estimated to account for participant variability in the recall probability for the four trials of the free recall experiment in Data Example 1. In general, if $\alpha, \beta > 1$, the beta distribution is unimodal, like those in Figure 3; however, for $\alpha = \beta = 1$, the beta becomes a uniform (flat) distribution; for smaller values of the parameters, U-shaped distributions occur.

Under the assumption that individual participant recall probabilities p_i come from a beta distribution, the row sums R_i of Equation 2 are independent and distributed as Bin(M, p_i), where in turn the p_i are i.i.d. with hyperdistribution given by Beta(α, β). These assumptions lead to the result that the row sums are i.i.d. (or exchangeable for a Bayesian) from a hierarchical model given by the beta-binomial distribution, Bb(M, α, β), whose marginal density is given by

$$\Pr(R = r \mid \alpha,\beta,M) = s_2\binom{M}{r}\Gamma(\alpha+r)\Gamma(\beta+M-r), \qquad (16)$$

where $r = 0, 1, \ldots, M$ and

$$s_2 = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+M)}.$$

The distribution has mean and variance given, respectively, by $E(R) = M\alpha/(\alpha+\beta)$ and

$$\mathrm{Var}(R) = \left[\frac{M\alpha\beta}{(\alpha+\beta)^2}\right]\left[\frac{(\alpha+\beta+M)}{(\alpha+\beta+1)}\right]. \qquad (17)$$

Equation 17 expresses the variance as a product of two bracketed terms. The first bracketed term can be viewed as the variance of the row sum under assumption based just on the means as in the binomial variance $M\mu_p(1-\mu_p)$ in the case that all participants have the same probability $p = \mu_p$. The second bracketed term is displayed in Equation 18 as OD, and it quantifies the overdispersion due to participant heterogeneity; namely,

$$OD(\alpha,\beta,M) = \frac{(\alpha+\beta+M)}{(\alpha+\beta+1)}. \qquad (18)$$

Thus, if the R_i are distributed as a random variable R which has distribution Bb(M, α, β), Equation 18 shows that the overdispersion grows linearly with the number of items per participant. Equation 18 also shows that for fixed M as $(\alpha+\beta)$ goes to infinity decreasing the variance of the beta distribution of p, Var(R) approaches the variance of a Bin(M, μ_p) random variable appropriate for explaining the row sums if the data do not indicate participant heterogeneity.

We used a Newton–Raphson algorithm programmed in MATLAB to obtain best fits to participant variability in the free recall data of Data Example 1, and we cross-checked the result with an available program in Minka (2000). Figure 2 illustrates the beta-binomial fit to the data from Trial 2 of the free recall experiment. The figure provides the empirical frequency distribution over participants of the number of correct recalls out of the 30 middle list items. The data are grouped into 10 bins, each of size three. The figure also provides the best-fitting binomial and beta-binomial distributions for the data. Notice that the beta-binomial distribution provides a good fit to the data, and, as expected, it has a larger variance than the best-fitting binomial on the basis of the assumption of homogeneity in the participants. The other three trials of the free recall experiment were similarly better fit with the beta-binomial distribution, and Figure 3 exhibits the best-fitting beta hyperdistributions of the participant recall probabilities for all four trials.

Sources for hierarchical multinomial models. In this short section, we provide references for the reader who might want to employ hierarchical data models to explore heterogeneity in a single source (participants or items) in the categorical data structure in Equations 9 and 10. These models are all in the spirit of the beta-binomial model discussed in the preceding section, and a lot of computational software for them is readily available. We think further data exploration with these methods may increase the likelihood that appropriate hierarchical specifications will be selected for a particular cognitive model; however,

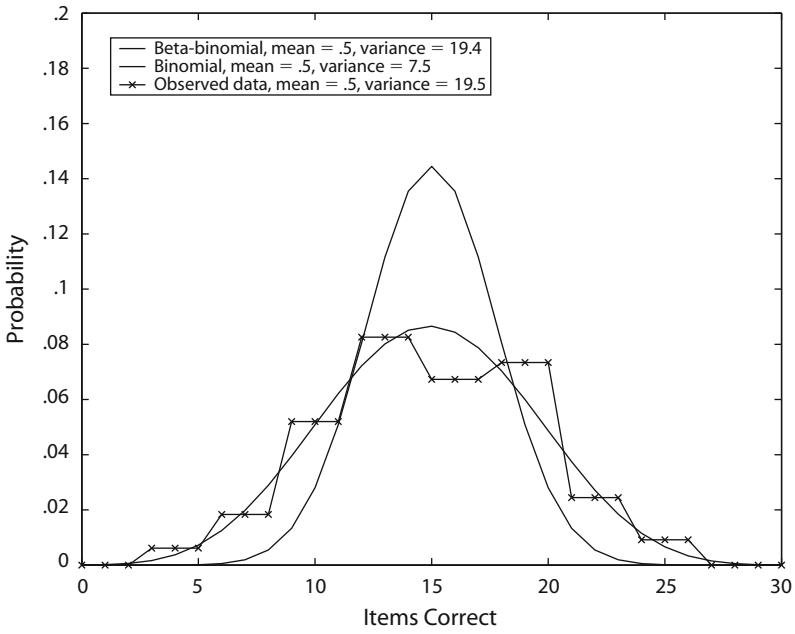

Figure 2. Comparison between observed data for trial two and best fit curves for the binomial and beta-binomial models for the free recall data in Data Example 1. Note that the above data and distributions are discrete but drawn as lines in order to better differentiate the plots. The raw data have been placed into bins of three.

until a large body of practical application has developed, it would be premature to suggest that this approach should necessarily be followed in dealing with heterogeneity. It is useful to reiterate, however, that the proposed analyses are neutral as to the appropriate cognitive model for the basic data-generating mechanism, because they depend only on the sampling theory for the categorical data structure.

There are a lot of references to facilitate data analysis with the Dirichlet-multinomial distribution and the special case of the beta-binomial distribution. Inference for

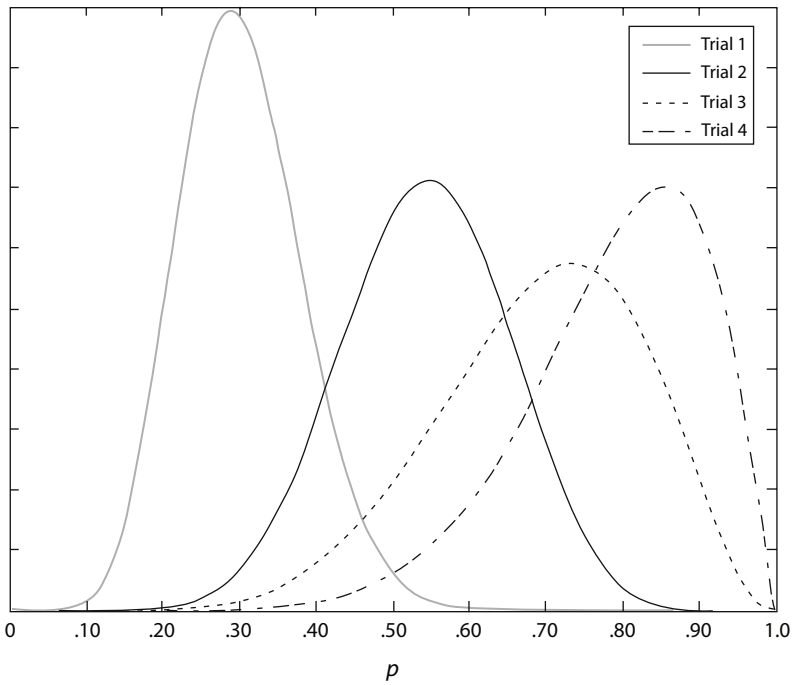

Figure 3. Density plots of the estimated beta distributions of participant recall probabilities p_i for the free recall data in Data Example 1 for all 40 items.

the beta-binomial model is discussed in Griffiths (1973), and Madden and Hughes (1994) describe a stand-alone program. There is also a built-in package for the beta-binomial in SAS. In the case of the Dirichlet-multinomial model, Mosimann (1962) discusses inferential issues, and Minka (2000) discusses parameter estimation routines that are implemented in a MATLAB toolbox known as Fastfit. There is an R package known as VGAM which can also fit the Dirichlet-multinomial. Kim and Margolin (1992) compare the multinomial model against overdispersed alternatives, and it covers the special case, the beta-binomial model, for the case of two categories (see also Garren, Smith, & Piegorsch, 2001). Fitting finite mixture multinomials is discussed in two books, one by Titterington et al. (1985), the other by McLachlan and Peel (2000). These books include estimation methods, discussions of model identifiability, and various decision rules for selecting the number of components to use to handle overdispersion. A frequently used approach to estimating mixture models involves the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), whereby one basically treats the assignment of cases to latent classes as "missing data" and alternates calculations of within-class maximum likelihood estimates (MLEs), given the class labels, with calculations of expected class labels (given these MLEs). Neerchal and Morel (2005) explicitly discuss both the Dirichlet multinomial and finite mixture models for handling overdispersion in categorical data, and they provide computational tools to estimate their models.

The Case of Heterogeneity in Both Participants and Items

If heterogeneity is detected for both participants and items, aggregation over either source is not a viable option. In this case, it is necessary to incorporate a specification of parameter variability into a model, since one cannot estimate a model from a sample consisting of a single response to a participant–item. As in the case of a single source of heterogeneity, one could proceed without further data exploration to specify random effects in both participants and items; but so far, there are only a few examples where variability in both participant and item parameters has been incorporated into a cognitive model—for example, Karabatsos and Batchelder (2003) and Rouder et al. (2007). These examples have used standard Bayesian hierarchical methods (e.g., Gelman et al., 2003; Gill, 2002), and Rouder and Lu (2005) provide a tutorial article on these methods written for cognitive modelers.

Another option is to engage in further data exploration before deciding on the best way to specify the random effects. One possibility for this sort of analysis is to use models in psychometric test theory. Unlike the case of cognitive modeling, variability in both participant and item parameters has been the standard situation in psychometric test theory for many years (e.g., Lord & Novick, 1968). The area of item response theory (IRT) consists of parametric and nonparametric models that allow heterogeneity in both participants and items (e.g., de Boeck & Wilson, 2004; Embretson & Reise, 2000; von Davier & Carstensen, 2007). In fact, the usual data structure for an IRT model is exactly the same as that for most cognitive models—namely, a participant $\times$ item matrix of categorical data, as described in Equations 2 and 9, where the response categories reflect various test item responses (e.g., correct/incorrect or multiple choice). IRT models, unlike cognitive models, specify individual differences in participant and item parameters as part of the base model; however, there are numerous examples of fixed effects and random effects assumptions on the participant and item parameters in the IRT literature. Generally, IRT models are very close to the basic sampling assumptions of the data structure involving versions of logit and log-linear models (e.g., Agresti, 2002). The complexity in IRT models is in their inference rather than their structure, because there is only one observation for each combination of participant and item. Because IRT models are developed for the data structure in Equation 9, these models may suggest statistically useful ways of modeling heterogeneity in both participants and items. It is beyond the scope of the present article to describe the IRT models in detail, and to show how they might be used to explore cognitive data exhibiting heterogeneity in both participants and items; but we conclude this section with some generalizations about the possible value of this approach.

It is our belief that cognitive modelers can gain in two ways by inspecting cognitive data with standard IRT models. First, hierarchical modeling is routine in the area of test theory, so it is a good place to study how the different approaches work in practice with relatively simple models such as the Dirichlet multinomial and finite mixture multinomial models, that have close ties to the nature of the data structure itself, for the single source of variability discussed earlier. In this way, one may be able to discover some of the properties of the overdispersion in the data before deciding how to incorporate hierarchical assumptions directly into a particular cognitive model.

A second way that analysis of cognitive data with an IRT model may prove productive is in selecting items that have the ability to discriminate the various cognitive processes postulated in a model. IRT models specify the performance of a participant to an item in terms of such factors as item difficulty and item discriminability, and these parameters function very much as the threshold and discriminability parameters, respectively, of psychometric functions (e.g., Klein, 2001; Kuss, Jäkel, & Wichmann, 2005). Suppose a model postulates a latent cognitive event that either does or does not occur during the manifest response to an item—for example, recognizing by familiarity or recollection, clustering related items, or guessing Category A instead of Category B. Further suppose that the data categories for a model can be partitioned into two sets: those that indicate the occurrence of the cognitive event, and those that do not. Then the responses to items can be scored dichotomously in terms of these two subsets of categories, and an analysis with an IRT model such as the two-parameter logistic (2-PL) model (e.g., Lord & Novick, 1968) can reveal which items should be used to determine whether or not the latent cognitive event in question occurred during the manifest response process.

In fact, such an approach might lead to more experimental designs in cognitive modeling where heterogeneous items rather than homogeneous items are used, and these designs coupled with appropriate hierarchical cognitive models might be useful in gaining sharper insights into underlying cognitive mechanisms.

GENERAL DISCUSSION

In this section, we will first discuss some of the consequences of relaxing two of the assumptions that were made for the analyses presented in the previous sections. The first concerns the possible impact of violation of independence (or exchangeability) in the observations in the participant–item data, and the second concerns the case where the data are of the continuous, not the categorical, type. In the final portion of this section, we will summarize our recommendations for how cognitive modelers should analyze categorical data prior to model analysis.

Violations of Independence

All the tests of homogeneity presented earlier examined whether or not certain statistics of the participants and/or items exhibited overdispersion relative to that expected by a baseline model based on the data sampling design. Overdispersion of a participant or item statistic relative to the baseline model was taken as evidence for heterogeneity. The tests all assumed that the participant–item response random variables are mutually stochastically independent, as represented in Equations 3 and 10. It turns out that violation of this assumption can result in either underdispersion or overdispersion of those statistics, even under the assumption that responses to participant–items are identically distributed. In this section, we will consider only violations of independence in the sequence of responses of a participant across items, because—barring collusion among participants—it is reasonable to assume that responses to an item, across participants, are independent.

Consider first the case of dichotomous data, where we developed tests for overdispersion in participants' row sums (see Equations 5, 6, and 7). As before, let

$$R_i = \sum_{j=1}^{M} X_{ij}$$

be a random variable for participant i's row sum. From a familiar result in probability theory (e.g., Hogg et al., 2005) we can write

$$\mathrm{Var}\left(R_i\right) = \sum_{j=1}^{M} \mathrm{Var}\left(X_{ij}\right) + 2\sum_{j=1}^{M}\sum_{j'=j+1}^{M} \mathrm{Cov}\left(X_{ij}, X_{ij'}\right), \quad (19)$$

where

$$\mathrm{Cov}\left(X_{ij}, X_{ij'}\right) = E\left(X_{ij} \cdot X_{ij'}\right) - E\left(X_{ij}\right)E\left(X_{ij'}\right).$$

If independence is assumed, each of the covariances is zero, but in general for dichotomous Bernoulli [Bin(1, p)] random variables

$$\mathrm{Cov}\left(X_{ij}, X_{ij'}\right) = \Pr\left(X_{ij} = 1, X_{ij'} = 1\right) - p_{ij}p_{ij'}, \quad (20)$$

and this quantity can be either positive or negative.

To simplify, suppose the X_{ij}s are identically distributed over both participants and items, as in hypothesis $\mathbf{H_1}$ in Equation 5, with $\Pr(X_{ij} = 1) = p$; $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, M$. Then it is easy to compute that, if independence is violated, Var(R) in Equation 19 could be either smaller or larger than the $\mathbf{H_1}$ target of $Mp(1 - p)$. In particular, negative covariances between the item responses within a participant will decrease the row variance, and positive covariances will inflate it. Either of these possibilities is reasonable in certain experimental contexts. For example, if a participant with capacity limitations on memory is only able to retain and recall a small number of items, correct recall of one item would tend to decrease the likelihood of recalling another item leading to negative covariances in Equation 20. This result could also follow from the phenomenon of output interference (e.g., Bjork, 1989). On the other hand, if retained items provide retrieval cues for other items, one might well find a positive rather than a negative covariance between items within a participant.

In some cases, it is possible to assess dependence in a participant–item array if the dependencies are systematic across participants over the item series. For example, if responses to items j and j' are negatively correlated across participants, this will be seen in the lack of independence in the 2 × 2 table of correct and incorrect responses to the two items over participants. Even within a participant, some types of correlations between items can be detected. For example, let us suppose that responses to items with adjacent serial positions are not independent. Define the alternation random variables for participant i

$$A_{ij} = \begin{cases} 1 \text{ if } X_{ij} \neq X_{i,j+1}, \\ 0 \text{ otherwise}, \end{cases} \quad (21)$$

for $j = 1, 2, \ldots, M - 1$.The effects of dependence between adjacent items will show up as a departure of the number of alternations,

$$A_i = \sum_{j=1}^{M-1} A_{ij},$$

in a sequence of item responses across the serial positions from that expected from a Bernoulli process. For example, suppose there are M items distributed as a Bernoulli process with $\Pr(X_{ij} = 1) = p_i$, for $j = 1, 2, \ldots, M$. Then $E(A_{ij}) = \Pr(A_{ij} = 1) = 2p_i(1 - p_i)$, so an expected baseline under independence is $E(A_i) = 2(M - 1)p_i(1 - p_i)$. Positive correlations between adjacent trials will decrease $E(A_i)$ over the baseline, and negative correlations will increase it.

Although certain types of violations of independence can be easily detected, others cannot. For example, suppose each participant has varying attention levels: high levels that support successful responding, and low levels that support errors. If the high attention states are located haphazardly both between and within participants, violations of independence will occur, but they will not be detectable by standard statistical tests. Unfortunately, such dependencies have the potential to either increase or decrease the mean of a statistic designed to assess over-

dispersion under an independence assumption. So far, our analysis of the effects of violation of independence has focused on the case of dichotomous data, but it is obvious that problems of the sort that we have discussed can occur for statistics for overdispersion in the case of data from multiple categories as well.

The Case of Continuous Type Data

All of the methods presented so far focus on the case of categorical data. The key idea was to test for overdispersion in the category frequency counts over participants or items. In essence, our tests concerned whether or not category frequency counts exhibited a higher variance across participants or items than that expected from a multinomial distribution in Equation 12, which is the natural data generating distribution in the case of i.i.d. observations that fall into a system of categories. In the case of continuous data, it is also possible to develop tests in the spirit of our approach to categorical data. However, as opposed to the categorical case, there is no single natural data-generating distribution, so the ability of a researcher to devise parametric tests of item or participant overdispersion depends on the data-generating distribution and the measure thought to be overdispersed.

For example, if the data are thought to be normally distributed, and if variability is a concern for the mean but not for the variance, a basic ANOVA may be applied where each participant is a level of the participant factor. In the case that the distribution is unknown and variability is suspected on only one dimension, basic nonparametric statistics may be applied. For example, concern for variability in the mean can be tested with a Kruskal–Wallis test (Kruskal & Wallis, 1952), whereas variability in the variance may be tested through an Ansari–Bradley test (Ansari & Bradley, 1960). These tests are available in many common statistical packages, including both R and SAS. In cases where the nature of dispersion is entirely unknown, a *k*-sample Kolmogorov–Smirnov test may be applied (e.g., Conover, 1965; Kiefer, 1959). This test considers differences in the empirical cumulative distribution functions (cdfs) across participants or items.

A limitation of the above tests is that they assume that variability is possible only from the tested source and that the other source (items or participants) is homogeneous. In cases of possible variability on both dimensions, we can use the permutation framework discussed earlier. The permutation test is based on a participant $\times$ item matrix of continuous type data $\mathbf{D} = (x_{ij})_{N\times M}$, where x_{ij} is the response of participant i to item j on some continuous measure. For the following discussion, we will test for overdispersion of the participant row means in the presence of possible overdispersion in the items; however, other statistics (e.g., the *k*-sample Kolmogorov–Smirnov) may also be applied to either the rows or columns. First, the mean is recorded for the complete data set and for each individual row. A difference measure is then calculated between the overall mean and individual means. For example, one could take the sum of squared difference between the overall mean and each row mean. Next, a Monte Carlo sample of permutations of the same type as described previously and in the Appendix can be simulated for the continuous data matrix, **D**. For each permuted data set, the above process is repeated where the overall mean is calculated and compared with the row means. The observed measure of deviation between the rows and overall data set is then compared with the distribution of deviations found for the permuted data sets. If the value for the observed data set is large relative to the permuted data set, this indicates that the data are overdispersed on the measure of interest.

Recommendations and Conclusions

The goals of this article have been to outline a series of simple tests to examine whether or not heterogeneity is present in categorical data from participants and/or items, and, in addition, to introduce the reader to several basic hierarchical models for modeling the identified variability. The chi-square test, as described in Equations 7 and 14, is suggested when variability is a concern on participants (items) but not items (participants). For cases where variability is possible on both items and participants, we recommend a permutation-type test in order to take into account the fact that variability from one source (e.g., participants) may shrink variability in the source of interest (e.g., items). The test is described in the first half of the article, and the Appendix supplies supporting MATLAB code for sampling permutations.

If observations over participants and items are independent, and the cognitive model of interest is fully identified, variability at the level of the category probabilities guarantees variability on one or more model parameters. Similarly, given sufficient power, a failure to identify variability in the category counts provides evidence for homogeneity across participants and items of all the parameters of the cognitive model. These results follow directly from the fact that an identified model of categorical data has a one-to-one correspondence between model parameters and possible category probability distributions, so model parameters may vary across participants (or items) if and only if category probabilities also vary across participants (or items).

Given the above result, if participant and/or item heterogeneity is identified at the level of the data, researchers must ensure that they take into account this variability when analyzing their cognitive model. In some cases, participant and/or item differences can be fit through the addition of a hierarchical level to the model. For more details on hierarchical modeling, refer to the discussion in the introduction along with relevant citations, and to the examples presented in the second half of this article. If hierarchical modeling is not possible due to the complexity of the nature of its specifications, we suggest that, at least, the behavior of the model under parameter heterogeneity should be tested through simulation.

Recently, there have been a number of developments in computational statistics that make it relatively easy to analyze hierarchical versions of parametric models. For example, programs such as WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) can estimate complex hierarchical models with limited time and effort. In light of these developments, one might argue that cognitive

models should always be equipped with random effects when they are used to analyze data. However, researchers should be cautious in applying hierarchical modeling if no variability is found in the category probabilities. It is necessarily the case that a hierarchical version of a parametric model has a greater capacity to overfit data than does the base model with no random effects. For example, if the variability of category probabilities obtained by fitting a hierarchical version of a cognitive model is large, but the tests we have proposed for variability at the level of the data nevertheless fail to reject participant and item homogeneity, this suggests that the base model is misspecified, and that the hierarchical version is simply overfitting the data.

Our recommendation is that categorical data should always be tested for participant and item heterogeneity, and that the results of these tests should be taken into consideration in a researcher's subsequent analysis of the data. Given the wide range of tools currently available in computational statistics to handle individual and item differences, there are few remaining excuses for cognitive psychologists to ignore these developments in analyzing their cognitive models.

AUTHOR NOTE

Work on this article was supported by two grants from the National Science Foundation: SES-0136115 to A. K. Romney and W.H.B. (Co-PIs) and SES-0616657 to X. Hu and W.H.B. (Co-PIs). In addition, we acknowledge the support from the Department of Cognitive Sciences and the Institute for Mathematical Behavioral Sciences for summer fellowship assistance to J.B.S. We thank David Riefer and the referees for helpful comments on the manuscript. Correspondence concerning this article should be addressed to W. H. Batchelder, Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100 (e-mail: whbatche@uci.edu).

REFERENCES

AGRESTI, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, **7**, 131-177.

AGRESTI, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

AGRESTI, A., CAFFO, B., & OHMAN-STRICKLAND, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, **47**, 639-653.

ALBERT, J. H. (1999). Criticism of a hierarchical model using Bayes factors. *Statistics in Medicine*, **18**, 287-305.

ALBERT, J. [H.], & CHIB, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, **92**, 916-925.

ANSARI, A. R., & BRADLEY, R. A. (1960). Rank sum tests for dispersion. *Annals of Mathematical Statistics*, **31**, 1174-1189.

ASHBY, F. G., MADDOX, W. T., & LEE, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity–choice model. *Psychological Science*, **5**, 144-151.

BATCHELDER, W. H. (1975). Individual differences and the all-or-none vs incremental learning controversy. *Journal of Mathematical Psychology*, **12**, 53-74.

BATCHELDER, W. H., CHOSAK-REITER, J., SHANKLE, W. R., & DICK, M. B. (1997). A multinomial modeling analysis of memory deficits in Alzheimer's disease and vascular dementia. *Journals of Gerontology*, **52B**, P206-P215.

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

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

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

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

BJORK, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309-330). Hillsdale, NJ: Erlbaum.

CLARK, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, **12**, 335-359.

CONGDON, P. (2005). *Bayesian models for categorical data*. New York: Wiley.

CONOVER, W. J. (1965). Several *k*-sample Kolmogorov–Smirnov tests. *Annals of Mathematical Statistics*, **36**, 1019-1026.

CURRAN, T., & HINTZMAN, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 531-547.

DE BOECK, P., & WILSON, M. (EDS.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

DECARLO, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, **109**, 710-721.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1-38.

EFRON, B., & TIBSHIRANI, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

EMBRETSON, S. E., & REISE, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

ESTES, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134-140.

EVANS, M., HASTINGS, N., & PEACOCK, J. B. (2000). *Statistical distributions* (3rd ed.). New York: Wiley.

GARREN, S. T., SMITH, R. L., & PIEGORSCH, W. W. (2001). Bootstrap goodness-of-fit test for the beta-binomial model. *Journal of Applied Statistics*, **28**, 561-571.

GELMAN, A., CARLIN, J. B., STERN, H. S., & RUBIN, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

GILDEN, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review*, **108**, 33-56.

GILL, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. New York: Chapman & Hall.

GRIFFITHS, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, **29**, 637-648.

HAIDER, H., & FRENSCH, P. A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 392-406.

HAYS, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

HEATHCOTE, A., BROWN, S., & MEWHORT, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, **7**, 185-207.

HINTZMAN, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, **87**, 398-410.

HINTZMAN, D. L. (1993). On variability, Simpson's paradox, and the relation between recognition and recall: Reply to Tulving and Flexser. *Psychological Review*, **100**, 143-148.

HOGG, R. V., MCKEAN, J. W., & CRAIG, A. T. (2005). *Introduction to mathematical statistics* (6th ed.). Upper Saddle River, NJ: Pearson.

HOWARD, M. W., & KAHANA, M. J. (2002). A distributed representa-

tion of temporal context. *Journal of Mathematical Psychology*, **46**, 269-299.

JONES, M., LOVE, B. C., & MADDOX, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 316-332.

KARABATSOS, G., & BATCHELDER, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika*, **68**, 373-389.

KARPIUK, P., JR., LACOUTURE, Y., & MARLEY, A. A. J. (1997). A limited capacity, wave equality, random walk model of absolute identification. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 279-299). Mahwah, NJ: Erlbaum.

KIEFER, J. (1959). *K*-sample analogues of the Kolmogorov–Smirnov and Cramér–von Mises tests. *Annals of Mathematical Statistics*, **30**, 420-447.

KIM, B. S., & MARGOLIN, B. H. (1992). Testing goodness of fit of a multinomial model against overdispersed alternatives. *Biometrics*, **48**, 711-719.

KLAUER, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, **71**, 7-31.

KLEIN, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, **63**, 1421-1455.

KRUSKAL, W. H., & WALLIS, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583-621.

KUSS, M., JÄKEL, F., & WICHMANN, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, **5**, 478-492.

LEE, M. D., & WEBB, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, **12**, 605-621.

LEHMANN, E. L., & ROMANO, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.

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

MADDEN, L. V., & HUGHES, G. (1994). BBD—Computer software for fitting the beta-binomial distribution to disease incidence data. *Plant Disease*, **78**, 536-540.

MCLACHLAN, G., & PEEL, D. (2000). *Finite mixture models*. New York: Wiley.

MINKA, T. (2000). *Estimating a Dirichlet distribution* (Tech. Rep.). Cambridge, MA: MIT. Available at research.microsoft.com/~minka/papers/dirichlet/.

MOORE, D. S., & MCCABE, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: Freeman.

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, **49**, 65-82.

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

NAVARRO, D. J., GRIFFITHS, T. L., STEYVERS, M., & LEE, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, **50**, 101-122.

NEERCHAL, N. K., & MOREL, J. G. (2005). An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. *Computational Statistics & Data Analysis*, **49**, 33-43.

RASCH, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

R DEVELOPMENT CORE TEAM (2005). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

RIEFER, D. M., KEVARI, M. K., & KRAMER, D. L. F. (1995). Name that tune: Eliciting the tip-of-the-tongue experience using auditory stimuli. *Psychological Reports*, **77**, 1379-1390.

RIEFER, D. M., KNAPP, B. R., BATCHELDER, W. H., BAMBER, D., & MANIFOLD, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, **14**, 184-201.

ROUDER, J. N., & LU, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, **12**, 573-604.

ROUDER, J. N., LU, J., SUN, D., SPECKMAN, P. [L.], MOREY, R. [D.], & NAVEH-BENJAMIN, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, **72**, 621-642.

ROUDER, J. N., SUN, D., SPECKMAN, P. L., LU, J., & ZHOU, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, **68**, 589-606.

SPIEGELHALTER, D., THOMAS, A., BEST, N., & LUNN, D. (2003). *WinBUGS user manual version 1.4*. Cambridge: MRC Biostatistics Unit.

THORNTON, T. L., & GILDEN, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review*, **12**, 409-441.

TITTERINGTON, D. M., SMITH, A. F. M., & MAKOV, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

VON DAVIER, M., & CARSTENSEN, C. H. (EDS.) (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

WAGENMAKERS, E.-J., FARRELL, S., & RATCLIFF, R. (2004). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin & Review*, **11**, 579-615.

APPENDIX

Below is sample code for the permutation test on participants for both MATLAB and R. Note that the code below provides just one of many ways to implement the permutation test.

MATLAB Code for Generating Permuted Data Matrices

```
reps=10000 %number of permutations
subjs=29 %number of participants in data set
items=50 %number of items in data set
cate=4 %number of response categories
permuteddata =zeros(subjs, items); %initiate variable to hold permuted
  %matrices
contingency=zeros(subjs, cate);
%initiate variable to hold contingency tables from permuted data sets
for i=1:reps
  for ii=1:subjs %randomly permutes each row of data matrix D
    permuteddata(ii,:)=datamatrix(ii,randperm(items));
  end
  for cc=1:cate %converts data matrix to Nx2 contingency table
    contingency(:,cc)=sum(permuteddata==cc)';
  end
  %the statistical test of interest
end
```

APPENDIX (Continued)

R Code for Generating Permuted Data Matrices

```
# The R code below tests for subject variability

reps<-10000 #number of permutations
subjs<-29 #number of participants in data set
items<-50 #number of items in data set
cate<-4 #number of response categories
permutedata <-matrix(0,subjs,items)
contingency<-matrix(0,cate, subjs)
#initiate variable to hold contingency tables from permuted data sets
for (i in 1:reps){
  for (ii in 1:items) { #randomly permutes each row of data matrix D
    permutedata[,ii]<-datamatrix[sample(subjs,subjs),ii]
  }
  for (cc in 1:cate) { #converts data matrix to Nx2 contingency table
    contingency[cc,]<-apply(permutedata==cc,1,sum)
  #calculated permuted contingency table
  }
  #any test of interest can then be run on the permuted contingency table
}
```

Study of the Statistical Power of the Permutation Test

Although it is beyond the scope of this article to provide a thorough analysis of the power of the permutation test, a series of simulations were performed to examine the relative efficiency of the permutation test in comparison with the standard chi-square test. For the purpose of this study, simulated data sets were drawn from a beta-binomial distribution (see Equation 16 and surrounding text for details on the beta-binomial). We considered 5 levels of participant variability for each of three different sample sizes. The resulting 15 simulations contained 1,000 data sets each. All simulated data sets were tested for participant variability using both the chi-square and permutation test at an alpha level of .05. The proportion of times that the permutation and chi-square test identified participant heterogeneity are presented in Table A1. Note that in all cases the permutation test and chi-square test provide similar power. Also, in the case of 80 participants and 80 items both tests are able to detect all simulated levels of variability with high accuracy. Finally, the false rejection rate matches the set alpha level of .05 in all cases. A larger simulation study may be able to detect errors in the false rejection rate, but at least in this case the biases are almost certainly small if present. Taken together, the above results provide evidence that the permutation test performs at least as well as the standard chi-square test in detecting participant heterogeneity in the absence of item variability.

Table A1
Estimated Power of the Permutation and Chi-Square Tests

	Number of Participants/Items		
	80	30	10
Permutation Test			
$\alpha = 0.2, \beta = 0.2$	1.00	1.00	1.00
$\alpha = 2, \beta = 2$	1.00	1.00	.83
$\alpha = 10, \beta = 10$	1.00	.96	.23
$\alpha = 20, \beta = 20$	1.00	.73	.12
$\alpha = 40, \beta = 40$	.99	.36	.08
No variability	.05	.05	.05
Chi-Square Test			
$\alpha = 0.2, \beta = 0.2$	1.00	1.00	1.00
$\alpha = 2, \beta = 2$	1.00	1.00	.82
$\alpha = 10, \beta = 10$	1.00	.96	.24
$\alpha = 20, \beta = 20$	1.00	.67	.13
$\alpha = 40, \beta = 40$	.99	.33	.09
No variability	.05	.05	.05

Notes—α and β are the parameters used for the simulating beta-binomial distribution. For the no-variability case, a binomial distribution was used, with $p = .5$.

(Manuscript received June 14, 2007;
revision accepted for publication December 26, 2007.)