Running head: HIERARCHICAL MULTINOMIAL MODELS

Hierarchical Multinomial Processing Tree Models: A Latent-Class Approach

Karl Christoph Klauer

Albert-Ludwigs-Universität Freiburg

PSYCHOMETRIKA, in press

Address information:

- Institut für Psychologie, Universität Freiburg, D-79085 Freiburg, Germany.
- E-mail: christoph.klauer@psychologie.uni-freiburg.de;
- Tel.: +49 761 2032469 (work); +49 761 5573917 (home);
- Fax: +49 761 2032417.

Abstract

Multinomial processing tree models are widely used in many areas of psychology. Their application relies on the assumption of parameter homogeneity, that is, on the assumption that participants do not differ in their parameter values. Tests for parameter homogeneity are proposed that can be routinely used as part of multinomial model analyses to defend the assumption. If parameter homogeneity is found to be violated, a new family of models, termed latent-class multinomial processing tree models, can be applied that accomodates parameter heterogeneity and correlated parameters, yet preserves most of the advantages of the traditional multinomial method. Estimation, goodness-of-fit tests and tests of other hypotheses of interest are considered for the new family of models.

In the last two decades, multinomial processing tree models, henceforth referred to as multinomial models, have been extensively used in many areas of psychology; an overview is given by Batchelder and Riefer (1999). They are usually tailored to a particular experimental task and can be used to test assumptions about the psychological processes that contribute to task performance and to assess the relative contribution of each process in a principled manner. Their increasing popularity is due to a number of advantages: Assumptions about psychological processes in a given experimental paradigm can often be cast into the form of a processing tree model in a natural manner. In addition, parameter estimation and hypotheses tests can be conducted by means of relatively straightforward maximum-likelihood techniques (Hu & Batchelder, 1994). Last but not least, multinomial models are often found to describe empirical data well.

To introduce the question of the present paper, consider a minimal multinomial model of the standard recognition task. A list of items is presented, and in a subsequent recognition test, these items are shown along with new items. For each item, participants are asked to decide whether the item is an old one, that was previously seen, or a new one. Ignoring the possibility of guessing, assume that participant t correctly recognizes an old item as old with probability D_t, and that the probability of n_t correct "old-"responses is given by the binomial distribution with parameters D_t and N, the fixed number of old items presented for recognition. In multinomial modeling, data from several, say T, participants are usually obtained, and the assumption of parameter homogeneity is made. That is, it is assumed that the different memory parameters are equal:
$D_1 = D_2 = \ldots = D_T =: \mu'$. This implies that the sum score $n_+ = \sum_{t=1}^{T} n_t$ also follows a binomial distribution with expected value $TN\mu'$ and variance $TN\mu'(1-\mu')$, defining a minimal multinomial model.

The present paper deals with the possibility that there is parameter heterogeneity, that is, in the present example that there are individual differences

in memory performance. What are the consequences of such differences for multinomial model analyses?

Following previous work on parameter heterogeneity (e.g., Riefer & Batchelder, 1991), assume that the participants are sampled from a population of potential participants in which the individual $D-$parameters are distributed according to a beta distribution with parameters α and β, both of them real numbers larger than zero. The beta distribution is a family of distributions on the interval (0,1) that is relatively tractable and accomodates a wide range of different distributions. Under this assumption, the so-called beta-binomial distribution, rather than the binomial distribution, governs the number of "old"-responses produced by a participant in the sample (Johnson, Kotz, & Kemp, 1993; chap. 2). It is instructive to reparameterize the beta distribution and the beta-binomial distribution by means of the two parameters μ and γ as follows: $\mu = \frac{\alpha}{\alpha+\beta}$, $\gamma = \frac{1}{1+\alpha+\beta}$. Both parameters take on values between zero and one. The parameter μ is the expected value of the beta distribution, that is, the population mean of parameter D. The variance of D is $\mu(1-\mu)\gamma$. The parameter γ quantifies the extent of heterogeneity.

In multinomial model analyses, possible individual differences in D are ignored, and it is assumed that the aggregated data n_+ follow a binomial distribution with parameters TN and μ'. Under this assumption, the maximum likelihood estimate of μ' is $\hat{D} = \frac{1}{TN} n_+$. If parameter homogeneity is violated, $\hat{D}$ is nevertheless an unbiased and consistent estimator of μ, the population average of D. Erdfelder (2000; chap. 5) has shown that parameter estimates will be consistent estimates of the population averages of the individuals' parameters for a certain subclass of multinomial models that he calls aggregation-invariant multinomial processing tree models, if, in addition, the model parameters are not correlated over persons. For more complex models, parameter estimation will in general be biased as a consequence of parameter heterogeneity so that maximum likelihood

estimates diverge systematically from the population averages of the individuals' parameters (Riefer & Batchelder, 1991); see the more formal analysis of the consequences of parameter heterogeneity presented in Appendix A.

For the time being, consider the variance of the estimate $\hat{D}$. Given parameter homogeneity, it is $\frac{1}{TN}\mu(1-\mu)$, estimated by $\frac{1}{TN}\hat{D}(1-\hat{D})$, because n_+ follows a binomial distribution. In contrast, under parameter heterogeneity, the variance is $\frac{1}{TN}\mu(1-\mu)\{1+(N-1)\gamma\}$. The actual variance is thus underestimated by a factor of $1+(N-1)\gamma$ when parameter heterogeneity is ignored. Given parameter heterogeneity, it is larger, by a factor of maximally N, than is to be expected on the basis of the binomial distribution.

If this overdispersion is ignored, and the too small binomial-distribution variance is used, standard errors and confidence intervals of model parameters will be estimated too small, and significance tests for differences between parameter values and aggregated frequencies will exhibit $\alpha-$errors above the nominal significance level as a consequence. These problems increase in severity as the level of heterogeneity (γ) increases and the numbers of data points (N) collected per participant increase as is readily apparent from the above example (the actual variance of the parameter estimate is underestimated by a factor of $1+(N-1)\gamma$). In the general case, such problems will be further exacerbated as the number of participants is increased (see Appendix A).

Furthermore, goodness-of-fit tests of multinomial models can be seen as tests of equality restrictions imposed upon the parameters of saturated models (Batchelder & Riefer, 1999). Given parameter heterogeneity, such goodness-of-fit tests will therefore also exhibit inflated levels of $\alpha-$errors and will often lead to the rejection of simple models, even if these adequately describe each individual's data, especially in cases where the number of data points collected per person is relatively large. In applications, this outcome often prompts researchers to work with more complex models involving more parameters to fit the aggregated data.

Not unfrequently, a saturated model is even used that describes the aggregated data perfectly. This strategy ensures that parameter heterogeneity will remain undetected and has the potential to distort the substantive conclusions drawn from subsequent analyses based on the fitting model.

In what follows, an extension of the multinomial method termed *latent-class* multinomial processing tree models will be discussed. Tests for parameter homogeneity will be derived that can be routinely applied in the course of traditional multinomial model analyses and that allow researchers to defend the assumption of parameter homogeneity. If parameter homogeneity is violated on the other hand, the latent-class extension provides a tractable method of applying multinomial models in a way that accomodates parameter heterogeneity, including correlated parameters, yet preserves most of the advantages of the multinomial modeling technique. Whereas the present paper is focused on interindividual differences between participants, items as a source of variability and dependencies are considered in the General Discussion. Before moving on to these topics, let us briefly turn to the so-called pair-clustering model that is used as a running example in this paper.

Example 1: The Pair-Clustering Model

The pair-clustering model is one of the best analyzed members of the family of multinomial models (e.g., Batchelder & Riefer, 1986, 1999; Riefer & Batchelder, 1991). It is based on a free-recall task in which participants are presented with a list of words that are related by categories. The items consist of several categorically related word pairs (e.g., oxygen and hydrogen), plus a number of singleton words. Word pairs and singletons are presented one word at a time, and participants are later asked to recall the list items in any order.

The recall events are scored into mutually exclusive response categories. For the word pairs, four categories, C_{11}, C_{12}, C_{13}, and C_{14}, are distinguished:

- C_{11}: Both words are recalled adjacently,
- C_{12}: both words are recalled, but not adjacently,
- C_{13}: only one word in the pair is recalled, and
- C_{14}: neither word in the pair is recalled.

The recall of singletons is scored into two categories C_{21}: "The singleton is recalled" and C_{22}: "The singleton is not recalled". The data are the counts n_{kj} with which each response category C_{kj} is observed, aggregated over participants and the N_1 word pairs and N_2 singletons in the list.

The model is based on four parameters that are the probabilities of storing a word pair as a cluster (c), the probability of a successful retrieval of a stored cluster (r), the probability of successful retrieval of a member of a word pair not stored as a cluster (u) and the probability of the successful retrieval of a singleton (a). Figure 1 shows the processing tree representation of the model. Word pairs are stored as a cluster with probability c. A stored cluster is retrieved with probability r in which case both words are recalled adjacently (response category C_{11}). If a stored cluster cannot be retrieved with probability $1 - r$, neither word of the word pair is retrieved (response category C_{14}). Thus, it is assumed that clustered items are accessible either as a word pair or not at all.

The model equations are:

$$\begin{aligned} p(C_{11}) &= cr \\ p(C_{12}) &= (1-c)u^2 \\ p(C_{13}) &= (1-c)2u(1-u) \\ p(C_{14}) &= c(1-r) + (1-c)(1-u)^2 \\ p(C_{21}) &= a \\ p(C_{22}) &= 1-a \end{aligned} \tag{1}$$

Frequently, a restricted model with $u = a$ is used.

Latent-Class Multinomial Processing Tree Models

A natural approach to accomodate parameter heterogeneity is to consider the model parameters as random rather than fixed effects. For this purpose, a *core* multinomial model will be defined as the model that describes a given participant's data with potentially different parameters for each person. This can be extended to a *hierarchical* multinomial model by specifying a distribution of the parameters to model parameter heterogeneity (Raudenbush & Bryk, 2002).

Multinomial processing tree models are models for response frequencies of pre-defined mutually exclusive response categories. In most cases, there are several independent category systems. For example, in the pair-clustering paradigm, responses are observed to two kinds of items, word pairs and singletons, that are scored into two category systems with four and two response categories, respectively. The separate category systems are modelled by separate subtrees of the multinomial model. Person t contributes frequency counts n_{kjt}, where $k = 1, \ldots, K$ runs over category systems, or equivalently subtrees, and $j = 1, \ldots, J_k$ runs over the J_k response categories of category system k. For each category system k, these frequencies are assumed to follow a multinomial distribution with parameters p_{kjt}, $j = 1, \ldots, J_k$, and N_k, the fixed number of responses obtained per person t and category system k.

A multinomial processing tree model consists of a description of the category probabilities p_{kjt} by means of S functionally independent parameters θ_{st}, $s = 1, \ldots, S$ (each θ_{st} being free to vary in $(0, 1)$): $p_{kjt} = p_{kj}(\boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_t$ is the vector of the S parameter values by person t. The functions $p_{kj}(\boldsymbol{\theta})$ have a simple form (e.g., Equation 1) that permits the application of a simple EM-algorithm for maximum-likelihood estimation of the model parameters (Hu & Batchelder, 1994). In the traditional analysis, it is assumed that the $\boldsymbol{\theta}_t$ are equal over persons, and the response-category frequencies aggregated over persons are then sufficient statistics

for the model parameters.

Allowing for different parameters for each person t, the vector of person-wise category counts $\mathbf{n}_t = (n_{11t}, \ldots, n_{1J_1t}, \ldots, n_{K1t}, \ldots, n_{KJ_Kt})'$ is still modelled by a vector-valued random variable $\mathbf{N}$ that follows a product-multinomial distribution:

$$P(\mathbf{N} = \mathbf{n}_t|\boldsymbol{\theta}_t) = \prod_{k=1}^{K} \left\{ \binom{N_k}{n_{k1t} \ldots n_{kJ_kt}} \prod_{j=1}^{J_k} [p_{kj}(\boldsymbol{\theta}_t)]^{n_{kjt}} \right\}. \tag{2}$$

The model from Equation 2 defines the core model. For each category system k, one of the n_{kjt} can be computed from the others, because their sum is fixed to N_k. Therefore, there are $J^* = \sum_k (J_k - 1)$ non-redundant category counts.

If the model parameters are distributed according to a distribution with probability measure μ, the responses of a randomly sampled participant are distributed according to:

$$P(\mathbf{N} = \mathbf{n}) = \int P(\mathbf{N} = \mathbf{n}|\boldsymbol{\eta}) d\mu(\boldsymbol{\eta}), \tag{3}$$

where $P(\mathbf{N} = \mathbf{n}|\boldsymbol{\eta})$ is given by the right side of Equation 2, in which the fixed values $\boldsymbol{\theta}_t$ are replaced by the variable of integration, $\boldsymbol{\eta}$, and $\mathbf{n}_t$ is replaced by $\mathbf{n}$.

A sample of T participants is described by T independent, identically distributed random variables $\mathbf{N}_t$, $t = 1, \ldots, T$, and the entire data set, $(\mathbf{n}_1, \ldots, \mathbf{n}_T)$ is thus modelled by

$$P\left((\mathbf{N}_1, \ldots, \mathbf{N}_T) = (\mathbf{n}_1, \ldots, \mathbf{n}_T)\right) = \prod_{t=1}^{T} \left\{ \int P(\mathbf{N}_t = \mathbf{n}_t|\boldsymbol{\eta}) d\mu(\boldsymbol{\eta}) \right\}. \tag{4}$$

In this framework, we define *latent-class* multinomial models by probability measures μ that distribute their probability mass over a finite number C of fixed parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C$. If $\lambda_c = \mu(\{\boldsymbol{\theta}_c\})$ is the size of class c, the model equation simplifies to:

$$P\left((\mathbf{N}_1, \ldots, \mathbf{N}_T) = (\mathbf{n}_1, \ldots, \mathbf{n}_T)\right) = \prod_{t=1}^{T} \left\{ \sum_{c=1}^{C} \lambda_c P(\mathbf{N}_t = \mathbf{n}_t|\boldsymbol{\theta}_c) \right\}. \tag{5}$$

This means that each participant is assumed to fall into one of C latent classes of proportional sizes λ_c. Within each class, parameter homogeneity holds, but across classes there can be parameter variability and correlations between parameters. In comparison to other kinds of distributions μ, the latent-class assumption leads to a relatively tractable class of models, as will be seen. In addition, other kinds of distributions can be approximated arbitrarily well by the latent-class assumption as the number of classes is increased. In practice, however, identifiability and interpretability concerns will limit the number of classes to relatively small numbers (see next section). As with any family of models, identifiability, estimation, goodness-of-fit tests and other hypotheses tests are areas of major interest, and we will consider these in turn.

Identifiability

The parameters ξ of a *latent-class* multinomial model with C classes are the parameters λ_c for the class sizes and the vectors $\boldsymbol{\theta}_c$ of core-model parameters for each class c. Latent-class multinomial models can be considered mixtures of product-multinomial distributions and as such, their parameters can be identified only up to permutations of the classes (Titterington, Smith, & Makov, 1985; chap. 3).

It is easy to show that a necessary condition for identifiability is that the core model is identified at least locally at each $\boldsymbol{\theta}_c$. In virtually all applications, globally identified multinomial models are used, and in the sequel, it is assumed that the core model is globally identified.

In a latent-class multinomial model, the category counts jointly follow a mixture of product-multinomial distributions, and each category count considered individually follows a mixture of binomial distributions. Furthermore, it is well known that mixtures of binomial distributions with parameters p_c and N and mixture coefficients λ_c are identified if and only if $N \geq 2C - 1$ (Titterington et al. 1985; chap. 3). Thus, if there are at least $2C - 1$ observations per person and

category system k, that is, if $N_k \geq 2C - 1$ for all k, the C mixture coefficients λ_c and the class-wise category probabilities $p_{kjc} = p_{kj}(\boldsymbol{\theta}_c)$ can be expected to be identified. It follows that if the core model is globablly identified, the core-model parameters $\boldsymbol{\theta}_c$ are identified for each class c. For example, if all N_k are larger than two, models with $C = 2$ and globally identified core model will be identified; with all N_k larger than four, models with three classes can be expected to be identified. In applications, non-identifiability is signalled by a singular Fisher information matrix.

Estimation

A simple EM-algorithm can be devised for the maximum-likelihood estimation of *latent-class* multinomial models, using principles proposed by Dempster, Laird, and Rubin (1977, Section 4.3). Let $\boldsymbol{\theta}_c^{(m)}$ and $\lambda_c^{(m)}$ be the current estimates of the model parameters. First, the posterior probability, ω_{tc}, of class c given participant t's data, is computed for each class and participant:

$$\omega_{tc} = \frac{\lambda_c^{(m)} P(\mathbf{N} = \mathbf{n}_t \mid \boldsymbol{\theta}_c^{(m)})}{\sum_{d=1}^{D} \lambda_d^{(m)} P(\mathbf{N} = \mathbf{n}_t \mid \boldsymbol{\theta}_d^{(m)})}. \tag{6}$$

New estimates for the class sizes are then computed as the averages of the ω_{tc},

$$\lambda_c^{(m+1)} = \frac{1}{T} \sum_{t=1}^{T} \omega_{tc}, \tag{7}$$

along with estimates of the expected missing class-wise frequency counts m_{kjc},

$$m_{kjc} = \sum_{t=1}^{T} \omega_{tc} n_{kjt}, \tag{8}$$

for $k = 1, \ldots, K$, $j = 1, \ldots, J_k$, and $c = 1, \ldots, C$. The new estimates of the parameters $\boldsymbol{\theta}_c^{(m+1)}$ are then obtained on the basis of these class-wise frequency counts. For this purpose, the m_{kjc} are entered as data into a traditional multinomial-model analysis of the core model applied to C independent groups that correspond to the C classes, with different model parameters in each group (e.g., Hu & Batchelder, 1994). The new estimates are simply the traditional

maximum-likelihood estimates that result from this analysis. Although the m_{kjc} are real numbers rather than integer counts, the EM-algorithm that is used in the traditional analysis (Hu & Batchelder, 1994) can be applied without modifications.[1]

Note for later reference that the EM-algorithm just described is applicable even if certain core-model parameters θ_{sc} are constrained to be equal across classes c. In applications, the EM-algorithm is often found to be slow in converging, and the present algorithm makes no exception. In the analyses discussed below, we found that a three-step procedure converged relatively fast: Starting from random initial parameter values, a couple of EM-iterations are computed. The preliminary parameter estimates that emerge are used as the inital values for a conjugate-gradient algorithm (Powell, 1977) maximizing the likelihood of the data with a relatively liberal stopping criterion. The resulting parameter estimates are again subjected to the EM-algorithm for final, polishing iterations. The entire procedure is repeated several times with different random starts to minimize the danger of running into local minima.

Note finally that for $C = 1$, the model reduces to a normal multinomial model, that is, to the core model itself. In this case, the category counts, aggregated over persons, are sufficient statistics, and parameter estimation can can proceed on the basis of the aggregated cell counts using the traditional method (Hu & Batchelder, 1994).

Goodness-of-Fit Tests

In multinomial modeling, goodness of fit is most often assessed by means of a log-likelihood ratio statistic in which the model under study is compared to a saturated model that describes the aggregated data perfectly. The saturated model is usually the unrestricted product-multinomial model or a reparametrization thereof, but for some models (such as the pair-clustering model), the saturated model still imposes inequality restrictions on the expected aggregated frequency counts that can be violated, but are then difficult to evaluate for significance. For

latent-class multinomial models, a saturated model is given by a multinomial over all the frequency patterns $\mathbf{n}$ that a participant can produce, but since at most T of the many possible frequency patterns have non-zero counts in real data sets, it makes little sense to fit this model as a baseline in goodness-of-fit tests reyling on asymptotic approximations.

Borrowing techniques from structural equation modeling (e.g., Browne, 1984; Muthén, 1993; Satorra, 1992), we propose to test latent-class multinomial models by assessing their capability to describe the first and second moments of the data. This amounts to testing the fit to the means, or equivalently, to the aggregated category counts as is done in the traditional multinomial approach (first moments) as well as to the variances and covariances of the person-wise category counts (second moments). Test statistics are consequently proposed for testing mean structure as well as variance-covariance structure. The latter test whether the observed variances and covariances of the person-wise category counts are accounted for by the model. Parameter heterogeneity and correlated parameters give rise to variances and correlations of category counts that differ from the expectations derived from traditional multinomial models with $C = 1$. The variance-covariance structure tests are therefore highly sensitive for deviations from parameter homogeneity when applied for the one-class model, whereas they test whether the observed extent of heterogeneity is well described when applied for latent-class multinomial models with $C > 1$.

Three test statistics, termed M_1, M_2, and M_3, are considered for mean structure testing, and two test statistics, termed S_1 and S_2, for variance-covariance structure testing. Consider first mean structure test M_3. M_3 adopts the log-likelihood ratio logic of the traditional multinomial modeling approach. Specifically M_3 compares the log-likelihood l_1 of the latent-class multinomial model of interest (computed as the logarithm of the probability given in Equation 5), evaluated at the maximum likelihood estimates, with the log-likelihood l_0 of a

latent-class model with the same number of classes and a saturated core model, evaluated at the maximum likelihood estimates obtained for the resulting model. Thus,

$$M_3 = -2(l_1 - l_0) \tag{9}$$

For $C = 1$, this is simply the above-mentioned log-likelihood ratio test of the traditional approach. Because the saturated model describes the mean category counts perfectly (setting aside possible inequality restrictions), the test basically tests whether the restricted model still describes the mean category counts adequately. Similarly, for $C > 1$, it is easy to see that the latent-class multinomial model with saturated core model describes the mean category counts perfectly, because the saturated core model can perfectly accomodate the missing class-wise category counts for each class separately. Consequently, the log-likelihood ratio statistic is likely to be highly sensitive to any deviations of the model predictions from the mean category counts that arise when the restricted model is used as core model. For this reason, we classify M_3 as a mean structure test although it will also show some sensitivity for detecting other kinds of deviations from the restricted latent-class multinomial model. M_3 is asymptotically distributed as χ^2 with $q_0 - q_1$ degrees of freedom where q_0 and q_1 are the numbers of functionally independent parameters of the two models compared (see Appendix C). If the model under study is itself based on the saturated core model, M_3 is equal to zero, reflecting the fact that such a model should perfectly describe the mean category counts.

The remaining statistics, M_1, M_2, S_1, and S_2 are based on direct comparisons of the observed means and variances and covariances with the model predictions for these moments. Let $\mathbf{m}$ be the vector of mean observed category counts with elements m_{kj} stacked one above the other. For each category system k, one of the m_{kj} can be computed from the others, because their sum is fixed to N_k, and we leave out these redundant means, so that $\mathbf{m}$ has $J^* = \sum_k (J_k - 1)$ elements.

Similarly, let $\mathbf{s}$ be the vector of observed variances and covariances between the person-wise category counts, in which only variances and covariances involving the J^* non-redundant categories are considered and only the lower left triangle of the resulting variance-covariance matrix is used. Thus, $\mathbf{s}$ is a vector of $\frac{1}{2}J^*(J^*+1)$ elements. Let ξ be the vector of the q functionally independent model parameters $\xi = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_C, \lambda_1, \ldots, \lambda_{C-1})'$, $q = S \times C + C - 1$, and let $\boldsymbol{\mu}(\xi)$ and $\boldsymbol{\sigma}(\xi)$ be the model predictions for $\mathbf{m}$ and $\mathbf{s}$ evaluated at ξ. Furthermore, denote by $\Gamma_1(\xi)$ and $\Gamma_2(\xi)$ the variance-covariance matrices of $\mathbf{m}$ and $\mathbf{s}$, respectively, as predicted under the latent-class model. $\Gamma_1(\xi)$ is a J^* by J^* matrix, whereas $\Gamma_2(\xi)$ is a $\frac{1}{2}J^*(J^*+1)$ by $\frac{1}{2}J^*(J^*+1)$ matrix. Because redundant category counts have been removed, both matrices are positive definite and can be inverted. In Appendix B, it is described how $\boldsymbol{\mu}(\xi)$, $\boldsymbol{\sigma}(\xi)$, $\Gamma_1(\xi)$, and $\Gamma_2(\xi)$ can be computed. Note that $\boldsymbol{\mu}(\xi)$ is a vector that stacks the expected category counts $\mu_{kj}(\xi)$.

The deviations of observed from predicted moments are captured by the vectors $\boldsymbol{\delta}_1 = \mathbf{m} - \boldsymbol{\mu}(\hat{\xi})$ for the means and $\boldsymbol{\delta}_2 = \mathbf{s} - \boldsymbol{\sigma}(\hat{\xi})$ for the variances and covariances. In both cases, the model predictions are evaluated at the maximum likelihood estimates $\hat{\xi}$ of the model parameters. Finally, we need the matrices $A_1(\xi)$ and $A_2(\xi)$ of the first derivatives of $\boldsymbol{\mu}(\xi)$ and $\boldsymbol{\sigma}(\xi)$ with respect to ξ. $A_1(\xi)$ and $A_2(\xi)$ are J^* by q and $\frac{1}{2}J^*(J^*+1)$ by q matrices, respectively. The test statistics are defined as follows:

$$\begin{aligned}
M_1 &= \boldsymbol{\delta}'_1 \left(\Gamma_1^{-1} - \Gamma_1^{-1} A_1 \left\{ A_1' \Gamma_1^{-1} A_1 \right\}^+ A_1' \Gamma_1^{-1} \right) \boldsymbol{\delta}_1, \\
M_2 &= \boldsymbol{\delta}'_1 \left(\Gamma_1 - A_1 \mathcal{I}^{-1} A_1' \right)^+ \boldsymbol{\delta}_1, \\
S_1 &= \boldsymbol{\delta}'_2 \left(\Gamma_2^{-1} - \Gamma_2^{-1} A_2 \left\{ A_2' \Gamma_2^{-1} A_2 \right\}^+ A_2' \Gamma_2^{-1} \right) \boldsymbol{\delta}_2, \\
S_2 &= \boldsymbol{\delta}'_2 \left(\Gamma_2 - A_2 \mathcal{I}^{-1} A_2' \right)^+ \boldsymbol{\delta}_2,
\end{aligned} \tag{10}$$

where all vectors and matrices are evaluated at the maximum likelihood estimates $\hat{\xi}$ of the model parameters, and the superscript + denotes the Moore-Penrose inverse (Magnus & Neudecker, 1988, chap. 2). $\mathcal{I}$ is the expected Fisher information

for the sample of T participants:

$$\mathcal{I} = T\, E\left[\nabla \log(P(\mathbf{n} \mid \xi))(\nabla \log(P(\mathbf{n} \mid \xi)))'\right], \tag{11}$$

where $\nabla \log(P(\mathbf{n} \mid \xi))$ is the vector of first derivatives of the logarithm of $P(\mathbf{N} = \mathbf{n} \mid \xi)$ from Equation 3 with respect to ξ and the expected value is taken over $\mathbf{n}$.

All four statistics are asymptotically distributed as χ^2 when the degrees of freedom are larger than zero and when degenerate cases are excluded, as discussed in Appendix C. The degrees of freedom are for M_1: $J^* - \text{rank}(A_1)$, for M_2: $\text{rank}(\Gamma_1 - A_1\mathcal{I}^{-1}A_1')$, for S_1: $\frac{1}{2}J^*(J^*+1) - \text{rank}(A_2)$, and for S_2: $\text{rank}(\Gamma_2 - A_2\mathcal{I}^{-1}A_2')$. In the next two sections, these statistics are further discussed by way of example.

Example 2: $C = 1$; Testing Parameter Homogeneity

When $C = 1$, the latent-class multinomial model simplifies to a traditional multinomial model, that is, to the core model, and parameter estimation can proceed on the basis of the aggregated frequency counts in the traditional manner. This means that the above test statistics can be computed without applying the modified estimation algorithm described above.

For $C = 1$, M_3 in fact equals the log-likelihood ratio statistic G^2 that is used for testing model fit in a traditional multinomial model analysis (Batchelder & Riefer, 1999). M_1 and M_2, on the other hand, differ from each other and from Fisher's X^2 statistic, $X^2 = T\ \sum_{k=1}^{K}\sum_{j=1}^{J_k}\frac{(m_{kj}-\mu_{kj}(\hat{\xi}))^2}{\mu_{kj}(\hat{\xi})}$, only by the choice of the generalized inverse of the limiting variance-covariance matrix of $\boldsymbol{\delta}_1$ (see Appendix C) for $C = 1$. Accordingly, all three statistics are asymptotically distributed as χ^2 with $J^* - q$ degrees of freedom. Because the model parameters of the core model are uniquely identified from the expected variances and covariances, $\text{rank}(A_2)$ is furthermore given by q, and the degrees of freedom of S_1 by $\frac{1}{2}J^*(J^*+1) - q$. As a consequence, the matrix $A_2'\Gamma_2^{-1}A_2$ has full rank and can be inverted, simplifying the

computation of S_1, because the inverse is of course also the Moore-Penrose inverse. Turning to S_2, $(\Gamma_2 - A_2\mathcal{I}^{-1}A_2')$ has full rank and can be inverted, so that the computation of S_2 is similarly simplified, and there are $\frac{1}{2}J^*(J^*+1)$ degrees of freedom for S_2. In addition, $\mathcal{I}$ has an especially simple form for $C = 1$, its (i, j) element is given by $T\sum_{k=1}^{K}\sum_{l=1}^{J_k}\frac{\partial\mu_{kl}(\xi)}{\partial\xi_i}\frac{\partial\mu_{kl}(\xi)}{\partial\xi_j}\mu_{kl}^{-1}$ (Bishop, Fienberg, & Holland, 1975, chap. 14). To summarize, for $C = 1$:

$$
\begin{aligned}
M_1 = &\quad \boldsymbol{\delta}_1'\left(\Gamma_1^{-1} - \Gamma_1^{-1}A_1\left\{A_1'\Gamma_1^{-1}A_1\right\}^{-1}A_1'\Gamma_1^{-1}\right)\boldsymbol{\delta}_1, && df = J^* - q,\\
M_2 = &\quad \boldsymbol{\delta}_1'\left(\Gamma_1 - A_1\mathcal{I}^{-1}A_1'\right)^{+}\boldsymbol{\delta}_1, && df = J^* - q,\\
S_1 = &\quad \boldsymbol{\delta}_2'\left(\Gamma_2^{-1} - \Gamma_2^{-1}A_2\left\{A_2'\Gamma_2^{-1}A_2\right\}^{-1}A_2'\Gamma_2^{-1}\right)\boldsymbol{\delta}_2, && df = \frac{1}{2}J^*(J^*+1) - q,\\
S_2 = &\quad \boldsymbol{\delta}_2'\left(\Gamma_2 - A_2\mathcal{I}^{-1}A_2'\right)^{-1}\boldsymbol{\delta}_2, && df = \frac{1}{2}J^*(J^*+1).
\end{aligned}
$$

Note that S_2 reduces to the well-known $C(\alpha)$ test for overdispersion in binomial counts (e.g., Chen, 1998) for $C = 1$ and $J^* = 1$. Furthermore, for $C = 1$, if a saturated core model is used that describes the aggregated category counts perfectly, S_1 and S_2 can be seen as "model-free" tests of parameter homogeneity that test whether the observed variance-covariance structure is consistent with a product-multinomial distribution of the raw data.

Consider as an example a real data set obtained in the pair-clustering framework. The pair-clustering paradigm leads to data that are scored in two category systems. $T = 63$ participants performed the experimental task twice, and the data of the first and the second trial were separately coded into a total of four category systems, the first two for the first trial, the second two for the second trial. Thus, $K = 4$, $J_1 = J_3 = 4$, and $J_2 = J_4 = 2$. There were $N_1 = N_3 = 10$ observations per person in category systems 1 and 3, and $N_2 = N_4 = 5$ observations per person in category systems 2 and 4. As core model, the pair-clustering model with $u = a$ is used with separate parameters for each trial. For Trial 1, these were $c^{(1)}, r^{(1)}$, and $u^{(1)}$; for Trial 2, $c^{(2)}$, $r^{(2)}$, and $u^{(2)}$. There are thus $J^* = 8$ non-redundant category counts and $q = 6$ parameters. The traditional

goodness-of-fit test tests whether the model adequately describes the aggregated, or equivalently, the mean category counts by means of the log-likelihood ratio statistic. It revealed a value of $M_3 = 1.85$, indicating a satisfactory model fit with two degrees of freedom, $p = .40$. Similarly, for the asymptotically equivalent statistics M_1 and M_2, $M_1 = M_2 = 1.72$ was observed, yielding essentially the same result, $df = 2$, $p = .42$. The traditional analysis would thus lead one to accept the pair-clustering model and subsequent hypotheses tests would use this model as baseline to test, for example, whether there were any differences in the different core-model parameters as a function of trial.

The statistics S_1 and S_2 can be routinely computed from the output of any traditional multinomial model analysis to defend the as yet untested assumption of parameter homogeneity. Violations of parameter homogeneity, that is, variability over persons in the parameters and possible correlations between the parameters over persons, should lead to overdispersion in the category counts and to inflated correlations between counts for different categories. S_1 and S_2 test whether the variances and covariances of the person-wise category counts are adequately described by the multinomial model under the assumption of parameter homogeneity. In the present case, the assumption of parameter homogeneity is untenable: $S_1 = 281.22$, $df = 30$, $p < .01$, and $S_2 = 322.81$, $df = 36$, $p < .01$.

Example 3: C=2; A Latent-Class Pair-Clustering Model

A latent-class multinomimal model analysis was therefore conducted with two latent classes and the pair-clustering model as core model. This model has $q = 13$ parameters. The parameter estimates and their asymptotic standard errors, estimated from the expected Fisher information matrix, are shown in Table 1. Although redundant, we also give the results for $\lambda_2 = 1 - \lambda_1$. It can be seen that the population of participants is split into two latent classes with proportional sizes $\lambda_1 = .21$ and $\lambda_2 = .79$. The core-model parameters quantify different aspects of participants' memory. Memory performance in all of these aspects increases from

Trial 1 to Trial 2 in each class, and the second class performs considerably worse than the first in all of these aspects.

Testing for mean structure, the model was found to describe the aggregated data well: The log-likelihood ratio test M_3 was $M_3 = 1.90$ with $df = 4$, $p = .75$; $M_1 = 0$ with $df = 0$; and $M_2 = 1.77$, $df = 2$, $p = .41$. This illustrates that the degrees of freedom of M_1 are at most as large as those of M_2.

The two-class model provides a much better account of the variances and covariances of the person-wise category counts: $S_1 = 37.06$, $df = 23$, $p = .03$, and $S_2 = 50.46$, $df = 36$, $p = .06$. This illustrates that S_1 will usually have fewer degrees of freedom than S_2. In the present case, the parameters are identified from the expected variances and covariances, implying that A_2 has full rank and that $A_2'\Gamma_2^{-1}A_2$ can be inverted, simplifying the computation of S_1. In addition, the degrees of freedom of S_1 are therefore simply given by $\frac{1}{2}J^*(J^*+1) - q = 36 - 13 = 23$. The matrix defining the quadratic form S_2, on the other hand, has again full rank, implying that the computation of S_2 can be similarly simplified and that its degrees of freedom are given by $\frac{1}{2}J^*(J^*+1) = 36$.

The test statistic S_1 still indicates a significant deviation of the model predictions from the observed variances and covariances in this example. A three-class solution provided an acceptable fit for both means and variances and covariances with respect to the conventional significance level of $\alpha = .05$. For the sake of simplicity, we accept the two-class solution as a reasonable first description of the data in the continuation of this example below.

When $C > 1$, computation of $\mathcal{I}$ is much more difficult than for $C = 1$. It requires the evaluation of the following sum:

$$\mathcal{I} = T \sum_{\mathbf{n}\in\Omega} \nabla \log(P(\mathbf{n} \mid \xi)) \left(\nabla \log(P(\mathbf{n} \mid \xi))\right)' P(\mathbf{n} \mid \xi), \tag{12}$$

where Ω is the set of vectors $\mathbf{n}$ of non-negative integer elements n_{kj} satisfying $\sum_{j=1}^{J_k} n_{kj} = N_k$ for all k. There are $\prod_k \binom{N_k+J_k-1}{N_k}$ such vectors, implying that for

large models with many categories and large N_ks, the expected information matrix cannot be computed within reasonable times. The so-called observed information matrix $\mathcal{I}_0$ with (i, j) element given by

$$-\sum_{t=1}^{T} \frac{\partial^2}{\partial\xi_i \partial\xi_j} \log(P(\mathbf{n_t} \mid \xi)) \tag{13}$$

can still be computed efficiently, and it can be used instead. However, we hesitate to recommend its use in computing M_2, because $\Gamma_1 - A_1\mathcal{I}_o^{-1}A_1'$ in general does not have the same rank as $\Gamma_1 - A_1\mathcal{I}^{-1}A_1'$. This makes it difficult, for one thing, to compute the correct degrees of freedoms for M_2. More importantly, the derivation of the asymptotic χ^2-distributions of M_2 requires that $\Gamma_1 - A_1\mathcal{I}_o A_1'$ has the correct rank (see Appendix C). It is therefore unknown whether the asymptotic result also holds when $\mathcal{I}$ is replaced by $\mathcal{I}_o$ or when other kinds of estimates of the variance-covariance matrix of the parameters such as bootstrap estimates (McLachlan & Peel, 2000; chap. 2) are used that do not guarantee the correct rank. Consequently, for large models and $C > 1$, we propose to use only M_1 and M_3 for testing mean structure, or to validate the result obtained with M_2 on the basis of $\mathcal{I}_o$ by parametric bootstrap techniques (Efron, 1982). A similar problem as for M_2 does not arise for the analogously constructed S_2 as explained in Appendix C.[2] Note finally that M_1, M_3, and S_1 do not require an estimate of the Fisher information.

As noted by an anonymous reviewer, most of the possible frequency patterns $\mathbf{n}$ will receive zero counts in real data. This means that asymptotic results will not hold for the counts of the possible frequency patterns $\mathbf{n}$ because the data are much too sparse. However, as discussed in Appendix C, the present estimation and goodness-of-fit results require asymptotic approximations only for the distributions of a few highly aggregated statistics; specifically of the first and second moments of the response-category frequencies, computed over persons, and of the so-called efficient scores, that is, of the first derivatives $\partial\xi_i \sum_t \log(P(\mathbf{n_t} \mid \xi))$,

$i = 1, \ldots, S$, of the log-likelihood function. Thus, the data are aggregated into a few summary statistics, and it is at this level of aggregation that asymptotic distributions are needed in justifying the present estimation and goodness-of-fit procedures. The situation can be compared to estimation and goodness-of-fit testing in structural equation modeling. In structural equation modeling, the data are based on continuous variables; as a consequence, almost every possible data point will receive a zero count. Nevertheless, as is well known, the asymptotic approximations that underlie maximum likelihood estimation and goodness-of-fit testing in this context are justified in many applications because they are based on highly aggregated statistics such as the first and second moments and the efficient scores rather than on the extremely sparse original data.

Hypotheses Tests

Starting from a latent-class multinomial model that fits the data, different hypotheses of interest can be tested. We consider two types of tests: 1) Tests diagnosing the source of parameter heterogeneity, and 2) tests of restrictions imposed on the parameters of the core model.

Diagnosing the Origin of Parameter Heterogeneity

To obtain more specific information about the source of parameter heterogeneity, it is possible to consider each parameter of the core model separately and to test, by means of a log-likelihood ratio test, whether it can be set equal across classes. If so, parameter homogeneity can be maintained for that parameter and the psychological process it stands for. For each such test, a latent-class model needs to be estimated that sets equal the parameter in question across classes. As already noted, the above EM-algorithm can still be used if a subset of the core-model parameters is set equal across classes. For the likelihood-ratio test, it is however required that the restricted model retains identifiability. This is usually ensured unless all core model parameters are simultaneously restricted to be equal across classes. Non-identifiability is also signalled by a singular (expected or

observed) Fisher information matrix for the restricted model.

Example 4: Diagnostic Tests for the Latent-Class Pair-Clustering Model. For the above pair-clustering model with two classes, it was tested whether homogeneity could be assumed for the first-trial parameters $c^{(1)}$, $r^{(1)}$, and $u^{(1)}$, by constraining all three of them to be equal across latent classes, that is, in Table 1, across the rows of that table simultaneously. The log-likelihood ratio statistic was 53.40 with 3 degrees of freedom, indicating that parameter homogeneity is already violated for the first-trial data as suggested by the parameter estimates and their standard errors shown in Table 1. Similarly, for the second-trial parameters, the test statistic was 80.26 with 3 degrees of freedom. The test was then conducted for each core-model parameter separately. This revealed that parameter homogeneity was violated for all parameters except the first-trial $r-$parameter; the associated log-likelihood ratio statistic was 0.16, $df = 1$, $p = .69$.

Tests of Restrictions Within Classes

In traditional multinomial model analyses, log-likelihood ratio tests are usually conducted to test psychologically motivated assumptions, once a baseline model has been accepted that fits the aggregate data. In the above pair-clustering model, for example, it might be hypothesized that practice gains from first to second trial are confined to certain memory processes (e.g., those captured by the model parameter c) and do not arise for others (e.g., those captured by the model parameter r). This hypothesis could be evaluated by separately testing (a) equality of the first-trial parameter $c^{(1)}$ and the second-trial parameter $c^{(2)}$ and (b) equality of the first- and second trial parameters $r^{(1)}$ and $r^{(2)}$.

It is possible to conduct analogous tests even if parameter homogeneity is violated, departing from a latent-class multinomial model with $C > 1$ that adequately describes the data. For this purpose, the core-model parameters in question are set equal within each class simultaneously, resulting in a restricted latent-class multinomial model. The log-likelihood ratio test, comparing the

baseline model and the restricted model, is a natural generalization of the corresponding tests conducted in traditional multinomial model analyses.

Example 5: Testing for Differences between Trials. For the above pair-clustering model with two classes, we tested whether the first- and second trial parameters of the core model could be set equal within each class simultaneously, that is, across corresponding columns in each row of Table 1. Thus, it was tested whether $c^{(1)} = c^{(2)}$, $r^{(1)} = r^{(2)}$, and $u^{(1)} = u^{(2)}$ could be maintained for both latent classes simultaneously. The log-likelihood ratio statistic was 87.05, $df = 6$, $p < .01$, indicating that the practice gains from first to second trial were significant. The effects of trials can of course be tested separately for each core-model parameter $c^{(i)}$, $r^{(i)}$, or $u^{(i)}$. The principle should be clear, and we omit the results for the sake of brevity.

Tests of Restrictions Within a Subset of Classes

The tests considered in the previous subsection simultaneously impose the core-model restrictions of interest within each latent class. Sometimes, researchers may be interested in performing even more specific tests involving only a subset of classes. For example, it might be of interest whether practice gains are confined to the first class of high performers in the above example. To test this assumption, it is necessary to test for equality of the first- and second trial model parameters in each latent class separately.

The log-likelihood ratio method cannot be used to construct such tests as explained next. The above EM algorithm can be used to fit a restricted model in which certain restrictions are imposed only on a subset of classes. This will lead to parameter estimates that satisfy the restrictions for the pre-specified number of classes. There is no control, however, over which latent classes will turn out to carry the restrictions in the maximum-likelihood solution. This lack of control over which class will turn out to be restricted compromises the log-likelihood ratio approach for the present question.

For this reason, it is recommended to use Wald's test (Rao, 1973, chap. 6) for post-hoc diagnostic tests that involve differential treatment of specific latent classes. Like the log-likelihood ratio test, Wald's test W is a large-sample test with an asymptotic χ^2-distribution. For the present purposes, it has the advantage that it can be computed on the basis of the parameter estimates obtained for the unrestricted model and an estimate of their variance-covariance matrix. In particular, it does not require the estimation of the restricted model. Consequently, the problem just described does not arise. Specifically, assume that we want to test $k < q$ independent restrictions $R_i(\xi) = 0$, $i = 1, \ldots, k$, imposed on the model parameters ξ, and let U be the k by q matrix of the first derivatives of the functions R_i. That the R_i are independent restrictions means that U has full rank k, and that $U'\mathcal{I}^{-1}U$ is invertible, where $\mathcal{I}$ is the expected Fisher information matrix or, when this is not available, the observed Fisher information. Let the (i, j) element of its inverse be u^{ij}. Wald's test is defined as

$$W = \sum u^{ij}(\hat{\xi})R_i(\hat{\xi})R_j(\hat{\xi}). \tag{14}$$

It basically tests whether the deviations that the maximum likelihood estimates $\hat{\xi}$ exhibit from the restrictions R_i are within the range of random fluctuations that is consistent with their standard errors. The test statistic W is asymptotically distributed as χ^2 with k degrees of freedom.

Example 6: $C = 2$; Testing Within Only One Class

For the above pair-clustering model with two classes, we tested whether the first- and second trial $c-$parameters of the first class of high performers (Class 1 in Table 1) could be set equal. This amounts to testing the restriction $c^{(1)} - c^{(2)} = 0$ for the $c-$parameters of the first class. For this restriction, it is easy to see that $U'\mathcal{I}^{-1}U$ is the estimated asymptotic variance of the difference of the maximum likelihood estimates, $\hat{c}^{(1)} - \hat{c}^{(2)}$, of the $c-$parameters of the first class. In terms of the estimated variances and covariances that are the elements of the

inverse of the Fisher information: $U'\mathcal{I}^{-1}U = \sigma^2(\hat{c}^{(1)}) - 2\sigma(\hat{c}^{(1)}, \hat{c}^{(2)}) + \sigma^2(\hat{c}^{(2)})$. Thus, $W = (\hat{c}^{(1)} - \hat{c}^{(2)})^2/(\sigma^2(\hat{c}^{(1)}) - 2\sigma(\hat{c}^{(1)}\hat{c}^{(2)}) + \sigma^2(\hat{c}^{(2)})) = 3.68$. According to the χ^2-distribution with $df = 1$, this value of W just misses conventional significance levels, $p = .06$.

Monte Carlo Study

A small Monte Carlo study was conducted to further illustrate the issues discussed so far. The standard pair-clustering model with $u = a$ was used for this purpose. The analyses are based on generated data sets for $T = 25$ participants that contribute $N_1 = 12$ and $N_2 = 6$ data points for word pairs and singletons, respectively. The core model has $S = 3$ parameters, c, r, and u, and there are $J^* = 4$ non-redundant category counts. Each analysis is based on 5,000 data sets with 25 simulated participants per data set.

Data were generated from a one-class model H_1 with $c = .7$, $r = .5$, and $u = .3$, and from a two-class model H_2 in which these parameter values characterize a first, large class with $\lambda_1 = .75$, and to which a second small class, $\lambda_2 = .25$, with parameter values $c_2 = .3$, $r_2 = .5$, and $u_2 = .7$ was added. Thus, there is a moderate amount of heterogeneity. Data generated from the one-class model H_1 were analyzed by means of a traditional one-class pair-clustering model ($C = 1$), data generated from the two-class model H_2 were analyzed both by means of the traditional model as well as by a two-class pair-clustering model ($C = 2$).

Table 2 shows the goodness-of-fit results for nominal significance levels $\alpha = .10$, .05, and .01. The probabilities of rejection are the proportions with which the different test statistics exceeded the critical value of the appropriate asymptotic χ^2-distribution over the 5,000 replications. The first three rows of Table 2 evaluate the actual significance levels that are obtained when the asymptotic distributions are used in a traditional analysis ($C = 1$) without parameter heterogeneity (H_1). It can be seen that the asymptotic results provide a satisfactory approximation even

for as few as $T = 25$ participants although both S_1 and S_2 are slightly too progressive.

The next three rows of Table 2 evaluate the test power for detecting the moderate level of heterogeneity built into H_2 when the data are analyzed by the traditional one-class model, $C = 1$. It can be seen that all test statistics exhibit some amount of sensitivity for detecting heterogeneity (see Appendix A). Yet, there is a good chance to fail to detect this kind of model violation using the traditional multinomial model test M_3 or the other mean-structure tests. For example, for $\alpha = 0.05$, the likelihood of an error of this kind is approximately $\beta = .43$. The variance-covariance structure tests S_1 and S_2 detect the heterogeneity almost with certainty.

The final three rows of Table 2 show the results when the heterogeneous data are analyzed by means of the appropriate latent-class multinomial model with $C = 2$. Thus, the probabilities of rejection reflect the actual significance levels that can be expected in this situation when the asymptotic distributions of the test statistics are relied upon. Statistic M_1 is useless in this situation, because it has zero degrees of freedom. As can be seen, performance of the other test statistics is again satisfactory despite the small T.

Table 3 shows the mean parameter estimates and their standard errors that emerged in these analyses. The actual standard errors (i.e., the standard deviations of the parameter estimates over the 5,000 generated data sets) and the mean model estimates of standard errors obtained from the Fisher information are shown. The first row of Table 3 demonstrates that the traditional analysis recovers the underlying parameter estimates and provides appropriate estimates of their standard errors when there is no parameter heterogeneity. The next row exemplifies the traditional analysis in the presence of parameter heterogeneity. The population averages of the parameter values across the two classes are .60, .50, and .40 for c, r, and u, respectively. It can be seen that the estimates of c and r

approach the respective averages quite well, whereas the estimate of u is biased away from its average. In addition, the model estimates of the standard errors somewhat underestimate the actual standard errors, especially in the case of u (see Appendix A). The final row shows the performance of the two-class pair-clustering model in analyzing data generated from H_2. It can be seen that the model recovers both underlying classes quite well. In addition, the estimates of the standard errors of the parameter estimates approximate the actual standard errors satisfactorily.

General Discussion

In applying multinomial processing tree models, it is assumed that the same parameters describe the data of each person in the sample. In many areas of application, the assumption may be unwarranted. For example, when reasoning or memory processes are modelled (for many examples of such models, see Batchelder & Riefer, 1999), the model parameters usually quantify different aspects of task performance. At the same time, people frequently differ substantially in their cognitive skills, leading to violations of parameter homogeneity. What is more, different performance aspects in the cognitive domain are usually correlated over persons, leading to correlations between the parameters that capture these different aspects. Frequently, guessing parameters are also incorporated in the models to account for participants' responses when they are in states of uncertainty. Guessing biases are often sensitive to a host of extraneous variables such as participants' knowledge or familiarity feelings, that are likely to lead to different biases for different participants. When parameter homogeneity is violated, multinomial model analyses can be misleading in many respects (see Appendix A). Tests for parameter homogeneity were proposed that can be computed from the output of traditional analyses and that allow researchers to defend the assumption of parameter homogeneity.

If parameter homogeneity is found to be violated, a natural solution is to

extend multinomial models to hierarchical models (Raudenbush & Bryk, 2002). Hierarchical multinomial models extend multinomial models by allowing for different parameters for each person and making an assumption about the distribution of parameter values over persons. Different distributions might be assumed such as beta distributions of the model parameters or a multivariate normal distribution of parameters transformed to range over the entire real line. Latent-class multinomial models are hierarchical models in which a discrete distribution of the model parameters is assumed. Compared to other possible distributions this leads to a relatively tractable family of models that allows for parameter heterogeneity as well as for correlated parameters.

In the latent-class approach, persons fall into one of several mutually exclusive classes. In each class, parameter homogeneity is assumed to hold. There are situations in which this is a realistic assumption such as when the population is split into relatively homogeneous subpopulations according to unknown discrete background variables such as level of education, pathology, and so forth. In many other cases a continuous distribution of parameters is more plausible. In such cases, the latent-class approach can only be claimed to provide a better approximation of the real state of affairs than traditional multinomial models. In fact, other kinds of distributions can be approximated arbitrarily well by the discrete distributions underlying latent-class multinomial models as the number of classes is increased. In practice, discrete approximations frequently provide satisfactory approximations of continuous distributions even if they are based on only a few points (e.g., Pinheiro & Bates, 1995). The goodness-of-fit tests developed in this paper allow one to assess the appropriateness of the approximation for a given data set.

Latent-class multinomial models can be used to conduct the same kinds of analyses that are of interest in traditional multinomial analyses. Thus, they allow researchers to proceed even if parameter homogeneity is found to be violated for the data set under study. Their application makes use of the original data set so

that there is no need for additional data collection. Computer software is being developed for model estimation, goodness-of-fit tests and hypotheses tests as discussed in this paper that uses the same format for the core-model description and person-wise data as the computer programs available for the traditional multinomial-model method (i.e,. *.eqn-files and *.mdt-files, respectively, Hu, 1991, 1998; Rothkegel, 1999) to be presented elsewhere. Note, however, that latent-class multinomial models with small numbers of classes may still fail to fit the data. In addition, the proposed goodness-of-fit measures do not only require stable estimates of the mean category counts as in the traditional method, but also of their variances and covariances as in structural equation modeling. This means that sample sizes like those realized in structural equation modeling should be used if the asymptotic approximations for the distributions of the goodness-of-fit tests are to be relied upon. If sample size is small, it is prudent to use Bootstrap techniques such as the modified Bootstrap method proposed by Bollen and Stine (1993) or the parametric Bootstrap method (Efron, 1982) to evaluate the significance of observed goodness-of-fit values.

On the other hand, very large samples give rise to another set of problems. Most distributions underlying real data are likely to be more complex than the fitted models, so that the models provide at best approximations of them. Under this premise, goodness-of-fit tests and model-selection procedures will in general favor complex models as sample size is increased (Linhart & Zuccini, 1986), that is, models with many latent classes in the present case. Model selection criteria such as Akaike's information criterion or the Bayes information criterion introduce a penalty for the complexity of the model and may help to alleviate this tendency when sample size is large. A review of these and related issues in the area of model selection is provided by Linhart and Zuccini (1986).

An alternative approach to deal with parameter heterogeneity that has sometimes been used in the literature is to analyze each person's data separately

and to consider the variance-covariance matrix of the person-wise parameter estimates. Although this approach has heuristic value, it confounds (a) substantive covariances and variances of the "true" person-wise parameter values with (b) spurious covariances and variances that go back to estimation error. Estimation error is likely to be sizeable due to the sparseness of the data at the level of individual participants. For example, no method is currently available in this approach to test whether the observed variance-covariance matrix of the person-wise parameter estimates is consistent with the assumption of parameter homogeneity or not.

Parameters may exhibit variability and correlations over persons. Similarly, items may be a source of heterogeneity. For example, the different words that are used in a pair-clustering recall task may systematically differ in their memorability. Latent-class multinomial models still incorporate the assumption that items are homogeneous. That is, if parameters are indexed by persons *and* items, it is assumed that the parameter values are equal across items within each person. Although potentially problematic, we believe that this assumption is less grave than the assumption of homogeneous persons. Experimenters can usually exert much more control over the items that they use than over the persons, and items are often carefully selected or constructed to ensure homogeneity. In addition, a simple remedy for the problem of item heterogeneity is often possible if item heterogeneity is feared to be a problem. Frequently large pools of appropriate items can be defined. If items are randomly sampled from large item pools for each participant anew, the principle of randomization ensures that any "random" item will be statistically equivalent to any other random item.

A natural and important next step in the study of latent-class multinomial models is to consider data with multiple independent groups of participants. For example, an experimental manipulation is often made between participants in work using multinomial models, and the question is then whether and how the

experimental groups differ in terms of the processes that are represented by the different model parameters. Two routes to dealing with multiple groups seem possible. One way is to analyze the data jointly with different model parameters for each group as is done in traditional multinomial models. Although this does not present any new statistical difficulties, the intended between-groups comparisons for given core-model parameters are rendered more difficult conceptually: Profiles of parameter values now have to be compared between groups (that is, profiles of the parameter value of the core-model parameter in question over the different classes). This is likely to be especially difficult, although not impossible, if the different experimental groups have to be described by different numbers of classes. For example, in most cases, the question of interest is whether or not the central tendency of a given core-model parameter θ_s differs between the groups. For a given group, the central tendency or expected value is $\sum_{c=1}^{C} \lambda_c \theta_{sc}$, and it is thus a function of the model parameters estimated for that group. The equality of these functions across groups can be tested by means of traditional methods such as Wald's test described above even if different numbers of classes are used for the different groups.

An alternative approach is to model the entire data set by means of the same latent classes and allowing for differences between groups only in the class sizes λ_c. That is, different class sizes λ_{cg} are allowed for each group g, but the same latent-class parameters $\boldsymbol{\theta}_c$, $c = 1, \ldots, C$, are used for all experimental groups. This approach is akin to the common group-structure approach of structural equation modeling of multiple groups (e.g., Kaplan, 2000, chap. 4). Differences between groups then emerge in the class sizes compared across groups. The core-model parameters of classes that are differentially represented in the different groups could then be directly compared to characterize the psychological processes that are and are not affected by the group factor. It remains to be seen which one of these approaches will prove to be more useful in applications.

References

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

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

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

Bollen, K. A. & Stine, R. A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 111–135). Newbury Park, CA: Sage.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Chen, J. (1998). Penalized likelihood-ratio test for finite mixture models with multinomial observations. *Canadian Journal of Statistics*, *26*, 583–599.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *39*, 1–38.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Erdfelder, E. (2000). *Multinomiale Modelle in der kognitiven Psychologie [Multinomial models in cognitive psychology]*. Unpublished habilitation thesis, Psychologisches Institut der Universität Bonn, Germany.

Hu, X. (1991). Statistical inference program for multinomial binary tree models [Computer software]. University of California at Irvine.

Hu, X. (1998). GPT – HomePage [Computer software and documentation]. Retrieved from http://xhuoffice.psyc.memphis.edu/gpt/.

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

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete multivariate distributions*. New York: Wiley.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1993). *Univariate discrete distributions* (2nd ed.). New York: Wiley.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, California: Sage.

Linhart, H. & Zuccini, W. (1986). *Model Selection*. New York: Wiley.

Magnus, J. R. & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. Chichester, England: Wiley.

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

Moore, D. S. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, *72*, 131–137.

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Pinheiro, J. C. & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*, 12–35.

Powell, M. J. D. (1977). Restart procedures for the conjugate gradient method. *Mathematical Programming*, *12*, 241–254.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods*. Thousand Oaks, California: Sage.

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

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, *31*, 696–700.

Satorra, A. (1992). Asymptotic robust inference in the analysis of mean and covariance structures. In P. V. Marsden (Ed.), *Sociological methodology 1992* (pp. 249–278). Oxford: Blackwell.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

Appendix A: Consequences of Parameter Heterogeneity

To demonstrate the consequences of parameter heterogeneity, consider the so-called pair-clustering model (Batchelder & Riefer, 1986, 1999; Riefer & Batchelder, 1991) described in the body of the paper. The traditional method proceeds from the counts n_{kj} with which each response category C_{kj} is observed, aggregated over participants and the N_1 word pairs and N_2 singletons in the list.

The model is based on the four parameters c, r, u, and a. Under parameter homogeneity, the maximum-likelihood estimates of u and a are given by (Riefer & Batchelder, 1991):

$$\hat{a} = \frac{n_{21}}{n_{21} + n_{22}}, \quad \text{and} \tag{15}$$

$$\hat{u} = \frac{2n_{12}}{2n_{12} + n_{13}}. \tag{16}$$

Frequently, one is interested in testing the model restriction $a = u$. This is done by a log-likelihood ratio test with one degree of freedom that is asymptotically equivalent to the following $z-$test (Rao, 1973; chap. 6e):

$$z = \frac{\hat{a} - \hat{u}}{\sqrt{Var(\hat{a} - \hat{u})}}. \tag{17}$$

Furthermore, given parameter homogeneity, it follows that $Var(\hat{a} - \hat{u}) = Var(\hat{a}) + Var(\hat{u})$, because both parameters are estimated from different data, that is, singletons data and word pairs data, respectively, and responses are assumed to be distributed independently within and across persons. To compute the $z-$value, the maximum likelihood estimates $\hat{a}$ and $\hat{u}$ are entered into the equations for the variances that are given below.

Standard asymptotic theory reveals that $\hat{a}$ and $\hat{u}$ are asymptotically distributed normally as the number of participants T and thereby the number of data points, TN_1 und TN_2, is increased, given parameter homogeneity. The means

of the asymptotic distribution are given by a and u, respectively, and the variances by[3]

$$Var(\hat{a}) = \frac{1}{TN_2}a(1-a), \quad \text{and} \tag{18}$$

$$Var(\hat{u}) = \frac{1}{TN_1}\frac{(2-u)(1-u)}{2(1-c)}. \tag{19}$$

Given parameter homogeneity and $a = u$, the above $z-$value is asymptotically distributed according to a standard normal distribution (with mean zero and standard deviation equal to one). What are the consequences of parameter heterogeneity on parameter estimation and on this $z-$test for equality of u and a, if $u = a$ still holds at the aggregate level of the means of the parameters a and u over persons?

Following Riefer and Batchelder (1991), assume that the model parameters $x = c, r, u$, and a follow independent beta distributions with parameters μ_x and γ_x (the parameters μ and γ of the beta distribution are defined in the introduction of the paper). It is easy to see that the probabilities with which a data point produced by a randomly sampled participant falls into one of the response categories are given by:[4]

$$\begin{aligned}
p(C_{11}) &= \mu_c\mu_r \\
p(C_{12}) &= (1-\mu_c)\mu_u(\mu_u(1-\gamma_u)+\gamma_u) \\
p(C_{13}) &= 2(1-\mu_c)\mu_u(1-\mu_u)(1-\gamma_u) \\
p(C_{14}) &= \mu_c(1-\mu_r)+(1-\mu_c)(1-\mu_u)(1-(1-\gamma_u)\mu_u) \\
p(C_{21}) &= \mu_a \\
p(C_{22}) &= 1-\mu_a
\end{aligned} \tag{20}$$

Comparing Equation 20 and the model equations for the pair-clustering model in the body of the paper (Equation 1), it is seen that the category

probabilities are no longer described by the original model. Nevertheless, it follows from the central-limit theorem that the statistics $\frac{1}{\sqrt{T}}(n_{kj} - TN_k p(C_{kj}))$ are jointly asymptotically normally distributed with mean zero as the number of participants T is increased. Using Equation 20, it can therefore be shown that $\hat{u}$ from Equation 16 is a consistent estimate of $\mu_u + \gamma_u(1 - \mu_u)$, while $\hat{a}$ from Equation 15 is a consistent estimate of μ_a. Thus, $\hat{u}$ is biased away from the population average μ_u of the individual $u-$values, whereas this is not the case for $\hat{a}$.

As a consequence, the numerator of the $z-$value of Equation 17 does not converge to zero, as the number of participants increases, but to the value $\gamma_u(1 - \mu_u)$, assuming that $\mu_a = \mu_u$ holds for the averages of a and u. Simultaneously, the denominator does converge to zero, and the $z-$test will almost surely reject the equality restriction $u = a$ even though on average the two are equal. The same is in fact true if $u = a$ holds for each individual. In the present case, the problem is further exacerbated by the fact that the variance that is used in the denominator of Equation 17 underestimates the actual variance when there is parameter heterogeneity. As a consequence, the variance of the $z-$value, based on this too small estimate, will exhibit a variance that is systematically larger than one, making large $z-$values more likely than is to be expected under a normal distribution with variance one. Figure A1 shows how the actual significance levels for the test diverge from the nominal level for levels of heterogeneity $\gamma \leq 0.2$. In Figure A1, it is assumed that $\mu_u = \mu_c = \mu_r = \mu_a = 0.5$, and $\gamma_u = \gamma_c = \gamma_r = \gamma_a = \gamma$. In this situation, values of γ of .2 imply that the standard deviations of parameter values across persons equal .22, corresponding to a very high level of heterogeneity. The three functions displayed were computed, in the order from bottom to top, with $(N_1, N_2, T) = (10, 5, 50)$, $(N_1, N_2, T) = (20, 10, 50)$, and $(N_1, N_2, T) = (20, 10, 100)$, and a nominal significance level α of 0.05 on the basis of the asymptotic normal distribution of $\frac{1}{\sqrt{T}}n_{kj}$. It can be seen that the actual significance levels quickly diverge from the nominal level as γ increases. It is also

seen that the problem increases as N_1, N_2, and T are increased.

In sum, parameter heterogeneity can cause the following problems:

- Parameter estimates can diverge systematically from the averages of the underlying person-wise parameters,
- confidence intervals that are estimated for parameters can be too small,
- significance tests for equality restrictions imposed upon parameters can exhibit inflated significance levels even if the restrictions hold on average, or for each individual, and
- goodness-of-fit tests may exhibit inflated significance levels even if the model holds for each individual.

When parameters are correlated over persons, the extent of such problems can vary widely depending upon the exact nature of the variance-covariance structure of parameters. This makes it difficult to predict the extent of these problems for any given application.

Appendix B: Moments Required for Goodness-of-Fit Tests

The predicted first moments $\boldsymbol{\mu}(\xi)$ are simply given by $\mu_{kj} = N_k \sum_c \lambda_c p_{kj}(\boldsymbol{\theta}_c)$. Furthermore, $\Gamma_1(\xi)$ and $\boldsymbol{\sigma}(\xi)$ contain the model predictions for the second central moments of the mean category counts and the person-wise category counts, respectively. Their computation requires the predicted second central moments of the person-wise category counts. The elements of $\Gamma_2(\xi)$ are the model predictions for the variances and covariances of the observed variances and covariances of person-wise category counts. To compute Γ_2, the predicted central fourth-order moments are needed. A typical element $\gamma_{ij,kl}$ of Γ_2 is the predicted covariance of (1) the observed covariance of the category counts for Categories i and j, s_{ij}, and (2) the observed covariance of the category counts for Categories k and l, s_{kl}:

$$\gamma_{ij,kl} = \frac{1}{T-1}\left\{\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}\right\} + \frac{1}{T}\left\{\sigma_{ijkl} - \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk}\right\}, \tag{21}$$

where σ_{st} is the predicted covariance of the category counts for Categories s and t as listed in $\boldsymbol{\sigma}(\xi)$ and σ_{ijkl} is the predicted fourth central moment of the involved four category counts (Browne, 1984).

It is convenient to compute the required predicted second-order and fourth-order central moments from the so-called mixed descending factorial moments of the category counts. These take on a particularly simple form for the product-multinomial distribution. For non-negative integer numbers u_{kj}, the mixed descending factorial moment of order $\mathbf{u} = (u_{11}, u_{12}, \ldots, u_{KJ_K})$ of the distribution of the person-wise category counts $\mathbf{N}$ is defined by (Johnson, Kotz, & Balakrishnan, 1997, Equation 34.14)

$$\mu'(\mathbf{u}) = E\left[N_{11}^{(u_{11})} N_{12}^{(u_{12})} \ldots N_{K,J_K}^{(u_{KJ_K})}\right] = \sum_{\mathbf{n}\in\boldsymbol{\Omega}} \prod_{k=1}^{K} \prod_{j=1}^{J_k} n_{kj}^{(u_{kj})} P(\mathbf{n} \mid \xi), \tag{22}$$

where $a^{(b)} = a(a-1)\ldots(a-b+1)$ for $b \leq a$ and $a^{(b)} = 0$ for $b > a$, and Ω is the support of the distribution of person-wise category counts, that is, the set of all

vectors $\mathbf{n}$ of length J with non-negative integer elements n_{kj} and $\sum_{j=1}^{J_k} n_{kj} = N_k$, $k = 1, \ldots, K$. It follows from Equation 35.5 in Johnson et al. (1997) that

$$\mu'(\mathbf{u}) = \left(\prod_{k=1}^{K} N_k^{\left(\sum_{j=1}^{J_k} u_{kj}\right)} \right) \sum_{c=1}^{C} \lambda_c \prod_{k=1}^{K} \prod_{j=1}^{J_k} p_{kj}^{u_{kj}}(\boldsymbol{\theta}_c). \tag{23}$$

Based on these decreasing moments, computed up to the fourth order ($\sum_k \sum_j u_{kj} \leq 4$), the non-central moments of order $\mathbf{u}$ can be computed as linear combinations of decreasing moments of order $\mathbf{u}$ and lower orders using the so-called Stirling numbers of the second kind as coefficients (Johnson et al., 1997, Equation 34.18). Given the non-central moments, it is straightforward to compute the required central moments using Equation 34.28 in Johnson et al. (1997). To compute the matrices A_1, A_2, and $\mathcal{I}$, the derivatives of the category probabilities $p_{kj}(\boldsymbol{\theta})$ are required. Expressions for the derivatives are given by Hu and Batchelder (1994).

Similar statistics as M_1, M_2, S_1, and S_2 have been proposed in the context of so-called asymptotic distribution-free structural-equation modeling (e.g., Browne, 1984). Unlike in that context, the variance-covariance matrices Γ_1 and Γ_2 are not estimated from the data directly, which for Γ_2 would require very large samples for stable results, but are computed as the model predictions as just described, evaluated at the maximum likelihood estimates of the model parameters.

Appendix C: Asymptotic Distribution of Goodness-of-Fit Tests

For deriving the asymptotic distribution of the proposed goodness-of-fit tests, it is necessary to exclude a couple of degenerate cases. First, it is assumed that the core model is globally identifed. Second, the true parameter value ξ_0 is assumed to fall in the interior of the parameter space, and the model is assumed to be identified at ξ_0 (up to possible permutations of the classes, cf. Section on Identifiability). For $C > 1$, this implies in particular that for $\xi_0 = (\boldsymbol{\theta}'_{0,1}, \ldots, \boldsymbol{\theta}'_{0,C}, \lambda_{0,1}, \ldots, \lambda_{0,C-1})'$, $\boldsymbol{\theta}_{0,c} \neq \boldsymbol{\theta}_{0,d}$ for all c and d with $c \neq d$. If $\boldsymbol{\theta}_{0,c} = \boldsymbol{\theta}_{0,d}$ for any two classes c and d, $\lambda_{0,c}$ and $\lambda_{0,d}$ are not identified, and latent classes c and d can be merged into one class so that the model reduces to one with fewer latent classes. So far, these are fairly standard assumptions of large-sample theory, and the asymptotic χ^2-distribution of the log-likelihood ratio test M_3 follows from standard large-sample theory under these assumptions, if the multinomial latent-class model with saturated core model is also identified.

The statistics M_1, M_2, S_1, and S_2 are constructed following a common principle described by Moore (1977): All four are quadratic forms $Y'B(\hat{\xi})Y$ in certain statistics Y that are asymptotically normally distributed with a certain variance-covariance matrix $\Sigma(\xi_0)$ if the model holds. From the asymptotic distribution of Y, it follows that the quadratic form $Y'B(\xi_0)Y$ is asymptotically χ^2-distributed with degrees of freedom given by the rank of $\Sigma(\xi_0)$, if $B(\xi_0)$ is a generalized inverse of $\Sigma(\xi_0)$ (Moore, 1977). This also holds if a consistent estimate of $B(\xi_0)$ is used rather than $B(\xi_0)$ itself, and the statistics M_1, M_2, S_1, and S_2 are based on consistent estimates of generalized inverses.

As will be seen below, the consistency of these estimates rests on the condition that the matrices A_1, A_2, $\Gamma_1 - A_1\mathcal{I}^{-1}A'_1$, and $\Gamma_2 - A_2\mathcal{I}^{-1}A'_2$ have constant rank at least in a neighbourhood of ξ_0. To ensure this, a strong identifiability condition is necessary. That the model is identified implies that the

latent category probabilities $p_{kj}(\boldsymbol{\theta}_{0,c})$ differ between any two latent classes for at least one k and j. This is still consistent, however, with equality of latent category probabilities p_{kj} for all categories j of a subset of category systems k between two (or more) latent classes. If this case should arise, it is possible that $A_1(\xi_0)$ and $A_2(\xi_0)$ have reduced rank relative to the case where the p_{kj} differ between latent classes for any two latent classes for at least one category j of each category system k. It is therefore assumed that for all k, $k = 1, \ldots, K$, and all $c \neq d$, $c, d = 1, \ldots, C$, there is at least one j with $p_{kj}(\boldsymbol{\theta}_{0,c}) \neq p_{kj}(\boldsymbol{\theta}_{0,d})$. In other words, differences between any two latent classes should be seen in all category systems or subtrees of the model simultaneously. For many cases, it is easy to see that this condition is sufficient for constant rank of the above matrices as discussed below. Note however that I have not found a proof of this assertion in the general case although I believe it is true. In applications, parameter estimates can be examined for violations of the condition. If differences between latent classes are assumed for, or found to be confined to, only certain subtrees, restricted models can be fitted that allow for differences between latent classes only in subsets of category systems k. In this case, too, constant rank is obtained, though at the lower level that can be the consequence of equalities between latent classes in certain subtrees of the model.

Under the assumption of constant rank, the asymptotic distributions of M_1 and M_2 are derived in the sequel. The derivations are similar for the analogously constructed S_1 and S_2. Let the J^* non-redundant observed person-wise category counts be given by $\mathbf{n}_t$, $t = 1, \ldots, T$, and their expected values by $\boldsymbol{\mu}(\xi)$. Let the statistics V_t and F_t be defined as

$$\begin{aligned} V_t &= \nabla \log P(\mathbf{n_t} \mid \xi), \\ F_t &= \mathbf{n}_t - \boldsymbol{\mu}(\xi), \end{aligned}$$

both evaluated at $\xi = \xi_0$. Furthermore, let $V = \frac{1}{\sqrt{T}} \sum_t V_t$ and $F = \frac{1}{\sqrt{T}} \sum_t F_t$. The statistics (V_t, F_t), $t = 1, \ldots, T$, are independently and identically distributed with

expected value zero and finite variance (Rao, 1973, chap. 6e). It follows from the central limit theorem that (V, F) is asymptotically normally distributed with mean zero as T goes to infinity. The variance-covariance matrix of the asymptotic distribution is given by the variance-covariance matrix of (V_t, F_t) that is independent of t. The variance-covariance matrix of V_t is simply $\mathcal{J}$, the expected Fisher information for a sample with $T = 1$ observation, evaluated at ξ_0 (Rao, 1973, chap. 6e1). The variance-covariance matrix of F_t is obviously $\Sigma_1(\xi_0)$, the variance-covariance matrix of the non-redundant person-wise category counts, $\mathbf{n}_t$, as predicted by the model. Note that $\mathcal{J} = \frac{1}{T}\mathcal{I}$, the expected Fisher information in the sample, and that $\Sigma_1(\xi_0) = T\Gamma_1(\xi_0)$. The covariance of the elements V_{it} and F_{jt} of V_t and F_t can be computed as follows:

$$\begin{aligned} E[V_{it}F_{jt}] &= E\left[\frac{1}{P(\mathbf{n} \mid \xi)}\frac{\partial P(\mathbf{n} \mid \xi)}{\partial \xi_i}(n_j - \mu_j)\right] \\ &= \sum_{\mathbf{n}} \frac{\partial P(\mathbf{n} \mid \xi)}{\partial \xi_i}(n_j - \mu_j) = \sum_{\mathbf{n}} \frac{\partial P(\mathbf{n} \mid \xi)}{\partial \xi_i} n_j \\ &= \frac{\partial}{\partial \xi_i}\sum_{\mathbf{n}} n_j P(\mathbf{n} \mid \xi) = \frac{\partial \mu_j}{\partial \xi_i}, \end{aligned}$$

noting in the second row of the equation that
$\mu_j \sum_{\mathbf{n}} \frac{\partial P(\mathbf{n} \mid \xi)}{\partial \xi_i} = \mu_j \frac{\partial}{\partial \xi_i} \sum_{\mathbf{n}} P(\mathbf{n} \mid \xi) = 0$. Thus, the asymptotic covariances are given by $A_1(\xi_0)$, and the entire asymptotic variance-covariance matrix of (V, F) by

$$\begin{pmatrix} \mathcal{J} & A_1' \\ A_1 & \Sigma_1 \end{pmatrix} \tag{24}$$

evaluated at $\xi = \xi_0$.

From here, the asymptotic distribution of $\sqrt{T}\boldsymbol{\delta}_1$ can be derived as follows: Let D be $D = \sqrt{T}(\hat{\xi} - \xi_0)$, where $\hat{\xi}$ are the maximum-likelihood estimates of ξ_0. It is well known that D is asymptotically equivalent to $\mathcal{J}^{-1}V$ (Rao, 1973, chap. 6e1). In addition, $\Delta = \sqrt{T}(\boldsymbol{\mu}(\hat{\xi}) - \boldsymbol{\mu}(\xi_0))$ is asymptotically equivalent to $A_1 D$ (Rao, 1973, chap. 6a2). It follows that $\Delta \stackrel{a}{=} A_1 D \stackrel{a}{=} A_1\mathcal{J}^{-1}V$. Because $\sqrt{T}\boldsymbol{\delta}_1 = \sqrt{T}(\mathbf{m} - \boldsymbol{\mu}(\hat{\xi})) = F - \Delta$, it holds that $\sqrt{T}\boldsymbol{\delta}_1 \stackrel{a}{=} F - A_1\mathcal{J}^{-1}V$. Hence,

$\sqrt{T}\boldsymbol{\delta}_1$ is asymptotically normally distributed with mean zero and variance-covariance matrix given by $\Sigma(\xi) = \Sigma_1 - A_1\mathcal{J}^{-1}A_1'$, evaluated at $\xi = \xi_0$.

Note that M_2 equals $(\sqrt{T}\boldsymbol{\delta}_1)'\Sigma^+(\sqrt{T}\boldsymbol{\delta}_1)$, evaluated at $\xi = \hat{\xi}$. If the Moore-Penrose inverse $(\Sigma(\hat{\xi}))^+$ that is used in this formula is a consistent estimate of $(\Sigma(\xi_0))^+$, it follows from Theorem 2a in Moore (1977) that M_2 is asymptotically distributed as χ^2 with degrees of freedom given by $\text{rank}(\Sigma(\xi_0))$. The function mapping parameters ξ onto the Moore-Penrose inverse of $\Sigma(\xi)$ is a continuous function at ξ_0, if $\text{rank}(\Sigma(\xi))$ is constant for all ξ in a neighbourhood of ξ_0 (Magnus & Neudecker, 1988, chap. 8.5). Because the maximum-likelihood estimates are consistent estimates of ξ_0, it follows under this condition that $(\Sigma(\hat{\xi}))^+$ is a consistent estimate of $(\Sigma(\xi_0))^+$ and that the degrees of freedom of M_2 are given by $\text{rank}(\Sigma(\hat{\xi}))$ almost surely, as the number of participants is increased. Thus, under the rank condition, M_2 is asymptotically distributed as χ^2 with $df = \text{rank}(\Sigma(\hat{\xi}))$. Note however, that we found the condition to be true only if the expected Fisher information is used in the computation of M_2, whereas the use of other consistent estimates of the Fisher information such as the observed Fisher information does not guarantee the asymptotically correct rank of the matrix defining the quadratic form M_2. A similar limitation does not arise for S_2, because the limiting variance-covariance matrix and consistent estimates of it based on the expected or the observed Fisher information usually have full rank in the interior of the parameter space.

Turning to M_1, let $B(\xi)$ be the matrix

$$B(\xi) = \Sigma_1^{-1} - \Sigma_1^{-1}A_1\left\{A_1'\Sigma_1^{-1}A_1\right\}^+A_1'\Sigma_1^{-1} \tag{25}$$

evaluated at ξ. Using standard results on the Moore-Penrose inverse (e.g., Magnus & Neudecker, 1988, chap. 2), it is not difficult to show that for each ξ, $\Sigma(\xi)$ is a generalized inverse of $B(\xi)$ and that $\text{rank}(B(\xi)) = J^* - \text{rank}(A_1(\xi))$. Hence, $Y = B(\xi_0)\sqrt{T}\boldsymbol{\delta}_1$ is asymptotically normally distributed with mean zero and

variance-covariance matrix $B(\xi_0)\Sigma(\xi_0)B(\xi_0) = B(\xi_0)$. It follows from the fact that $\Sigma(\hat{\xi})$ is a consistent estimate of $\Sigma(\xi_0)$, by using Theorem 2a in Moore (1977), that $Y'\Sigma(\hat{\xi})Y$ is asymptotically distributed as χ^2 with $J^* - \text{rank}(A_1(\xi_0))$ degrees of freedom. If $B(\hat{\xi})$ is a consistent estimate of $B(\xi_0)$, $M_1 = (\sqrt{T}\boldsymbol{\delta}_1)'B(\hat{\xi})\Sigma(\hat{\xi})B(\hat{\xi})(\sqrt{T}\boldsymbol{\delta}_1)$ is asymptotically equivalent to $Y'\Sigma(\hat{\xi})Y = (\sqrt{T}\boldsymbol{\delta}_1)'B(\xi_0)\Sigma(\hat{\xi})B(\xi_0)(\sqrt{T}\boldsymbol{\delta}_1)$, from which the asymptotic χ^2-distribution of M_1 follows. $B(\hat{\xi})$ is a consistent estimate of $B(\xi_0)$, if the Moore-Penrose inverse $(A_1\mathcal{J}^{-1}A_1')^+$ is a continuous function of ξ_0. This is true if rank$(A_1(\xi))$=rank$(A_1\mathcal{J}^{-1}A_1')$ is constant in a neighbourhood of ξ_0 (Magnus & Neudecker, 1988, chap. 8.5): For $C = 1$, rank$(A_1(\xi))$ is simply the number of non-redundant parameters q for any ξ in the interior of the parameter space, and M_1 has $J^* - q$ degrees of freedom; for $C > 1$, the number of parameters will frequently be larger than J^*, and $A_1\mathcal{I}A_1'$ will then have full rank J^* in a neighbourhood of any ξ_0 that satisfies the above-mentioned strong identifiability condition. An exception to this rule is given for models with a core model that imposes the same equality restrictions on the expected category counts in each latent class. The rank of A_1 will then be smaller than J^* by the number of independent such restrictions. In the former case, M_1 will follow a degenerate distribution with $df = 0$ that places all of its probability mass on zero; in the latter case, M_1 will follow a non-degenerate χ^2-distribution.

Note finally that for the analogously constructed S_1, the numbers of parameters q will usually be smaller than $\frac{1}{2}J^*(J^*+1)$, implying that S_1 will follow a non-degenerate χ^2-distribution. In particular, if the model parameters are uniquely identified from the predicted variances and covariances of the person-wise category counts, rank(A_2) equals q, and S_1 will have $\frac{1}{2}J^*(J^*+1) - q$ degrees of freedom.

Author Note

Karl Christoph Klauer, Psychologisches Institut.

The author thanks Bill Batchelder, Edgar Erdfelder, Thorsten Meiser, and Christoph Stahl for helpful comments on a previous version of this paper. The author is also grateful to Edgar Erdfelder for making available the data set analyzed in this paper.

Correspondence concerning this article should be addressed to K. C. Klauer at the Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany. Electronic mail may be sent via Internet to christoph.klauer@psychologie.uni-freiburg.de.

Footnotes

[1]Where closed-form estimates are available for core-model estimation, inputting the m_{kjc} into the estimation equations will immediately yield the desired parameter estimates $\boldsymbol{\theta}_c^{(m+1)}$.

[2]It should be noted, however, that the observed information matrix will usually be a poorer estimate of the asymptotic variance-covariance matrix than the expected information matrix so that the latter should be used whenever it is available.

[3]The variance of $\hat{u}$ is obtained from the asymptotic variances and covariances of n_{12} and n_{13} that are involved in its estimation by means of the fact that the asymptotic variance of $\sqrt{TN_1}\frac{2n_{12}}{2n_{12}+n_{13}}$ is given by the following equation (Rao, 1973, chap. 6a):

$$\begin{aligned} & \left[\frac{\partial}{\partial_x}\left(\frac{2x}{2x+y}\right)\right]^2 |_{(x,y)=(p(C_{12}),p(C_{13}))}\sigma^2\left(\frac{n_{12}}{\sqrt{TN_1}}\right) \\ +2 \quad & \frac{\partial}{\partial_x}\left(\frac{2x}{2x+y}\right)|_{(x,y)=(p(C_{12}),p(C_{13}))}\frac{\partial}{\partial_y}\left(\frac{2x}{2x+y}\right)|_{(x,y)=(p(C_{12}),p(C_{13}))}\sigma\left(\frac{n_{12}}{\sqrt{TN_1}},\frac{n_{13}}{\sqrt{TN_1}}\right) \\ + \quad & \left[\frac{\partial}{\partial_y}\left(\frac{2x}{2x+y}\right)\right]^2 |_{(x,y)=(p(C_{12}),p(C_{13}))}\sigma^2\left(\frac{n_{13}}{\sqrt{TN_1}}\right), \end{aligned}$$

where σ^2 and σ denote the asymptotic variances and covariances, respectively, of the terms in parentheses.

[4]For example, $p(C_{12}) = E_{(u,c)}[(1-c)u^2] = E_c[1-c]E_u[u^2] =$
$(1-\mu_c)((E_u[u])^2 + \sigma^2(u)) = (1-\mu_c)(\mu_u^2 + \mu_u(1-\mu_u)\gamma_u)$.

Table 1

Parameter Estimates and Standard Errors (SE) of the Latent-Class Pair-Clustering Model with Two Classes

		Trial 1			Trial 2		
Class	λ	$c^{(1)}$	$r^{(1)}$	$u^{(1)}$	$c^{(2)}$	$r^{(2)}$	$u^{(2)}$
1	.21	.68	.45	.36	.82	.68	.63
(SE)	.05	.07	.07	.06	.04	.05	.06
2	.79	.23	.35	.15	.45	.44	.28
(SE)	.05	.12	.19	.02	.06	.06	.03

Table 2

Probablities of Rejection in Percent for Data Analyzed with $C = 1$ and $C = 2$ Generated by H_1 and H_2

Nominal α	M_1	*df*	M_2	*df*	M_3	*df*	S_1	*df*	S_2	*df*
— Data Generated from H_1, Analyzed with $C = 1$ —										
.10	10.0	1	10.0	1	10.5	1	10.9	7	11.5	10
.05	4.6	1	4.6	1	5.4	1	6.0	7	7.1	10
.01	0.8	1	0.8	1	1.0	1	2.0	7	2.5	10
— Data Generated from H_2, Analyzed with $C = 1$ —										
.10	69.1	1	69.1	1	68.2	1	99.6	7	99.7	10
.05	57.7	1	57.7	1	56.5	1	99.3	7	99.5	10
.01	35.0	1	35.0	1	33.1	1	99.0	7	99.4	10
— Data Generated from H_2, Analyzed with $C = 2$ —										
.10	0	0	9.1	1	10.3	2	9.0	3	8.0	10
.05	0	0	4.4	1	5.5	2	5.1	3	4.7	10
.01	0	0	0.6	1	1.0	2	1.4	3	1.8	10

Table 3

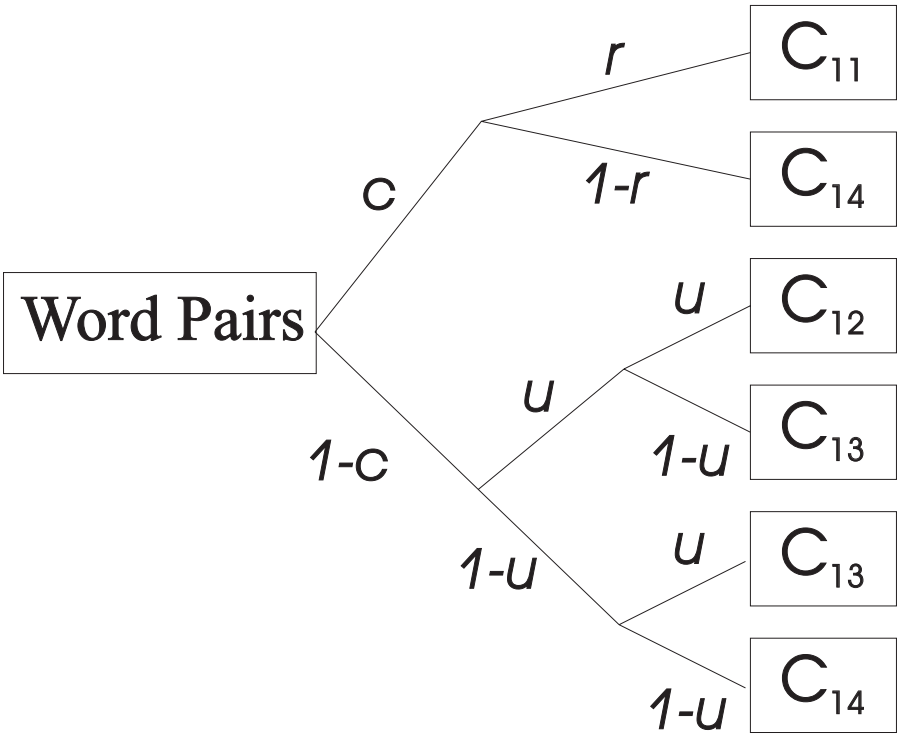
Parameter Estimates with Actual (SE_a) and Estimated (SE_e) Standard Errors in Percent

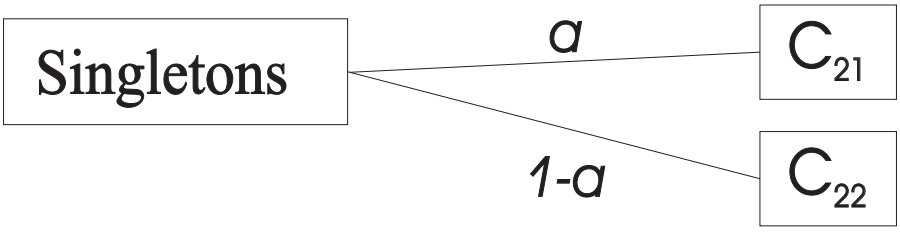
Class	λ	SE_a	SE_e	c	SE_a	SE_e	r	SE_a	SE_e	u	SE_a	SE_e
					— Data Generated from H_1, Analyzed with $C = 1$ —							
1	100	0	0	69.7	5.1	5.0	50.4	4.8	4.8	30.0	3.4	3.4
					— Data Generated from H_2, Analyzed with $C = 1$ —							
1	100	0	0	60.7	5.1	4.3	49.6	4.5	4.6	44.9	5.5	3.4
					— Data Generated from H_2, Analyzed with $C = 2$ —							
1	74.7	9.0	8.7	69.6	6.1	6.0	50.6	5.9	5.7	29.9	4.2	4.0
2	25.3	9.0	8.7	30.0	7.4	7.3	51.8	14.9	14.9	69.8	5.5	5.2

Figure Caption

Figure 1. Processing tree representation of the pair-clustering model. Parameter c is the probability of storing a word pair as a cluster, r is the probability of successful retrieval of a stored cluster, u is the probability of a successful retrieval for a member of a word pair not stored as a cluster, and a is the probability of a successful retrieval of a singleton word.

Figure A1. Asymptotic probabilities of rejecting the hypothesis $u = a$ (actual significance levels) as a function of the extent of heterogeneity γ for levels of heterogeneity $\gamma \leq 0.2$. The three functions displayed were computed, in the order from bottom to top, with $(N_1, N_2, T) = (10, 5, 50)$, $(N_1, N_2, T) = (20, 10, 50)$, and $(N_1, N_2, T) = (20, 10, 100)$. The nominal significance level α is 0.05.


Word Pairs
c
1-c
r
1-r
u
1-u
u
1-u
C11
C14
C12
C13
C13
C14


Singletons
a
1-a
C21
C22

Probability of Rejection