When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under sub-optimal conditions.

Mijke Rhemtulla[1], Patricia É. Brosseau-Liard[2], and Victoria Savalei[2]

[1]University of Kansas

[2]University of British Columbia

Author Note

This research was supported in part by a graduate fellowship from the Social Sciences and Humanities Research Council (SSHRC) to P. Brosseau-Liard, a SSHRC Banting Postdoctoral Fellowship to M. Rhemtulla, and a SSHRC grant to V. Savalei.

Correspondence concerning this article should be addressed to Mijke Rhemtulla, Center for Research Methods and Data Analysis, University of Kansas, 1425 Jayhawk Blvd. Watson Library, 470. Lawrence, KS 66045. E-mail: mijke@ku.edu

Abstract

A simulation study compared the performance of robust normal theory maximum likelihood (ML) and robust categorical least squares (cat-LS) methodology for estimating confirmatory factor analysis models with ordinal variables. Data were generated from 2 models with 2-7 categories, 4 sample sizes, 2 latent distributions, and 5 patterns of category thresholds. Results revealed that factor loadings and robust standard errors were generally most accurately estimated using cat-LS, especially with fewer than 5 categories; however, factor correlations and model fit were assessed equally well with ML. Cat-LS was found to be more sensitive to sample size and to violations of the assumption of normality of the underlying continuous variables. Normal theory ML was found to be more sensitive to asymmetric category thresholds, and was especially biased when estimating large factor loadings. Accordingly, we recommend cat-LS for datasets containing variables with fewer than 5 categories, and ML when there are 5 or more categories, sample size is small, and category thresholds are approximately symmetric. With 6-7 categories, results were similar across methods for many conditions; in these cases, either method is acceptable.

Keywords:

CFA, Categorical Indicators, Confirmatory Factor Analysis, Maximum Likelihood, Categorical Least-Squares, Robust Statistics

Confirmatory factor analysis (CFA) models are among the most popular types of structural equation models (SEMs) in psychology (Crowley & Fan, 1997; Marsh & Hau, 2007; Martens, 2005). A CFA model hypothesizes linear relationships between the observed variables y and the latent factors f, specified by a set of latent regression equations $y = \Lambda f + \varepsilon$, where Λ is the matrix of factor loadings and ε is an error term containing both measurement error and item-specific variance in the observed indicators.

Classic estimation methods in SEM, such as normal theory maximum likelihood (ML), are based on the assumption that the observed variables are measured on a continuous scale. When both indicators and factors are assumed to have continuous distributions, a standard CFA model that hypothesizes a set of linear regressions of observed indicators on latent factors is plausible. However, researchers often have to work with observed variables that can only take a limited number of values. This is especially true in the social sciences, where psychological constructs such as attitudes are frequently measured on Likert scales (e.g., “strongly disagree, disagree, agree, strongly agree”) and cognitive tests are frequently measured with binary (correct/incorrect) responses.

Because responses on ordinal variables are typically coded numerically in ascending order, it is easy for researchers to ignore the categorical nature of the variables and to treat them as continuous, applying continuous normal theory ML to estimate model parameters. However, this approach can lead to biased parameter estimates, as well as incorrect standard errors and model test statistics, especially when the number of categories is small (e.g., Johnson & Creech, 1983). This is because the standard continuous CFA model is fundamentally misspecified when applied to ordinal variables, which cannot be linear functions of continuous factors. This bias becomes smaller as the number of categories becomes larger, because the variables approach continuity.

The theoretically correct alternative to normal theory ML is to treat ordinal variables directly as ordinal. In order to reconcile the linear CFA model with the ordinal nature of the variables, one of two assumptions must be made: a) that underlying each categorical variable y is a normally distributed continuous variable y^{*1}, and the CFA model describes the relationship between y^{*} and the latent factors f (e.g., Muthén, 1993; Muthén, du Toit, & Spisic, 1997) or b) that the model relating the probability of the observed responses y to the values of the latent factors f is a generalized linear model with an ordered probit (or an appropriately scaled logit) link function (e.g., Baker & Kim, 2004; Bartholomew & Knott, 1999; Skrondal & Rabe-Hesketh, 2004). The first set of assumptions is more commonly made for categorical SEM methods, which are limited-information methods that rely on polychoric correlation estimates. The second set of assumptions is primarily used for item response theory (IRT) methods, which are full information methods using raw categorical variables. However, these two sets of assumptions are mathematically equivalent (Muthén & Asparouhov, 2002; Takane & de Leeuw, 1987; Wirth & Edwards, 2007). For instance, a correct/incorrect response on a test item can be thought of as a so-called artificial dichotomy (Pearson, 1901, 1904) because one can hypothesize that underlying this response is the continuous variable of topic knowledge. Equivalently, one can posit that the probability of the correct response on this test item depends on the person’s score on the underlying latent factor of “ability” via the probit link function. In psychology, the existence of underlying continuous variables is a common assumption when analyzing categorical variables, and this is the paradigm adopted in the present paper. However, this assumption is not strictly necessary to apply categorical methodology, and a probit link function assumption can be invoked instead (Muthén, 2003; Muthén & Asparouhov, 2002).

There are several different methods for fitting latent variable models to data containing ordinal variables. As already mentioned, the most significant distinction is between *full information* and *limited information* methods (e.g., Maydeu-Olivares & Joe, 2005). Full information methods model the entire multivariate categorical distribution of the observed variables, or, equivalently, they use subjects' entire response pattern to extract information about model parameters. (MML; Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988; Bock & Lieberman, 1970). Since the models are equivalent (Takane & de Leeuw, 1987), conversion formulas from IRT parameters to CFA parameters are readily available. Wirth and Edwards (2007) provide an excellent overview of modern developments and challenges in the area of full information categorical estimation methods.

Limited information methods, on the other hand, only use the low order margins (typically, univariate and bivariate frequencies) to estimate model parameters, losing information available in higher-order frequencies (e.g., the joint probability of three variables at once is not modeled). These methods typically make the "underlying continuous variable" assumption. They proceed by first estimating thresholds (values that divide the continuous distribution of y^* into categories to obtain the categorized variable y) and polychoric correlations (estimates of the correlations among the continuous variables y^*), either separately or simultaneously (Christoffersson, 1975; Lee, Poon, & Bentler, 1990, 1995; Muthén 1978, 1984; Olsson, 1979a) The CFA model is then fit to the matrix of polychoric correlations (Jöreskog, 1994; Lee, Poon, & Bentler, 1990, 1995; Muthén 1984, 1993).

Full information methods have a theoretical advantage over limited information methods, in that they produce more efficient parameter estimates because they use all of the information available in the data (Maydeu-Olivares & Joe, 2005, 2006; Joe & Maydeu-Olivares, 2011). However, this theoretical advantage does not appear to often translate into a practical advantage. In practice, the advantage of full information methods is slight at best, and several studies show that limited information methods perform as well or better in practice. For binary items, a frequently cited reference is Knol & Berger (1991), who state that, "for multidimensional data a common factor analysis on the matrix of tetrachoric correlations performs at least as well as the theoretically appropriate multidimensional item response models" (p. 457). Forero and Maydeu-Olivares (2009) conducted the most recent and thorough Monte Carlo comparison of full- and limited-information methods to date, and found that the differences between the methods were negligible in most conditions, with limited-information methods producing slightly better parameter estimates and full-information methods producing slight better standard errors. In their excellent review of the relevant literature, they also note that their finding is consistent with most previous research.

The research question explored in the present study is as follows: in the context of CFA model estimation with ordinal data, how many categories are necessary before continuous methodology performs as well categorical methodology? In this study, a limited information categorical method is used for comparison to the continuous methodology. Given small differences between limited- and full-information methods this choice is unlikely to affect our conclusions about the number of categories for which continuous methodology begins to perform comparably to categorical methodology. There are additional practical reasons to prefer limited information estimation. The computational burden of full information methods can be substantial, particularly when there is more than one latent factor. Full information categorical methodology is not available in all SEM programs (to our knowledge, only in M*plus* and Mx), and thus psychology researchers unfamiliar with IRT software may not have access to it. Finally, full information model test statistics tend to perform extremely poorly given the frequently sparse nature of the full contingency tables (Maydeu-Olivares & Joe, 2005, 2006; Joe & Maydeu-Olivares, 2011)[2]. In contrast, if continuous and limited-information categorical methodology is accompanied by the appropriate robust corrections to standard errors and test statistics, as is done in this study, test statistics are easily available for both continuous and limited-information categorical methods (Maydeu-Olivares & Joe, 2005, 2006; Satorra & Bentler, 1994; Muthén, 1993), and they are directly comparable (their degrees of freedom are the same). We now describe the particular methods studied in more detail.

Limited Information Categorical Estimation Methods with Robust Corrections

The limited information categorical method used in this study is the categorical least squares (cat-LS) method with robust corrections to standard errors and test statistics. Related limited information methods are also described for context. As already mentioned, limited information methods typically proceed by first estimating variables' thresholds and polychoric correlations. Parameter estimates of the CFA model are then obtained by fitting this model directly to the matrix of polychoric correlations. The estimation procedure used to fit the CFA model to the matrix of polychoric correlations distinguishes the various limited information methods and determines whether corrections are necessary to the default standard errors and test statistic.

The oldest in this family of methods, categorical weighted least squares (cat-WLS), uses the inverse of the estimated covariance matrix of polychoric correlations, $\hat{V}^{-1}$, as the weight matrix in estimation (e.g., Muthén, 1978, 1984). Let ρ (θ) represent the structure of the correlation matrix[3] under the CFA model, and let r represent the p (p-1)/2×1 vector of polychoric correlation estimates, where p is the number of variables. Then, cat-WLS obtains CFA model parameter estimates $\hat{\theta}$ by minimizing $F_{cat-WLS} = (r - \rho(\theta))'\hat{V}^{-1}(r - \rho(\theta))$. Standard errors are obtained from the diagonals of the asymptotic covariance matrix of $\hat{\theta}$, given by $(\hat{\Delta}'\hat{V}^{-1}\hat{\Delta})^{-1}$, where $\hat{\Delta}$ is the matrix of model derivatives evaluated at the parameter estimates. The test statistic is $T_{cat-WLS} = (N-1)F_{cat-WLS}(\hat{\theta})$, where N is sample size; this statistic is asymptotically chi-square distributed when the model is true. Note that different software programs may implement this approach slightly differently (Jöreskog, 1994; Lee, Poon, & Bentler, 1990, 1995; Muthén, 1984, 1993; Muthén et al., 1997). The cat-WLS method is asymptotically efficient within the class of categorical data methods relying on polychoric correlations[4]. Cat-WLS is conceptually similar to the asymptotically distribution free (ADF) estimation method for continuous data (Browne, 1984), and similarly breaks down unless the sample size is very large (DiStefano, 2002; Dolan, 1994; Flora & Curran, 2004; Hoogland & Boomsma, 1998; Lei, 2009; Maydeu-Olivares, 2001; Potthast, 1993; Yang-Wallentin, Joreskog & Luo, 2010).

The recommended limited information categorical estimation methods for small to medium samples are diagonal weighted least squares (cat-DWLS) and least squares (cat-LS, a.k.a., unweighted least squares)[5]. These methods are more stable than cat-WLS in smaller samples because they do not require inverting $\hat{V}$ (Flora & Curran, 2004; Maydeu-Olivares, 2001). Because both of these methods are less efficient than cat-WLS, they require robust corrections to the standard errors and test statistics (Satorra & Bentler, 1994; Muthén, 1993; Muthén et al., 1997). Cat-LS obtains model parameter estimates $\tilde{\theta}$ by minimizing the LS fit function $F_{cat-LS} = (r - \rho(\theta))'(r - \rho(\theta))$. Default standard errors are no longer accurate and the correct covariance matrix of $\tilde{\theta}$ is the robust covariance matrix: $(\tilde{\Delta}'\tilde{\Delta})^{-1}\tilde{\Delta}'\hat{V}\tilde{\Delta}(\tilde{\Delta}'\tilde{\Delta})^{-1}$. The "naïve" test statistic $T_{cat-LS} = (N-1)F_{cat-LS}(\tilde{\theta})$ is also no longer appropriate, as it is not chi-square distributed. Two popular adjustments to the test statistic exist: the mean-corrected (scaled) chi-square and the mean-and-variance corrected (adjusted) chi-square (Satorra & Bentler, 1994; Muthén, 1993). With categorical variables, the adjusted chi-square appears to perform better (Maydeu-Olivares, 2001). For cat-LS, the adjusted chi-square is computed as $T_{cat-MV} = \frac{tr(\tilde{U}\hat{V})}{tr(\tilde{U}\hat{V}\tilde{U}\hat{V})} T_{cat-LS}$, where $\tilde{U} = I - \tilde{\Delta}(\tilde{\Delta}'\tilde{\Delta})^{-1}\tilde{\Delta}'$. It is referred to a chi-square distribution with $k \approx \frac{[tr(\tilde{U}\hat{V})]^2}{tr(\tilde{U}\hat{V}\tilde{U}\hat{V})}$ degrees of freedom, rounded to the nearest integer. The mean and variance of this statistic match those of a chi-square distribution with *k* degrees of freedom when the CFA model is correct[6]. Note that the robust corrections still require the computation of $\hat{V}$, but not its inverse. The equation for the mean-corrected chi-square T_{cat-M} is omitted as we do not study it in the present paper.

The cat-DWLS estimator is similar to cat-LS except that it minimizes a weighted sum of squares, $F_{cat-DWLS} = (r - \rho(\theta))'\hat{D}^{-1}(r - \rho(\theta))$, where $\hat{D}$ is a diagonal matrix with elements of $\hat{V}$ on the diagonal. Robust corrections for cat-DWLS are similarly defined; these formulas are not presented here[7]. Recent evidence suggests that cat-LS performs better than cat-DWLS in terms of estimated parameter values, standard errors, and coverage, although it may exhibit lower convergence rates in some conditions (Forero, Maydeu-Olivares, and Gallardo-Pujol, 2009). Differences between cat-LS and cat-DWLS are generally small (Yang-Wallentin et al., 2010).

It is worth emphasizing that the robust corrections to standard errors that accompany cat-LS and cat-DWLS methods are corrections for loss of efficiency due to the fact that full cat-WLS estimation was not performed. It is sometimes mistakenly stated that robust corrections accompanying cat-LS and cat-DWLS allow for relaxing of the assumption that underlying observed categorical variables are a set of normally distributed continuous variables, that is, that robust corrections somehow adjust for the possible nonnormality in the underlying continuous variables (Savalei, in press). This is false; robust corrections accompanying categorical estimators still require the assumption that the underlying continuous variables are normally distributed (or equivalently, that observed categorical variables and latent factors can be connected via ordered probit regression). The confusion likely stems from the fact that the original development of the robust corrections (Satorra & Bentler, 1988; 1994) was to adjust for loss of efficiency in the continuous normal theory ML estimator due to nonnormality in the data.

Continuous Normal Theory ML with Robust Corrections

In this study, we compare cat-LS with robust corrections to the most popular continuous data estimation method, normal theory ML, accompanied by robust corrections for nonnormality (Satorra & Bentler, 1994). It is a good practical strategy to apply robust corrections to normal theory ML when variables are categorical, because these variables are, by definition, nonnormal. Any categorical variable is nonnormal by virtue of being discrete rather than continuous. Categorical variables are likely to produce nonzero kurtosis estimates, depending on the frequency of the middle categories, and category asymmetry will further lead to nonzero skewness. Robust ML is a covariance structure method, and thus begins with obtaining the sample covariance matrix of the data, *S*, ignoring the categorical nature of the data. Standardizing *S* would yield the matrix of Pearson product-moment correlations. Let $\Sigma(\theta)$

represent the structure of the covariance matrix under the CFA model. Then, normal theory ML obtains CFA model parameter estimates $\breve{\theta}$ by minimizing the fit function $F_{ML} = tr\{\Sigma^{-1}(\theta)S\} - \ln\left|\Sigma^{-1}(\theta)S\right| - p$. While this equation is not very intuitive, asymptotically (in large samples) it can be written as a quadratic form in model residuals with the weight matrix $\breve{W}$, called the normal-theory weight matrix. The correct standard errors when the data are continuous but nonnormal are obtained from the robust covariance matrix: $(\breve{\Delta}'\breve{W}^{-1}\breve{\Delta})^{-1}\breve{\Delta}'\breve{W}^{-1}\breve{\Gamma}\breve{W}^{-1}\breve{\Delta}(\breve{\Delta}'\breve{W}^{-1}\breve{\Delta})^{-1}$, where $\breve{\Gamma}$ is an estimate of the fourth-order moments matrix of the raw data (see Bentler, 2008, for a definition of the typical element of this matrix). The default test statistic, $T_{ML} = (N-1)F_{ML}(\breve{\theta})$, is no longer valid when data are nonnormal. The same two adjustments, the scaled and the adjusted test statistics, can be computed instead. With continuous nonnormal data, the mean corrected (scaled) chi-square is the most popular statistic (Satorra & Bentler, 1994). This statistic, computed as $T_{ML-M} = \dfrac{df}{tr(\breve{U}\breve{\Gamma})}T_{ML}$, where $\breve{U} = \breve{W}^{-1} - \breve{W}^{-1}\breve{\Delta}(\breve{\Delta}'\breve{W}^{-1}\breve{\Delta})^{-1}\breve{\Delta}'\breve{W}^{-1}$, is referred to a chi-square distribution with *df* degrees of freedom, although it only approximates this distribution in the mean. The mean-and-variance corrected (adjusted) chi-square for continuous nonnormal data is computed as $T_{ML-MV} = \dfrac{tr(\breve{U}\breve{\Gamma})}{tr(\breve{U}\breve{\Gamma}\breve{U}\breve{\Gamma})}T_{ML}$, and is referred to a chi-square distribution with $k \approx \dfrac{[tr(\breve{U}\breve{\Gamma})]^2}{tr(\breve{U}\breve{\Gamma}\breve{U}\breve{\Gamma})}$ degrees of freedom, rounded to the nearest integer (Muthén, 1993).

Normal theory ML with robust corrections for nonnormality is based on the assumption that observed data are continuous, albeit nonnormal. Parameter estimates $\breve{\theta}$ obtained by minimizing F_{ML} will be biased downward when the data are categorical, because a variable with a limited number of possible values (e.g., the integers 1 through 4) will necessarily be more weakly related to a latent factor than if it were measured continuously (Bollen & Barb, 1981; Olsson, 1979b). Thus, under the assumption that continuous variables underlie the observed categorical variables, the matrix of Pearson product-moment correlations will be an underestimate of the correlation matrix among the underlying continuous variables. Equivalently, under the alternative assumption that each categorical variable is related directly to the latent factors via an ordered probit (or logit) link function, negatively biased estimates will result from the fact that continuous methodology assumes a linear relationship between the variables and the factors, whereas the true link is probit. Thus, given either assumption, the relation between variables and factors is misspecified, and the resulting estimates will be biased. As the number of categories per variable increases, variables approach continuity and this bias decreases.

Goals of the Present Study

Researchers often use continuous methods such as normal theory ML in spite of the variables' categorical nature. While it is theoretically incorrect to do this, researchers usually work under the assumption that, given a sufficiently large number of categories, categorical variables are sufficiently similar to continuous variables to produce good results. While several studies have explored the question of how many categories are enough to treat categorical variables as continuous, the advent of robust corrections for both continuous and categorical estimation warrants a reassessment of this issue. No study has yet compared the performance of continuous and categorical estimation methods with their respective robust corrections, and a thorough investigation of this question will allow researchers to decide which of the most current methods is best for their data.

The main goal of the present study is to provide this much-needed comparison. We compare robust ML, a continuous methodology with corrections for nonnormality that is widely used and performs well under a variety of circumstances, to robust cat-LS, one of the best currently available categorical methodologies (Forero et al., 2009; Yang-Wallentin, et al., 2010) that provides correct standard errors and test statistics. A secondary aim of our investigation is to evaluate the relative performance of the two methods in conditions that generally pose difficulties for estimation or violate the underlying assumptions of both methods. To this end, we included a range of conditions including different sample sizes, model sizes, and varying levels of category threshold asymmetry. Additionally, categorical variables were generated by categorizing underlying normal as well as nonnormal distributions. In the condition where the underlying continuous variables are nonnormal, cat-LS should also result in biased parameter estimates. The comparison between cat-LS and ML is particularly interesting in this case, as both methods are wrong but one may do better than the other. We compare the relative performance of cat-LS and normal theory ML parameter estimates, the quality of robust standard errors, and the rejection rates of the adjusted test statistics. The results of this investigation will provide an answer to the question of "how many categories are enough to treat data as continuous" that is sensitive to the characteristics of a particular dataset.

Literature Review

Relative performance of continuous and categorical estimators

A number of studies have examined the performance of continuous and/or categorical estimators with ordinal data. We first summarize studies that did not apply robust corrections to standard errors and test statistic. Only the quality of parameter estimates is therefore relevant to the present investigation.

Several studies that have included continuous ML estimation (of either product-moment correlation matrices or covariance matrices) have found parameter estimates to be underestimated when the number of categories is very small (e.g., 2-3). This bias tends to diminish as the number of categories increases, such that when the number of categories reaches four or five, most studies pronounce ML parameter estimates to be accurate (Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Johnson & Creech, 1983; Muthén & Kaplan, 1985; but see Babakus, Ferguson, & Jöreskog, 1987, and Dolan, 1994, for evidence that even 5 categories may not be enough to eliminate bias in continuous ML estimation). In addition, a number of studies have assessed the performance of the limited information categorical methods cat-DWLS and cat-LS with ordinal data; these have tended to focus on variables with 2 or 5 categories. These studies have found that in most situations, both methods lead to unbiased parameter estimates (Beauducel & Herzberg, 2006; Dolan, 1994; Flora & Curran, 2004; Forero et al., 2009; Lei, 2009; Muthén et al., 1997; Nussbeck, Eid & Lischetzke, 2006; Rigdon & Ferguson, 1991; Yang-Wallentin et al., 2010).

Because several of the studies mentioned above were conducted before robust corrections were readily available, they found, predictably, that uncorrected or “naïve” standard errors were too small and the uncorrected chi-square test statistic was inflated, for both uncorrected ML and uncorrected categorical methodology. More recent studies have investigated the performance of cat-DWLS and cat-LS with robust corrections to standard errors. These have reported very little bias in estimated standard errors (Flora & Curran, 2004; Forero et al., 2009; Lei, 2009; Maydeu-Olivares, 2001; Nussbeck et al., 2006; Yang-Wallentin et al., 2010). Studies that have compared cat-DWLS to cat-LS have either reported no difference (Yang-Wallentin et al., 2010) or a slight advantage of cat-LS over cat-DWLS (Forero et al., 2009; Maydeu-Olivares, 2001). To our knowledge, only one study has reported standard errors resulting from robust continuous ML estimation of ordinal items (DiStefano, 2002). This investigation found that robust ML standard errors were negatively biased to a much smaller degree than their non-robust counterparts with 5-category ordinal data; however, considerable bias still appeared in certain conditions.

When it comes to robust test statistics, Maydeu-Olivares (2001) compared the performance of $T_{cat\text{-}M}$ and $T_{cat\text{-}MV}$, and found $T_{cat\text{-}MV}$ to outperform $T_{cat\text{-}M}$ in small samples (N = 100) with both cat-DWLS and cat-LS. Several studies have found that $T_{cat\text{-}MV}$ based on the cat-DWLS estimator performs well with 2- and 5-category data and a sample size of at least 200 (Flora & Curran, 2004; Lei, 2009; Nussbeck et al., 2006; Muthén et al., 1997). There is some evidence that $T_{cat\text{-}MV}$ as used with cat-DWLS may be robust to mild departures from underlying normality (Flora & Curran, 2004), and that it is not affected by category asymmetry (Lei, 2009; Nussbeck et al., 2006; Muthén et al., 1997; Yang-Wallentin et al., 2010).

Two studies have examined the performance of $T_{ML\text{-}M}$ with continuous ML estimation (DiStefano, 2002; Green, Akey, Fleming, Hershberger, & Marquis, 1997). Both of these investigations reported markedly better results with the scaling correction than without it; however, in many situations even the scaled test statistic did not adequately control type I error. In particular, with fewer than 4 categories and underlying nonnormal continuous distributions, $T_{cat\text{-}M}$ failed to control type I error (Green et al., 1997).

Beauducel and Herzberg (2006) described the only study that systematically compared continuous and categorical methodologies at a range of numbers of categories. These authors assessed the relative performance of continuous normal theory ML and cat-DWLS when CFA models were fit to data, varying model size (1 factor, 5 indicators to 8 factors, 40 indicators), sample size (from 250 to 1000) and number of categories (2 to 6). Results of this study tended to favor categorical methodology for parameter estimation, except for the estimation of factor correlations, which were consistently slightly overestimated with cat-DWLS. The findings regarding standard errors and test statistics are limited by the fact that robust corrections were applied to cat-DWLS but not to continuous ML. Robust Cat-DWLS standard errors were smaller than those produced by ML; however, neither type of standard errors was compared to the respective empirical standard deviation of parameter estimates to evaluate accuracy. Uncorrected ML type I error rates were consistently too high. Without a direct comparison of the performance of robust standard errors and test statistics with *both* methods, it is difficult to decide whether the small advantage of the categorical method in factor loading estimation is sufficient to advocate its use with 5 or more categories.

Factors Influencing Estimation

Finally, there are several factors that may especially influence the relative performance of categorical and continuous methods applied to ordinal data, such as variable category thresholds, whether the continuous distributions underlying the observed categorical variables are normal, and model size. These factors are manipulated in the present study. Research findings that address the influence of these factors on estimation are summarized below.

Threshold variability. Several researchers have evaluated the robustness of estimation techniques to the variability of thresholds used to categorize underlying normal data. With both categorical and continuous estimation, parameter estimates tend to be less accurate when category thresholds are distributed asymmetrically around the mean (Babakus et al., 1987; DiStefano, 2002; Dolan, 1994; Lei, 2009; Rigdon & Ferguson, 1991). There is some evidence that continuous estimation methods may be more affected by threshold asymmetry (Babakus et al., 1987; DiStefano, 2002); the correlation pattern created by items with varying thresholds can result in spurious "difficulty factors" and create model misfit (Bernstein & Teng, 1989; Ferguson, 1941). However, varying threshold conditions have received little attention in the literature, and it is presently unknown exactly how much impact varying thresholds have on model estimation.

Several studies have found that uncorrected standard errors and test statistics for cat-ML, cat-WLS, and cat-DWLS are affected by asymmetrically-distributed thresholds (e.g., Babakus et al. 1987; Dolan, 1994; Potthast, 1993; Rigdon & Fergusson, 1991), but this finding has no bearing on the performance of robust standard errors and test statistics. Forero et al. (2009), found that highly asymmetric thresholds (e.g., 2-category data where more than 90% of the distribution fell into one category) led to negatively biased robust standard error estimates with cat-DWLS and to a much lesser extent, cat-LS. Lei (2009) found that category asymmetry had little effect on cat-DWLS robust standard errors, but it led to higher type I error rates for cat-DWLS versions of $T_{cat\text{-}M}$ and $T_{cat\text{-}MV}$. Yang-Wallentin et al. (2010) used mildly asymmetrical category distributions and reported that category asymmetry made no difference to parameter estimates, robust standard errors, or robust test statistics under cat-LS or cat-DWLS. Finally, Green et al., (1997) studied the performance of $T_{ML\text{-}M}$ with asymmetric category distributions and found its performance to be good across a wide range of conditions, except when different items had widely varying threshold distributions. Collectively, these studies suggest that continuous and categorical estimation techniques may be robust to mildly asymmetric threshold values but not extremely asymmetric ones.

Underlying normality. When an ordinal item is obtained by categorizing a nonnormal variable, it becomes impossible to recover the underlying distribution[8]. For example, a trait such as *sadness* may be positively skewed, but if the binary item, “I feel sad” is endorsed by 50% of the sample, there is no way to distinguish a skewed distribution with a low threshold from a normal distribution with a threshold at its center. For this reason, categorical estimation is typically implemented under the assumption of underlying bivariate normality (or equivalently, of the probit link function connecting the probability of each category and the latent factors). When this assumption is violated, categorical methods such as cat-LS and cat-DWLS may produce biased parameter estimates, which will also affect standard errors and test statistics.

Flora & Curran (2004) manipulated the type of continuous distribution underlying the observed ordinal variables. They imposed 1 or 4 symmetric thresholds (to create 2- or 5-category distributions, respectively) on simulated continuous distributions that were either normal or had skewness of up to 1.25 and kurtosis of up to 3.75. Using cat-DWLS with robust corrections, they found that the estimated polychoric correlations and model parameter estimates displayed increasing positive bias as the underlying distribution became more skewed, though the absolute level of bias remained low. In addition, the cat-DWLS version of $T_{cat\text{-}MV}$ had slightly higher type I error rates with increasing skew. This study did not, however, include continuous estimation methods, so we cannot gauge the relative performance of the two types of methods when the underlying normality assumption is violated. Green et al. (1997) found that $T_{ML\text{-}M}$ was robust to most departures from underlying normality, including both underlying uniform and skewed distributions. The one exception to this was when some items had an underlying positive skew and others had an underlying negative skew; in this case, $T_{ML\text{-}M}$ tended to be too high. These authors did not investigate parameter estimates or standard errors, nor did they examine categorical estimation methods.

Model Size. Most of the research cited above has manipulated model size. The most pervasive effect of model size is simply that larger models are harder to fit, resulting in higher rates of non-convergence and improper solutions (e.g., Flora & Curran, 2004; Yang-Wallentin et. al., 2010; but see Forero et al., 2009). Some research (e.g., Flora & Curran, 2004; Potthast, 1993) has found that larger models result in too-high type I error rates and underestimated standard errors. Large models are difficult to estimate when the sample size is not sufficiently large. However, Beauducel and Herzberg (2006) report a case where both continuous ML and cat-DWLS were relatively successful at estimating an 8-factor, 40-indicators model with as sample size as low as 250. One aim of the present study is to compare categorical and continuous methods applied to ordinal data with small samples (*N* = 100 to 150) and for different model sizes (10 and 20 indicators).

In summary, few studies have systematically varied the number of categories when comparing the performance of categorical and continuous methods, and none have investigated the relative performance of robust corrections applied to both continuous and categorical methods. Furthermore, no study has assessed the relative performance of continuous and categorical methods in the more extreme conditions of small samples, violations of underlying normality, and asymmetric thresholds. The present study aims to fill these gaps in the literature. We compare cat-LS and normal theory ML estimation of CFA models with categorical variables ranging from 2-7 categories under all of the above described conditions, examining parameter bias and efficiency as well as the performance of test statistics. The results of this investigation lead to recommendations to researchers regarding the choice of continuous vs. categorical methods for ordinal data.

Method

To compare the performance of continuous ML and cat-LS methods with ordinal data, we simulated 1000 data sets for each of 480 conditions, formed by fully crossing the following five factors:

- CFA model size (2 levels: 10 indicators or 20 indicators)

- Underlying distribution (2 levels: normal, nonnormal)
- Number of categories (6 levels: 2-7 categories)
- Threshold symmetry (5 levels: Symmetry; Moderate Asymmetry, Moderate Asymmetry-alternating, Extreme Asymmetry, Extreme Asymmetry-alternating)
- Sample size (4 levels: *N* = 100, 150, 350, 600)

CFA Model Size

Model 1 was a 2-factor CFA model with 5 indicators per factor, for a total of 10 indicators. Factor loadings for the 5 indicators of each factor were .3, .4, .5, .6, and .7. These values have been used in previous simulation studies (e.g., Beauducel & Herzberg, 2006; DiStefano, 2002; Flora & Curran, 2004). The factor correlation was set to .3. The model was identified by fixing the variance of each latent variable to 1. Generated continuous variables had unit variance (prior to categorization). Model 2 was identical to Model 1, but with 10 indicators per factor. Model 1 had 34 degrees of freedom; Model 2 had 169 degrees of freedom. Note that for Model 2, the degrees of freedom are greater than the two smallest studied sample sizes, and estimation might be particularly difficult in these conditions (e.g., Yuan & Bentler, 1998).

Underlying Distribution

Categorical variables were generated by categorizing continuous ones. The continuous variables were either normal or nonnormal with univariate skew of 2 and kurtosis of 7. Previous research by Flora and Curran (2004) found that categorical estimation methods were fairly robust to moderate levels of underlying nonnormality (in their case, univariate skew of up to 1.25, and univariate kurtosis up to 3.75); we sought to extend this finding by selecting a more extreme level of nonnormality. The shape of the underlying distribution is predicted to affect cat-LS parameter estimates, because this method assumes that the continuous distributions underlying the observed ordinal data are normal (or equivalently, that the probit link function correctly describes the relationship between the categorical variables and their underlying latent factors). Continuous ML estimation may be affected by underlying nonnormality as well. Keeping the thresholds used for categorization the same but changing the shape of the underlying distribution that is being categorized changes the relative frequencies of each observed category in the resulting categorical variable. This can introduce additional skewness or kurtosis into observed categorical variables, making them more “nonnormal”, or it can balance opposite-direction skewness or kurtosis, making the resulting categorical variables more “normal”. Additionally, the new relative frequencies may be such that the relationship between the observed categorical variables and the latent factors is even more nonlinear, leading to greater bias when a linear latent regression model (i.e., the CFA model) is fit to data.

Number of Categories

Our chief goal was to explore a range of number of categories to see whether and at what point the categorical method would cease to produce noticeably better results than the continuous method. To this end, the continuous latent response distributions were categorized into 2, 3, 4, 5, 6, or 7 categories.

Threshold Symmetry

Since previous research has found that most methods performed worse when category thresholds were asymmetric (Babakuset al., 1987; DiStefano, 2002; Dolan, 1994; Forero et al., 2009; Lei, 2009; Potthast, 1993; Rigdon & Ferguson, 1991; cf. Yang-Wallentin et al., 2010), we included both symmetric and asymmetric thresholds.

In the Symmetry condition, category thresholds were distributed symmetrically around 0 and spaced evenly to evenly divide the distance between -2.5 and 2.5 standard deviations around the mean (e.g., for the 4-category condition, thresholds were -1.25, 0, and 1.25, resulting in 11%, 39%, 39%, and 11% of normally distributed data falling into each category). In the Moderate Asymmetry condition, category thresholds were chosen such that the peak of the distribution fell to the left of center. In the Extreme Asymmetry condition, category thresholds were created so that the lowest category would always contain the largest number of cases, and all other categories contained a much smaller (and decreasing) number of cases. A table of thresholds is available in the supplementary materials. Finally, the Moderate Asymmetry-Alternating and Extreme Asymmetry-Alternating conditions had thresholds that were identical to the two asymmetric conditions, except that the direction of the asymmetry was reversed for odd-numbered variables, simulating a situation where different items on a scale have very different levels of difficulty. This situation is expected to make it difficult for either estimation method to estimate a positive correlation.

For the underlying normal data, the expected proportion of cases falling into each category for any given set of thresholds can be analytically determined; Figure 1 depicts these frequencies in the Symmetry, Moderate Asymmetry, and Extreme Symmetry conditions (the two Alternating conditions are omitted because underlying distributions are normal and thus symmetric). For the underlying nonnormal data, the proportion of cases falling into each category is difficult to determine analytically; these proportions were estimated from 1,000,000 cases in each condition. Figure 2 depicts these proportions in the five thresholds conditions. Additionally, Table 1 presents the skew and kurtosis of the resulting categorical variables.

Sample Size

Four sample sizes were used. In psychology, sample sizes smaller than *N* = 200 are common. Thus, we included two small sample sizes (*N* = 100 and *N* =150) and two medium sample sizes (*N* = 350 and *N* = 600).

Data Generation and Analysis

Continuous data (normal and nonnormal) were generated in EQS 6.1 (Bentler, 2008) using methods developed by Fleishman (1978) and Vale and Maurelli (1983). EQS was also used to categorize the data.

Data in all cells in the study were analysed twice: 1) with covariance-based continuous ML with robust corrections for nonnormality (Satorra & Bentler, 1994), and 2) with polychoric correlation-based cat-LS with robust corrections (Muthén, 1993)[9]. Both EQS6.1 (Bentler, 2008) and M*plus* 6.11 (Muthén & Muthén, 1998-2010) were used for the analyses; however, only M*plus* results are presented in this article.

The cat-LS method as implemented in M*plus* first estimates variables' thresholds, then the matrix of polychoric correlations, then the parameter estimates of the CFA model (Muthén, 1984). The robust corrections are then computed following Muthén (1993) and Muthén et al. (1997). M*plus* automatically adds .5 to any zero-frequency cells found in bivariate contingency tables to estimate the polychoric correlation matrix (but not the asymptotic covariance matrix $\hat{V}$). EQS6.1 uses a different estimation method due to Lee et al. (1995); this method estimates thresholds and polychoric correlations simultaneously. EQS also adds .5 to any zero-frequency cells by default to estimate both polychoric correlations and the asymptotic covariance matrix $\hat{V}$. With some exceptions, M*plus* and EQS produced very similar parameter estimates and robust standard errors, and thus only M*plus* results are discussed and presented. When differences occurred they are noted in text. However, the rejection rates of the EQS and M*plus* versions of the test statistic $T_{cat\text{-}MV}$ differed, particularly when the number of categories was large. M*plus*'s $T_{cat\text{-}MV}$ performed better throughout, and thus only M*plus* results are presented[10].

Results

Results are presented in Figures 3-9. Due to an overwhelming amount of data, the results in all figures are collapsed across some conditions, and results for certain conditions are omitted. Full results are available in supplementary materials. In particular, and somewhat surprisingly, model size (10 vs. 20 indicators) had the smallest effect on the quality of parameter estimates, standard errors, and test statistics. In all figures, therefore, results are collapsed across model size. Where there are substantial differences as a function of model size, these are noted in the text. Results for sample sizes of 150 and 350 are omitted from all Figures, because, in general, results of N = 150 were similar to those with N = 100, and results with N = 350 were similar to N = 600). Threshold conditions of Moderate Asymmetry and Moderate Asymmetry-Alternating are also omitted; both Moderate threshold conditions showed similar patterns to their Extreme counterparts, but with better performance overall.

Below, we summarize our findings with respect to four outcomes: 1) convergence rates, improper solutions (i.e., Heywood cases), and outliers, 2) quality of factor loading and factor correlation estimates, in terms of relative bias and efficiency, 3) quality of robust standard errors, in terms of relative bias and coverage, and 4) quality of test statistics, in terms of type I error rate and power. Results for 2) and 3) are summarized only for loadings whose population values were λ = .3 or λ = .7 (recall that there were 2 loadings of each size in the small model and 4 loadings of each size in the large model; the presented results are averaged across all loadings of the same population value). The results for the intermediate loading values (λ = .4, .5, .6) can be found in supplementary materials.

Convergence Failures, Heywood Cases, and Outliers

A high rate of convergence failures, improper solutions, and outliers is an undesirable characteristic of an estimation method, both because these situations lead to uninterpretable results and also because researchers frequently interpret such outcomes as revealing poor model fit (Chen, Bollen, Paxton, Curran, & Kirby, 2001). Most convergence failures and improper solutions (i.e., when cat-LS estimation produced a factor loading greater than 1 or continuous ML estimation produced a standardized factor loading greater than 1) appeared with small samples (*N* = 100 or 150). Remaining conditions had at most 1 convergence failure out of 1000 replications, and all but 4 cells had fewer than 2% Heywood cases (the other 5 cells had between 2% and 8% condition codes, and were found in asymmetric threshold conditions). Both convergence failures and improper solutions occurred most frequently with 2 categories and decreased in frequency as the number of categories increased; with more than 4 categories, at least 99% of cases converged in every condition, and improper solutions were limited to 5% in all but two conditions. Continuous ML estimation resulted in more convergence problems than cat-LS estimation; cat-LS estimation resulted in a greater number of improper solutions. In general, the number of convergence failures and improper solutions in the larger Model 2 were less than half those found in Model 1, and in the vast majority of conditions, Model 2 had no failures of either type. Complete data on rates of non-convergence and Heywood cases can be found in the supplementary materials. Non-convergent and Heywood cases were removed from subsequent analyses. Finally, outliers were defined as any case that produced a standard error greater than 1. In only one case out of 420 conditions and 2 methods did such an outlier *not* correspond to an improper solution. This one case was additionally excluded from the analysis.

Parameter Estimates

Bias. Relative bias for parameter estimates was defined as $RB = \frac{(\bar{\theta}_{est} - \theta)}{\theta}$, where θ is the true parameter value and $\bar{\theta}_{est}$ is the average estimated value of the parameter across all replications in a given cell. Consistent with other literature, we consider estimates to be substantially biased if |RB| > .10 (e.g., Flora & Curran, 2004). While we present average parameter estimates rather than relative bias estimates in the figures, the absolute value of the bias can be inferred from the distance of each estimate from its true parameter value, and we note when the relative bias exceeded 10% in text.

Figures 3 and 4 present mean estimated parameter values for λ = .3 and λ = .7 for underlying normal data and underlying nonnormal data respectively. These figures make clear that with 2 categories, ML factor loading estimates are all substantially negatively biased, regardless of sample size. As the number of categories increases, ML estimates gradually become less biased, and by 5 categories relative bias is always less than 10%. Cat-LS estimates, on the other hand, are largely accurate with 2 to 4 categories, and remain accurate with 5-7 categories. There is only one exception to this pattern: in the Extreme Asymmetrical Alternating threshold condition when *N* = 100 and data are binary, cat-LS estimates display considerable bias, particularly when loadings are high. Although ML estimates are consistently within acceptable levels of bias as long as the number of categories is at least 5, ML estimation is affected to a much greater extent than cat-LS estimation by the distribution of category thresholds. The skew in the observed variables that results from extreme category asymmetry presents a greater challenge for ML estimation. Finally, comparing Figures 3 and 4, we see that the shape of the underlying distribution affects cat-LS estimates more so than ML. When the underlying distribution is nonnormal, all cat-LS parameter estimates take on a slight positive bias (around 4%), except when there are just two categories. In this case, and especially with extremely asymmetrical thresholds, parameters are underestimated. This is to be expected, because cat-LS assumes the underlying continuous distributions are normal, and cat-LS estimates are not consistent when this assumption is violated. Model size had little effect on the accuracy of parameter estimates: the pattern of results in Figures 3 and 4 equally describes Models 1 and 2.

Figure 5 presents the mean parameter estimates of the factor correlation (population value is .3) for both normal and nonnormal underlying data. As is clear from the figure, both methods generally yield accurate results. In particular, ML yields unbiased estimates in almost every condition, regardless of sample size, number of categories, symmetry and underlying distribution. Cat-LS estimates show little bias across most conditions, except when the combination of factors includes binary data, extremely asymmetrical thresholds, and a small sample. In these conditions, the factor correlation as estimated by cat-LS is much higher than the true value, and this bias is greater when the underlying data are nonnormal. These are the same conditions that correspond to factor loading bias with cat-LS. Comparing the two methods, cat-LS generally produces slightly more accurate estimates than ML with 2-4 categories, and slightly less accurate estimates with 5-7 categories[11]. This comparison is hardly meaningful, however, given that bias is almost never greater than 5% with either method. Model size again has little effect on estimates of the factor correlation.

Efficiency. We evaluated the size of empirical standard deviations of standardized parameter estimates as a measure of efficiency. These results are not presented but are available in the supplementary materials. As expected, with both methods, parameters are estimated more efficiently with increasing sample size and an increasing number of categories; this increase in efficiency is approximately equal across the two methods. Both methods yield less efficient estimates when thresholds are asymmetric. ML yields more efficient factor loading estimates than cat-LS when factor loadings are small. When factor loadings are large and sample size is small, cat-LS produces more efficient estimates than ML; when factor loadings and sample size are large, ML estimates are more efficient. ML factor correlation estimates are more efficient than cat-LS estimates when thresholds are symmetrically distributed, but in more complex threshold conditions, cat-LS estimates are more efficient. It should be noted that efficiency comparisons only make sense when there is little or no bias in parameter estimates. For this reason, the tendency of ML factor loadings to be more efficient than those of cat-LS when *N* is large is not always meaningful, particularly with few categories, where ML is most biased.

Robust Standard Errors

We evaluated the performance of robust standard errors in terms of bias relative to empirical standard errors (i.e., standard deviations of unstandardized parameter estimates) as well as in terms of the coverage of 95% confidence intervals. Coverage results are presented in Figures 6-8; bias results are not presented, but both bias and coverage are discussed below.

Bias. Relative bias for robust standard error estimates is defined as $RB = \frac{(SE_{est} - SE_{emp})}{SE_{emp}}$,

where SE_{est} is the average estimated robust standard error in a given cell and SE_{emp} is the empirical standard deviation of parameter estimates in the corresponding cell, which is used as a proxy for the true parameter standard error.

Both ML and cat-LS robust standard error estimates display consistently negative bias; that is, they are on average smaller than the empirical standard deviations of the associated parameter estimates. When the sample size is small, this bias is often substantial. In particular, ML standard errors are from 8 to 30% (average is 15%) smaller than empirical standard errors when the sample size is small, and cat-LS standard errors are from 3 to 37% (average is 13%) smaller than empirical standard errors when the sample size is small. The extent of this bias is not affected by threshold condition or underlying distribution.

On the whole, the two methods produce comparable standard error estimates. Cat-LS produces better robust standard errors for factor loadings, and ML produces better robust standard errors for factor correlations. This finding is consistent across number of categories.

Coverage. Figures 6 and 7 present coverage rates for λ = .3 and λ = .7 when the underlying continuous distribution is normal (Figure 6) and nonnormal (Figure 7). Figure 8 presents coverage for the factor correlation for both types of underlying distribution. Coverage is defined as the proportion of 95% confidence intervals (created using robust standard error estimates) around the estimated parameter value that include the true parameter value. As such, 95% is the optimal value of coverage, and coverage below 90% is considered inadequate (Collins, Schafer & Kam, 2001; Enders & Peugh, 2004). The largest observed coverage value was 95.3%. Because coverage is a joint measure of parameter estimate bias and standard error bias, it can be low if parameter estimates are biased, standard error estimates are too small, or a combination of these.

Overall, the performance of the two methods in terms of coverage is similar to their performance in terms of parameter estimate bias. The more drastically parameter estimates diverge from the true parameter value (Figures 3-5), the more coverage drops to unacceptable levels (Figures 6-8). As with parameter bias, when factor loadings are high, cat-LS outperforms ML in almost every condition. The difference between methods is most pronounced when the number of categories is small (due to greater bias in parameter estimates) and when the sample is large (due to smaller standard errors). ML coverage rates are affected by both the underlying distribution and threshold symmetry. When thresholds are symmetric, ML coverage rates with 5 or more categories are around 90%. When thresholds are not symmetric, ML coverage rates with 5 or more categories and *N* = 600 range from 46% (5 categories, underlying normal, Extreme Asymmetry-Alternating thresholds) to 89% (7 categories, underlying nonnormal, Extreme Asymmetry thresholds). These results suggest that confidence intervals around ML parameter estimates may not be reliable if there is evidence that categories are asymmetrically distributed. Cat-LS coverage rates are more predictable: when the underlying distribution is normal, coverage is always high (greater than 90%); when it is nonnormal, coverage becomes lower, but never drops below 83% when *N* = 600 (when *N* = 100, there is a single condition in which coverage is poor, corresponding to Extreme Asymmetry-Alternating thresholds, underlying nonnormal distribution, and 2 categories. Poor coverage in this case is due to a high degree of bias in the parameter estimates). Cat-LS coverage is best with few categories.

The ML coverage is much better for factor loadings whose population values are low. With 5 or more categories, coverage in all conditions is at least 87%. With 4 categories, coverage is at least 83%, and with 2-3 categories, it is highly variable, ranging from 60 to 94% depending on underlying distribution and threshold symmetry. Cat-LS coverage rates of low factor loadings are very high across all conditions, except in the one condition where significant parameter estimate bias was present, that is, with binary data, N = 100, underlying nonnormality, and alternating extremely asymmetric thresholds. In this condition cat-LS coverage is just 65%, but ML coverage is even lower at 63%.

Model size has a slight effect on the relative performance of the methods: both estimators have slightly poorer coverage in the 20-indicator model than in the 10-indicator model. The difference between models is bigger when the coverage rate is worse for both methods. For instance, when coverage for the smaller model is 85% or higher, it may be 1-2% lower for the larger model. When coverage for the smaller model is very poor (e.g., 40%), it is often drastically lower for the larger model (e.g., 17%).

Figure 8 presents coverage results for the factor correlation as a function of threshold condition, number of categories, estimation method, and sample size. When N = 600, the two methods provide similar coverage rates, which tend to vary from .90 to .95. When N = 100, both cat-LS and ML have poorer coverage when the number of categories is small and the thresholds are asymmetric, but ML coverage rates are better than those of cat-LS, particularly when the threshold distribution is not symmetric.

Test Statistics

With continuous-variable data, we examined the mean-adjusted statistic $T_{ML\text{-}M}$ and the mean-and-variance-adjusted statistic, $T_{ML\text{-}MV}$ (Satorra & Bentler, 1994). It was found that $T_{ML\text{-}M}$ produced systematically higher type I error rates than $T_{ML\text{-}MV}$ across every condition; thus, we do not present the results for T_{M}. However, this finding implies that the good performance of $T_{ML\text{-}M}$ that has been reported in simulation studies (e.g, Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992) depends heavily on the assumption that the data are continuous nonnormal rather than categorical. With categorical-variable data, we examined only the mean-and-variance adjusted statistic $T_{cat\text{-}MV}$, because M*plus* does not provide $T_{cat\text{-}M}$ with cat-LS estimation[12]. However, $T_{cat\text{-}MV}$ generally performed very well.

Type I error. Empirical type I error at $\alpha = .05$ is defined as the proportion of converged replications that generated a *p*-value less than .05. Type I error rates for ML-based $T_{ML\text{-}MV}$ and cat-LS-based $T_{cat\text{-}MV}$ are presented in Figure 9. With one exception, both cat-LS and ML produce test statistics with reasonable type I error control. The exception is the case of Extreme Asymmetry-Alternating thresholds, with few categories and a large sample, when the type I error rate associated with $T_{ML\text{-}MV}$ is highly inflated. This inflation is seen most clearly in the larger model and when the underlying distribution is nonnormal; in this case, the type I error rate with 2-3 categories is nearly .6[13]. Normal theory ML methodology actually assumes that a linear model holds for observed data, and thus the model is fundamentally misspecified. Thus, technically, $T_{ML\text{-}MV}$'s rejection rates measure not type I error but power to detect this misspecification (Maydeu-Olivares, Cai, & Hernandez, 2011). The assumption behind using continuous methodology with categorical variables is that, at least as the number of categories gets large, this power will be low enough to function as a type I error rate. As Figure 9 shows, this is largely true, except with few categories, $N = 600$, and Extreme Asymmetry-Alternating thresholds, where the increased sample size and the worst type of categorization results in $T_{ML\text{-}MV}$ having fairly high power to detect that the underlying linear model is misspecified (i.e., data were wrongly assumed to be continuous).

Setting aside this most problematic condition, type I error rates rarely rise above .1, and they are frequently at or below .05 for both methods. With 6-7 categories and a small sample size, cat-LS exhibits the worst type I error control, frequently reaching around .1 or a little above.

In general, type I error rates associated with cat-LS increase as the number of categories increases, whereas those associated with ML are stable. In the Extreme Asymmetry and Extreme Asymmetry-Alternating threshold conditions, type I error rates of $T_{cat\text{-}MV}$ spike at 3 categories. This spike is higher for the larger model (around 24%); it is not clear what causes this particular anomaly. Type I error rates tend to be around 1% lower in the larger model than the smaller model; wherever a particularly poor condition leads to a high type I error, it is worse in the larger model.

Power. While this study is mainly concerned with correctly specified models, we conducted a brief evaluation of the relative power of the ML-based and the cat-LS-based robust test statistics to detect at least a major model misspecification. We fit a 1-factor model to the data generated by Model 1 (the 10-indicator, 2-factor model) for the subset of conditions in which the underlying distribution was normal and thresholds were symmetrically distributed. Under these conditions, type I error rate was well controlled by both statistics, allowing for a power examination. Table 2 displays the results. Values lower than 80% are bolded. With large samples (N = 350 to 600), both statistics have virtually perfect power to detect a misspecified model. With N = 100 or 150 and 4 or more categories, cat-LS has slightly greater power than ML; with 2-3 categories, the two methods have comparably low power rates.

Discussion

The aim of the present study was to revisit the question, how many categories should variables have before they may be treated as continuous? Though many studies have investigated aspects of this question, none have directly compared the best currently available robust methodologies for both types of data, and none have systematically manipulated multiple factors to simulate a variety of challenging conditions. We investigated the most commonly used continuous nonnormal estimation method, robust ML, and one of the best categorical estimation data methods, robust cat-LS. To determine the number of categories at which method choice ceases to matter, we studied variables with 2 through 7 categories. To test the robustness of each method against various assumption violations and non-ideal scenarios, we manipulated the distribution of underlying data, the thresholds used to categorize it, model size, and sample size. The results of this investigation allow us to make broad conclusions about the relative utility of the two studied methods.

Overall, our results confirm that cat-LS performs extremely well with up to 7 categories, and in a variety of conditions. The only problematic conditions for cat-LS involved the intersection of underlying nonnormality and small samples. However, once the number of categories in the data reached 5, continuous robust ML frequently performed as well as (and occasionally better than) robust cat-LS. In brief, we conclude that cat-LS is a good estimation choice for ordinal data with as many as 7 categories, but that with 5 or more categories robust ML is an acceptable choice as well, particularly when the thresholds are not wildly asymmetric. We provide a more detailed discussion of our findings below.

2 to 4 Categories

Our results confirm the conventional wisdom that, with just 2-4 categories, continuous methodology is generally not recommended. With few categories, robust ML consistently underestimates factor loadings and parameter standard errors. Together, these two shortcomings lead robust ML to have unacceptably low coverage for factor loadings. Consistent with previous studies, however, ML produced unbiased estimates of the factor correlation. These results suggest that it is the measurement model parameters that are most affected by wrongly assuming that a linear model describes the relations between categorical variables and latent factors. The *structural* model parameters (in this case, factor correlations), are not affected, and if the structural parameters are of greatest interest, robust ML can be an acceptable choice even with 2- to 4-category data, and is in fact preferred when the sample size is small. While cat-LS was largely superior to ML with 2-4 categories, it showed mild upward bias in non-ideal conditions (underlying nonnormal distribution, asymmetric thresholds, small samples).

When it comes to test statistics, $T_{cat\text{-}MV}$ performs well in general, but tends to under-reject when data are binary. This underrejection, however, does not translate into loss of power relative to $T_{ML\text{-}MV}$, which does not have this tendency. $T_{cat\text{-}MV}$ also tends to over-reject models with 3-category data when the underlying data distribution is nonnormal. ML-based $T_{ML\text{-}MV}$ performs about as well as $T_{cat\text{-}MV}$, except with alternating threshold asymmetry, where ML rejection rates are occasionally extremely high. This finding means that even though parameter estimates are generally downward biased when robust ML is applied to data with 2-4 categories, the continuous ML robust test statistic can still be useful to evaluate the overall model fit. Both statistics had relatively low power to detect serious model misspecification with small samples and 2-3 category variables; this finding suggests that researchers fitting models to such severely categorized data should aim to obtain a larger sample size than is typically recommended for continuous data.

5 to 7 Categories

While previous research has generally agreed that it is a bad idea to use ML with fewer than 5 categories, evidence has been unclear when 5 or more categories are present. Past findings have suggested that categorical methodology can outperform continuous methodology with 5, 6, and even 7 categories (e.g., Beauducel & Herzberg, 2006; Dolan, 1994); however, no study has compared categorical methods to the best-available robust continuous methods. With 5-7 categories, cat-LS continues to display more accurate factor loading estimates than ML, although both methods produce estimates within the range of acceptable bias. When category thresholds are roughly symmetric, ML is as good as or better than cat-LS. In particular, when there are 7 categories, ML is frequently preferable to cat-LS. As category thresholds become more asymmetric, ML parameter estimates become increasingly negatively biased, which in turn affects confidence interval coverage. In no condition, however, did we observe that the relative bias of ML parameter estimates was greater than 10% with 5 or more categories. The worst bias was observed when factor loadings were high: the higher the loading, the greater the bias in ML estimates. As with 2-4 categories, ML estimates of the structural model parameters (in this case, the correlation between two factors) were extremely accurate, producing marginally better estimates than cat-LS.

When it comes to coverage, the two methods produce similarly good coverage rates for low factor loadings and factor correlations. When factor loadings are high, cat-LS confidence intervals are almost always more accurate. This is especially true when the underlying distributions are normal but the category thresholds are asymmetric. When these two conditions hold, cat-LS coverage rates are between 4% and 49% higher. When thresholds are symmetric, or when the underlying distribution is nonnormal, the two methods produce comparable coverage rates, although cat-LS coverage continues to be just slightly better.

Finally, when it comes to type I error and power of the robust test statistic, both methods produced comparably good results. With 5 categories, $T_{ML\text{-}MV}$ produces slightly worse rejection rates than $T_{cat\text{-}MV}$ in the Extreme Asymmetry Alternating conditions. There are virtually no differences between the type I error rates of the two statistics with 6 categories. With 7 categories, $T_{ML\text{-}MV}$ produces slightly more accurate rejection rates than $T_{cat\text{-}MV}$ in smaller samples. $T_{cat\text{-}MV}$ is slightly more powerful than $T_{ML\text{-}MV}$ but only when N =100.

Overall, these findings suggest that, particularly if other reasons to prefer ML estimation are present, there is little problem in doing so with 5 or more categories.

Additional Considerations

In summary, the present findings support the common wisdom that categorical methodology is most necessary when variables have 2 - 4 categories. Additionally, categorical methodology outperforms continuous methodology with up to 7 categories when category thresholds are asymmetric and factor loadings are high. Robust continuous methodology performs as well as cat-LS with 5 to 7 categories, and even slightly better with 7 categories and a small sample size, provided that thresholds are approximately symmetric.

Other factors also play a role in the relative performance of ML and cat-LS. Cat-LS parameter estimates are more sensitive than robust ML parameter estimates to violations of the assumption of normality of underlying continuous variables: when the underlying distribution is nonnormal, cat-LS parameter estimates increase by a few tenths of a standardized loading, leading to small but systematic overestimation of the parameter value. For both ML and cat-LS, conditions that lead to large negative bias in parameter estimates when the underlying distribution is normal lead to even worse bias when the distribution is nonnormal. Coverage and type I error rates are not very affected by underlying nonnormality when cat-LS is used, and they are inconsistently affected (sometimes for the better, sometimes for the worse) when ML is used. The practical significance of these results may be limited, however, as the assumption that each categorical variable is a categorization of an underlying continuous normally distributed variable is not testable. Researchers may wish to use additional caution when interpreting results obtained by cat-LS when it is likely that this assumption is violated. Thoughtful consideration of a given construct as well as the population of interest may reveal whether the normality assumption is plausible. For example, if a categorical variable measures frequency of drug use, it is unlikely that the underlying continuous variable is normally distributed in the general population, although it may be in some clinical populations.

Our findings regarding model size generally indicate little difference between methods as a function of model size. ML parameter estimates are insensitive to model size, and cat-LS parameter estimates are only affected when the sample size is small. In this case, the smaller model shows less bias. This suggests that categorical methodology is more sensitive to the extreme situation when a model's degrees of freedom are larger than the sample size; this may be because the asymptotic covariance matrix of polychoric correlations is more difficult to estimate in these conditions. Both methods have poorer coverage but lower type I error rates in the larger model.

There are several reasons why applied researchers may continue using continuous estimation methods, even with ordinal data. Continuous methods are older and hence more familiar to researchers. ML with robust corrections for nonnormality is no longer an esoteric method but is routinely used in applications. While limited information categorical methods with robust corrections are rapidly gaining popularity, they remain newer, less familiar, and less integrated with other developments. For example, the treatment of missing data is largely straightforward with continuous methodologies but remains a challenge with categorical methodology (e.g., M*plus* defaults to pairwise deletion when data are declared as categorical). Although it would be possible to conduct multiple imputation on a categorical-variable dataset before analyzing, the standard recommendation is to impute continuous values for missing data even when the original variables are categorical, which would produce a dataset containing a mix of continuous and categorical observations (Enders, 2010). Most applied researchers are also limited in their choice of software, and it may be that the software package they have access to only implements one type of methodology (e.g., the R package *lavaan* 0.4-13 supports robust ML methodology, but not yet categorical methodology). The results of the present study indicate that reliance on continuous methodology in the presence of ordinal data will produce acceptable results when the number of categories is 5 or higher. If no other choice exists but the data have fewer categories, the researcher should interpret the estimated measurement model parameters as severe underestimates.

One limitation of our results is that different software packages can implement the same methodology somewhat differently, and this particularly applies to the limited information categorical methodology, which has historically been developed separately by authors of various software packages (Jöreskog, 1994; Lee et al,, 1990, 1995; Muthén 1984, 1993; Muthén et. al., 1997). Our recommendations are thus somewhat specific to users of M*plus*. However, we also analyzed these data using EQS, and the results for parameter estimates and standard errors were largely similar, although categorical test statistics produced by EQS behaved worse. A second limitation is that our results are based on CFA models, and as such they should not be generalized to more complex models, such as multiple group models or mixture models (e.g.., Bauer, 2005; Bauer & Curran, 2004). More research is necessary to determine how the choice of continuous vs. categorical methodology impacts estimation and inference in such models.

In addition to studying models that are more complex than classic CFA models, future research should study models with a larger structural component, to fully test the hypothesis that continuous methods estimate structural parameters accurately even when the number of categories is low. A more thorough investigation of power is also warranted. Finally, a more detailed comparison of cat-LS and other categorical methods, such as cat-DWLS may also be in order; while Forero et al. (2009) made such a comparison and ruled in favor of cat-LS, these authors did not examine factor correlations or test statistics.

Conclusion

The present study summarizes the relative performance of the best available robust continuous and categorical methodologies for CFA with categorical variables. Our findings confirm that when observed variables have fewer than 5 categories, robust categorical methodology is best. With 5 to 7 categories, both methods yield acceptable performance; the choice between methods will depends on other aspects of the data, including sample size, model size, the symmetry of the observed distribution, the likely underlying distribution of the constructs being measured, and the results that are of most interest to the researcher. These findings should guide applied researchers in their choice of method.

References

Babakus, E., Ferguson, C. E. J., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24*, 222-228. doi:10.2307/3151512

Baker, F. B., & Kim, S. -H. (2004). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Bartholomew, D. J., & Knott, M. (1999). *Latent Variable Models and Factor Analysis, 2nd Edition*. London: Arnold.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods, 10*, 305-316.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*, 3-29.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186–203. doi:10.1207/s15328007sem1302_2

Bentler, P. M. (2008). EQS structural equation modeling software. Encino, CA: Multivariate Software.

Bentler, P. M., & Yuan, K. H. (2011). Positive definiteness via off-diagonal scaling of a symmetric indefinite matrix. *Psychometrika, 76*, 119-123. doi: 10.1007/s11336-010-9191-3

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105,* 467-477.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. Applied *Psychological Measurement*, *32*, 771–775.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179-197.

Bollen, K. A., & Barb, K. H. (1981). Pearson's R and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

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, F., Bollen, K., Paxton, P., Curran, P.J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research, 29,* 468-508.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5-32. doi:10.1007/BF02291477

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330-351.

Crowley, S. L., & Fan, X. (1997). Structural equation modelling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment, 68*, 508-531. doi:10.1207/s15327752jpa6803_4

Curran, P. J., West, S. G., & Finch, J. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29. doi:10.1037/1082-989X.1.1.16

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*, 327-346. doi:10.1207/S15328007SEM0903_2

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309–326.

Enders, C. K. (2010). Applied missing data analysis. New York: Guilford Press.

Enders, C. K., & Peugh, J. L. (2004). Using an EM Covariance Matrix to Estimate Structural Equation Models with Missing Data: Choosing an Adjusted Sample Size to Improve the Accuracy of Inferences. *Structural Equation Modeling, 11*, 1-19. doi:10.1207/S15328007SEM1101_1

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6*, 323-329.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532. doi:10.1007/BF02293811

Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466–491. doi:10.1037/1082-989X.9.4.466

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded-response models: Limited versus full-information methods. *Psychological Methods, 14*, 275-299. doi: 10.1037/a0015825

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*, 625-641. doi:10.1080/10705510903203573

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4,* 108-120. doi: 10.1080/10705519709540064

Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26,* 329–367.

Hu, L.-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351-362. doi:10.1037/0033-2909.112.2.351

Joe, H., & Maydeu-Olivares, A. (2011). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika, 75,* 393-419. doi: 10.1007/S11336-010-9165-5

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review, 48*, 398-407. doi:10.2307/2095231

Jöreskog, K.G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59,* 381–390. doi:10.1007/BF02296131

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457-477. DOI: 10.1207/s15327906mbr2603_5

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1990). A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika, 55,* 45-51. doi:10.1007/BF02294742

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology, 48*, 339–358.

Lei, P.-W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity, 43*, 495-507. doi:10.1007/s11135-007-9133-z

Marsh, H. W., & Hau, K-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32*, 151-170. doi:10.1016/j.cedpsych.2006.10.008

Martens, M. P. (2005). The use of structural equation modelling in counselling research. The *Counselling Psychologist*, *33*, 269-298. doi:10.1177/0011000004272260

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66,* 209–227. doi:10.1007/BF02294836

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the Fit of Item Response Theory and Factor Analysis Models. *Structural Equation Modeling, 18,* 333-356. doi: 10.1080/10705511.2011.581993

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009-1020. doi:10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713–732.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479-515.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables, *Psychometrika, 43,* 551-560. doi:10.1007/BF02293813

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132. doi:10.1007/BF02294210

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

Muthén, B. O., (2003, August 18). Re: Underlying Normality and Polychoric Correlations [Online Forum Comment]. Retrieved from http://www.statmodel.com/discussion/messages/23/65.html

Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in M*plus*. M*plus* Web Note #4. Retrieved from http://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Muthén, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika, 53,* 563-577. DOI: 10.1007/BF02294408

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

Muthén, L. K., & Muthén, B. O. (1998-2010). Mplus User’s Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analysing multitrait–multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: What sample size is needed for valid results? *British Journal of Mathematical and Statistical Psychology, 59,* 195–213. doi:10.1348/000711005X67490

Olsson, U. (1979a). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44,* 443-460.

Olsson, U. (1979b). On the robustness of factor analysis against crude classifications of the observations. *Multivariate Behavioral Research, 14,* 485-500.

Pearson, K. (1901). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 195, 1-47, 405. doi:10.1098/rsta.1900.0022

Pearson, K. (1904). Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation. Drapers' Company Research Memoirs, Biometric Series 1.

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. British Journal of Mathematical and Statistical Psychology, 46, 273-286.

Rigdon, E. E., & Fergusson, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28*, 491–497. doi:10.2307/3172790

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis, *Proceedings of the Business and Economic Statistics Section,* 308-313.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), Latent variable analysis: Applications to developmental research (pp. 399–419). Thousand Oaks, CA:

Sage.

Savalei, V. (in press). Understanding Robust Corrections in SEM. *Structural Equation Modeling*.

Skrondal, A., & Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Bocal Raton, FL: Chapman & Hall/CRC.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Vale, C.D., & Maurelli, V. A. (1983) Simulating multivariate non-normal distributions. *Psychometrika, 48*, 465-471. doi:10.1007/BF02293687

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58-79. doi:10.1037/1082-989X.12.1.58

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17,* 392-423. DOI: 10.1080/10705511.2010.489003Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51*, 289-309.

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modeling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology,* 64, 107-133.

Footnotes

[1] More precisely, we assume that the y^*'s are jointly normally distributed.

[2] In fact, M*plus* does not print the full information test statistic when variables have more than 4 categories.

[3] In CFA models, the mean structure (with continuous data) or the threshold structure (with categorical data) is saturated and is not explicitly modeled. In multiple group models, the model may be expanded to include a vector of thresholds and the correlation matrix (e.g., Millsap and Yun-Tein, 2004).

[4] The cat-WLS method is still less efficient than full information categorical methods, because those methods use information from the entire p-variate contingency table of the data, not just the first and second order marginals. What is meant here is that cat-WLS is more efficient than cat-LS or cat-DWLS, or any other method that uses a weight matrix that is not a consistent estimate of V^{-1}.

[5] It is also possible to fit a model to the polychoric correlation matrix using normal theory ML or GLS fit functions, but rates of non-convergence and improper solutions are higher (Lei, 2009; Rigdon & Ferguson, 1991), and these methods additionally require a positive definite input matrix. See, however, Yuan, Wu, and Bentler (2011) and Bentler and Yuan (2011) for new solutions to this problem.

[6] In M*plus*, cat-ULS with robust standard errors and the mean-and-variance adjusted test statistic is activated by ESTIMATOR = ULSMV. In EQS, cat-ULS is activated by METHOD = LS,ROBUST. In LISREL, it is necessary to first compute the polychoric correlation matrix in PRELIS (MA = PM), which is then used as input to the LISREL program. ULS is specified using ME = UL. In all programs, variables must to be declared to be categorical.

[7] In M*plus*, cat-DWLS with the mean-and-variance adjusted test statistic is activated by ESTIMATOR=WLSMV. In LISREL, it is activated using ME = DW. This method is not available in EQS.

[8] When data are binary and the tetrachoric correlations are estimated jointly rather than bivariately, some underlying normality tests are possible (Muthén & Hofacker, 1988), but they have not gained popularity due to difficulty of interpretation.

[9] A quarter of all conditions, including those Model 1 conditions where the underlying data were normally distributed, were also analyzed with marginal maximum likelihood analysis (in M*plus*, this is activated by ESTIMATOR = ML, and with cat-DWLS (ESTIMATOR = WLSMV). This was done to ensure that cat-LS was in fact the best available categorical estimator, as suggested in previous research (e.g., Forero et al., 2009; Forero & Maydeu-Olivares, 2009; Yang-Wallentin et al.). Results from these analyses are available in the supplementary materials, but in short, they confirm that cat-ULS is the best available categorical estimation method as the number of categories increases beyond 2-3.

[10] We suspect that these differences are due to different computations of the asymptotic covariance matrix $\hat{V}$ and in particular differences in the treatment of zero-frequency cells when computing this matrix.

[11] Results for the factor correlation were also biased in EQS under similar conditions, particularly category asymmetry, but the bias was higher and not limited to 2 categories. The highest produced overestimate was .67, with 2 categories and Extreme Asymmetry-Alternating thresholds; though overestimates also occurred with 7 categories (e.g., .60; Extreme Asymmetry). In EQS, ML produced consistently better estimates of factor correlations.

[12]EQS prints both robust categorical statistics but their performance is not as good as the performance of M*plus*'s mean-and-variance adjusted statistic, particularly with larger number of categories. With 2-4 categories, type I errors of both EQS test statistics tend to stay below .06, but they are frequently closer to 0 than to .05. Across most thresholds conditions, EQS' test statistics drop to close to zero with 5 or more categories; the exception to this is the threshold condition Extreme Asymmetry-Alternating, where both type I error rates leap up to as high as 80%.

[13]It is worth noting that M*plus*' implementation of the robust corrections is slightly different from EQS's implementation. As a result, type I error rates produced by Mplus are typically a few percentage points higher than those of EQS. The differences are slightly greater with a small sample and the larger model, where they are in the 3-8% range.

Table 1

Skew and Kurtosis of Observed Categorical Variables by Threshold Distribution, Underlying Distribution, and Number of Categories

		threshold distribution									
underlying distribution	no. cats	Symmetry		Mod. Asym		Mod. Asym-Alt		Ext. Asym		Ext. Asym-Alt	
		S	K	S	K	S	K	S	K	S	K
normal	2	0.00	-2.00	0.59	-1.65	-0.58	-1.66	1.97	1.88	-1.97	1.87
	3	0.00	-0.54	0.13	-1.09	-0.13	-1.09	1.42	0.45	-1.41	0.44
	4	0.00	-0.53	0.69	-0.23	-0.69	-0.23	1.10	-0.25	-1.10	-0.26
	5	0.00	-0.47	0.59	-0.21	-0.59	-0.21	0.91	-0.58	-0.90	-0.59
	6	0.00	-0.42	0.62	-0.11	-0.61	-0.11	0.80	-0.68	-0.80	-0.69
	7	0.00	-0.41	0.52	-0.29	-0.52	-0.29	0.79	-0.61	-0.78	-0.62
skew = 2 kurtosis = 7	2	0.50	-1.75	1.11	-0.77	-0.22	-1.95	2.26	3.13	-4.05	14.42
	3	0.00	0.27	0.29	-0.96	-0.03	-0.59	1.84	1.76	-1.25	0.56
	4	0.92	-0.05	1.08	0.44	-0.13	-0.65	1.57	0.94	-0.69	-0.82
	5	0.73	-0.16	1.10	1.04	0.21	-0.80	1.38	0.48	-0.42	-1.11
	6	0.80	0.19	1.52	1.92	0.17	-0.61	1.28	0.30	-0.25	-1.19
	7	0.93	0.30	1.33	1.17	0.32	-0.39	1.27	0.38	-0.17	-1.19

Note. Values in this table were obtained by generating samples of size *N* = 1 000 000 for each condition and recording the skew and kurtosis of the observed distributions. Mod. Asym = Moderate Asymmetry; Mod. Asym-Alt = Moderate Asymmetry-Alternating; Ext. Asym = Exteme Asymmetry; Ext. Asym-Alt = Extreme Asymmetry-Alternating. S = skew; K = kurtosis.

Table 2

Observed Power of T_{MV} Statistic to Detect a Majorly Misspecified (1-factor) Model

	2 categories		3 categories		4 categories		5 categories		6 categories		7 categories	
N	ML	cat-LS	ML	cat-LS	ML	cat-LS	ML	cat-LS	ML	cat-LS	ML	cat-LS
100	**.458**	**.424**	**.627**	**.615**	**.771**	.824	.815	.886	.846	.935	.857	.942
150	**.685**	**.698**	.878	.895	.955	.969	.974	.989	.981	.991	.989	.997
350	.997	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Type I error was assessed by fitting a 1-factor model to 2-factor simulated data. Conditions where power is less than 80% are bolded.

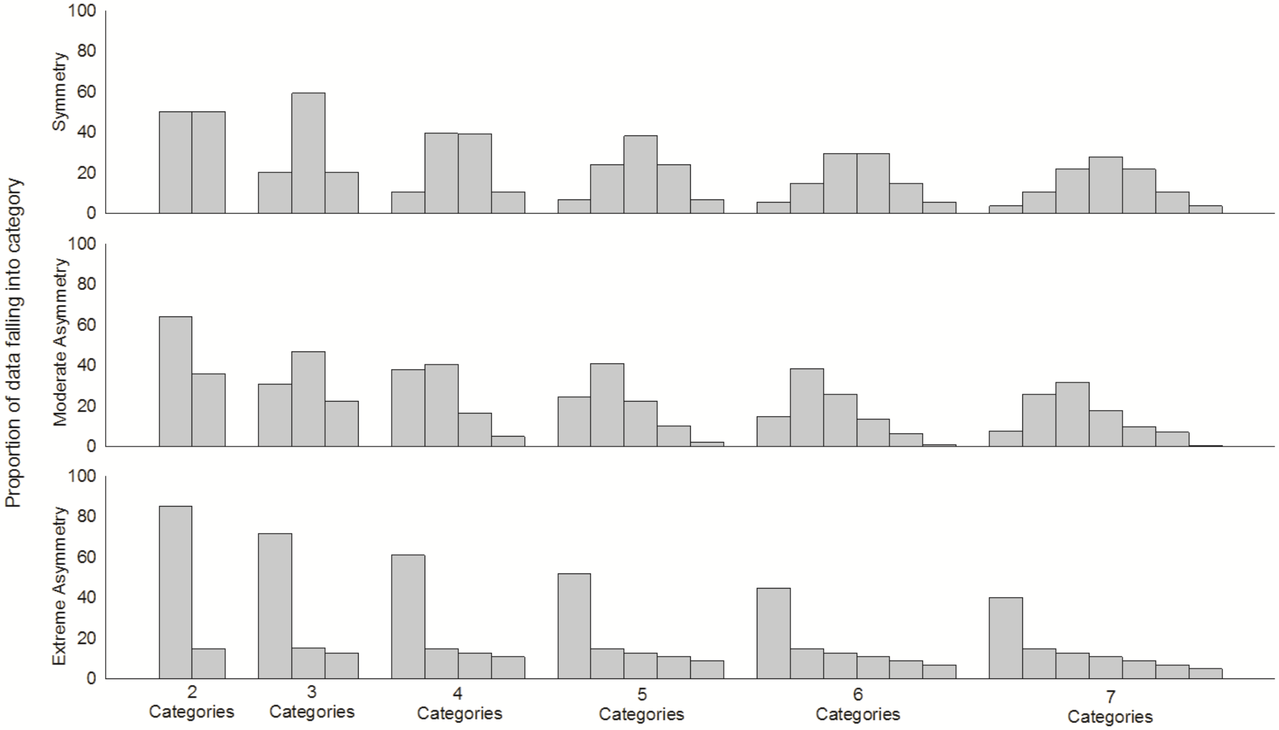

Figure 1. Distributions of observed data when thresholds imposed on normally distributed data.

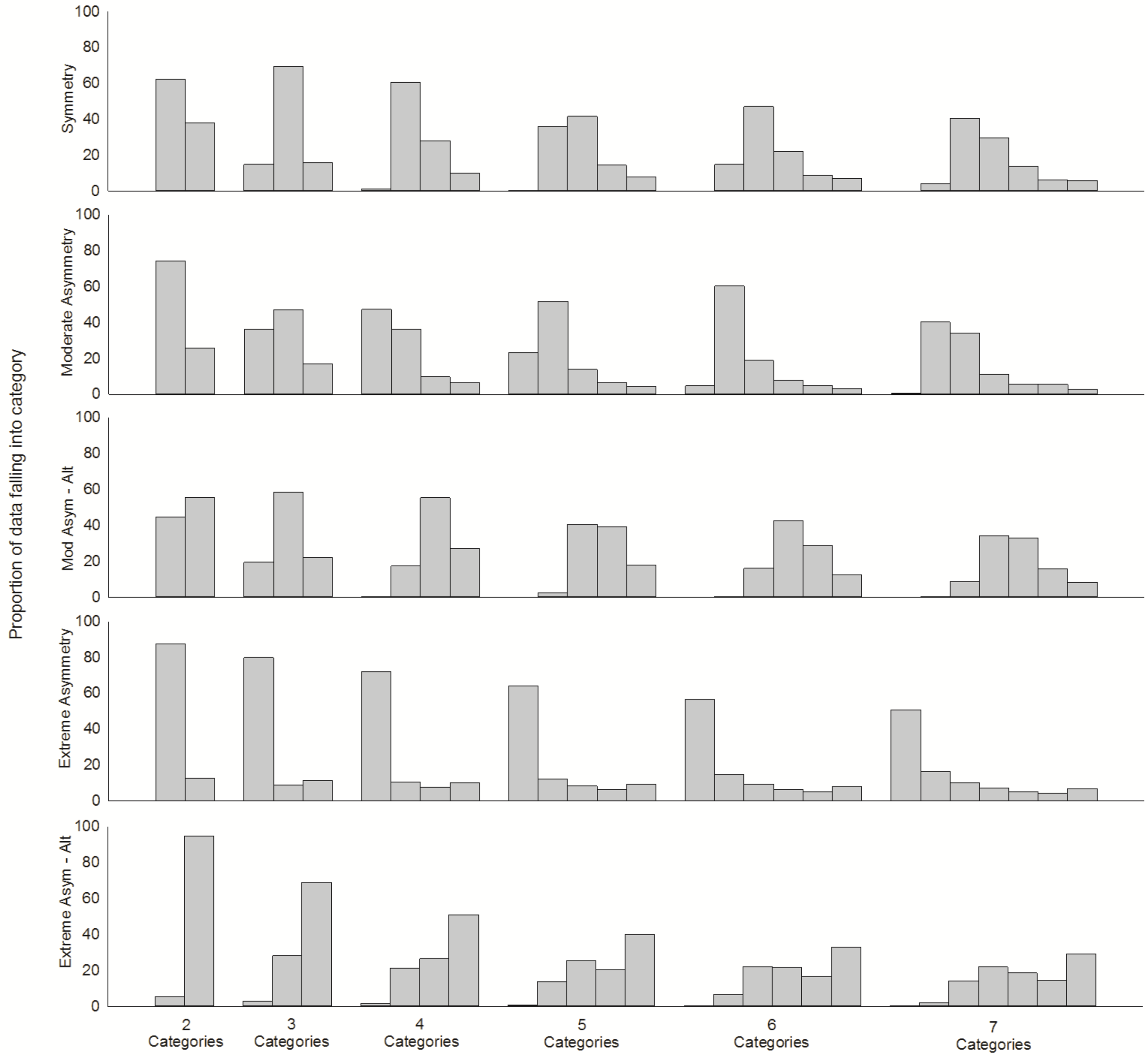

Figure 2. Distributions of observed data when thresholds imposed on non-normally distributed data (skew 2, kurtosis 7).

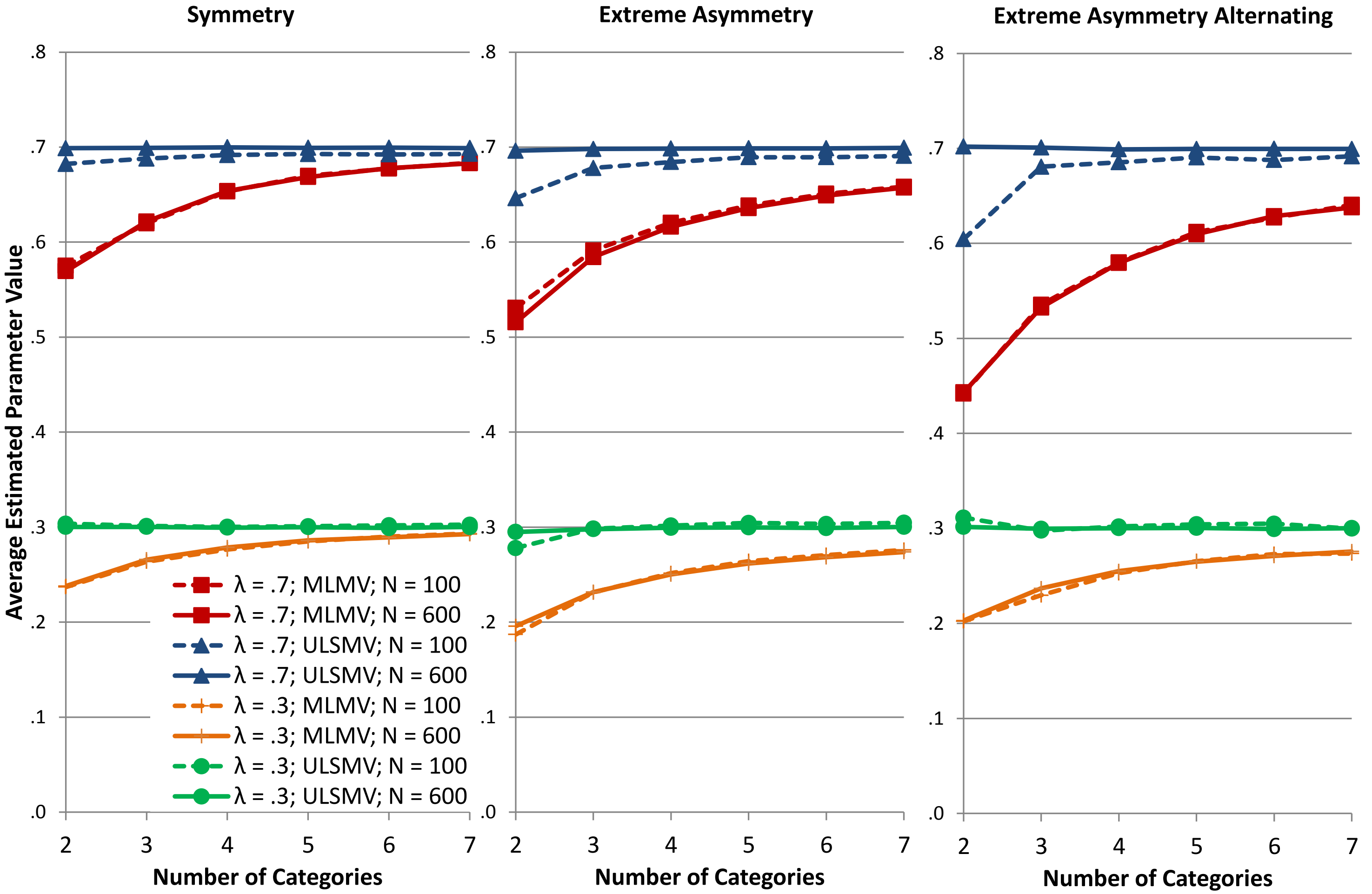

Figure 3. Parameter estimates (factor loadings, underlying distribution is normal). Values presented correspond to conditions in which the underlying distribution is normal. Values are averaged across model size, and across all loadings for which the true parameter value was the same. The upper set of lines represents results for a true parameter value of .7. In this set, estimators are denoted by color and marker shape and sample sizes are denoted by line type (see legend). The lower set of lines represents results for a true parameter value of .3. In this set, estimators are again denoted by color and marker shape and sample sizes are again denoted by line type. Vertical panels represent different levels of threshold symmetry.

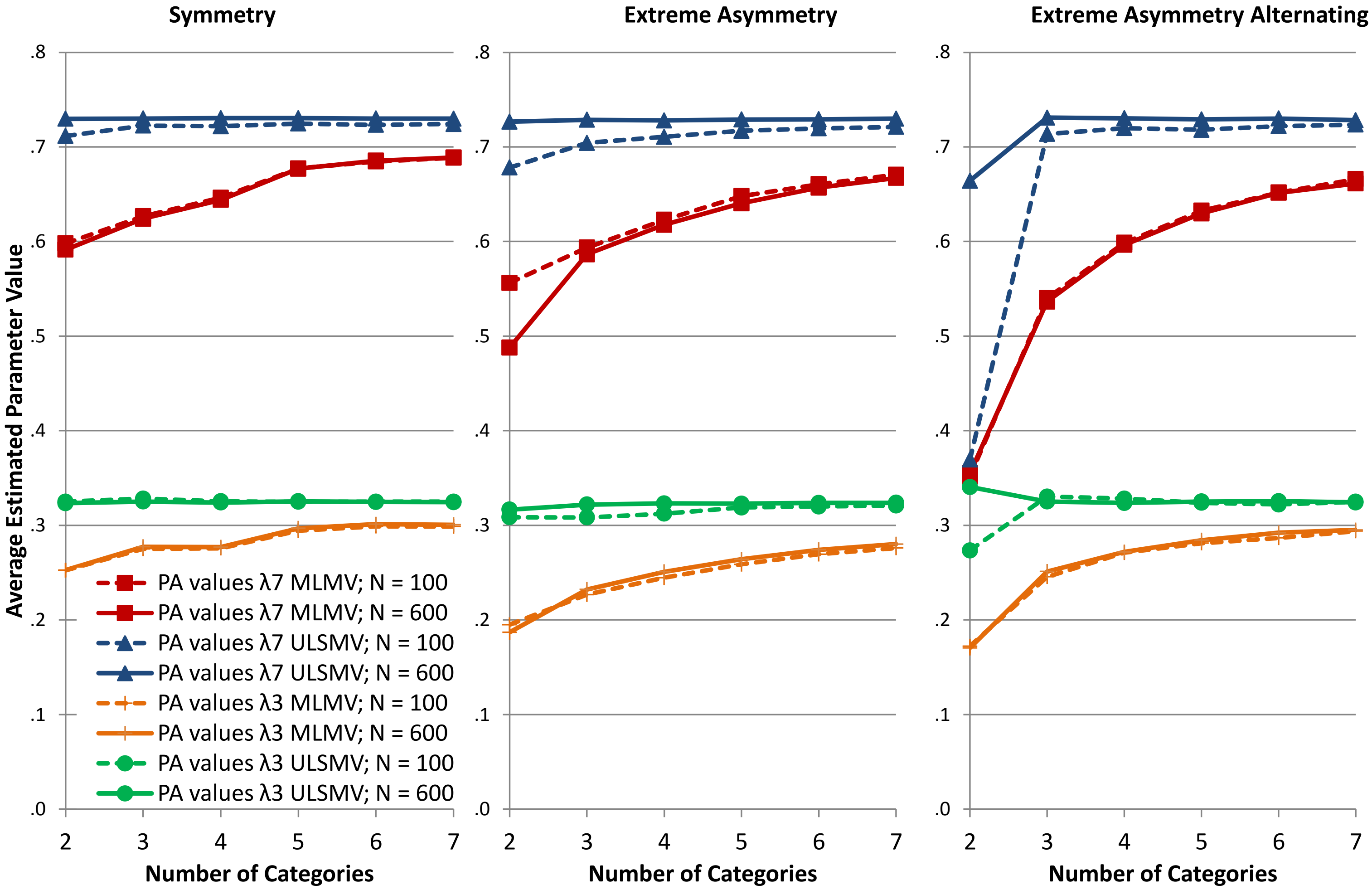

Figure 4. Parameter estimates (factor loadings, underlying distribution is nonnormal; S2, K7). Values presented correspond to conditions in which the underlying distribution is nonnormal (S2, K7). Values are averaged across model size, and across all loadings for which the true parameter value was the same. The upper set of lines represents results for a true parameter value of .7. In this set, estimators are denoted by color and marker shape and sample sizes are denoted by line type (see legend). The lower set of lines represents results for a true parameter value of .3. In this set, estimators are again denoted by color and marker shape and sample sizes are again denoted by line type. Vertical panels represent different levels of threshold symmetry.

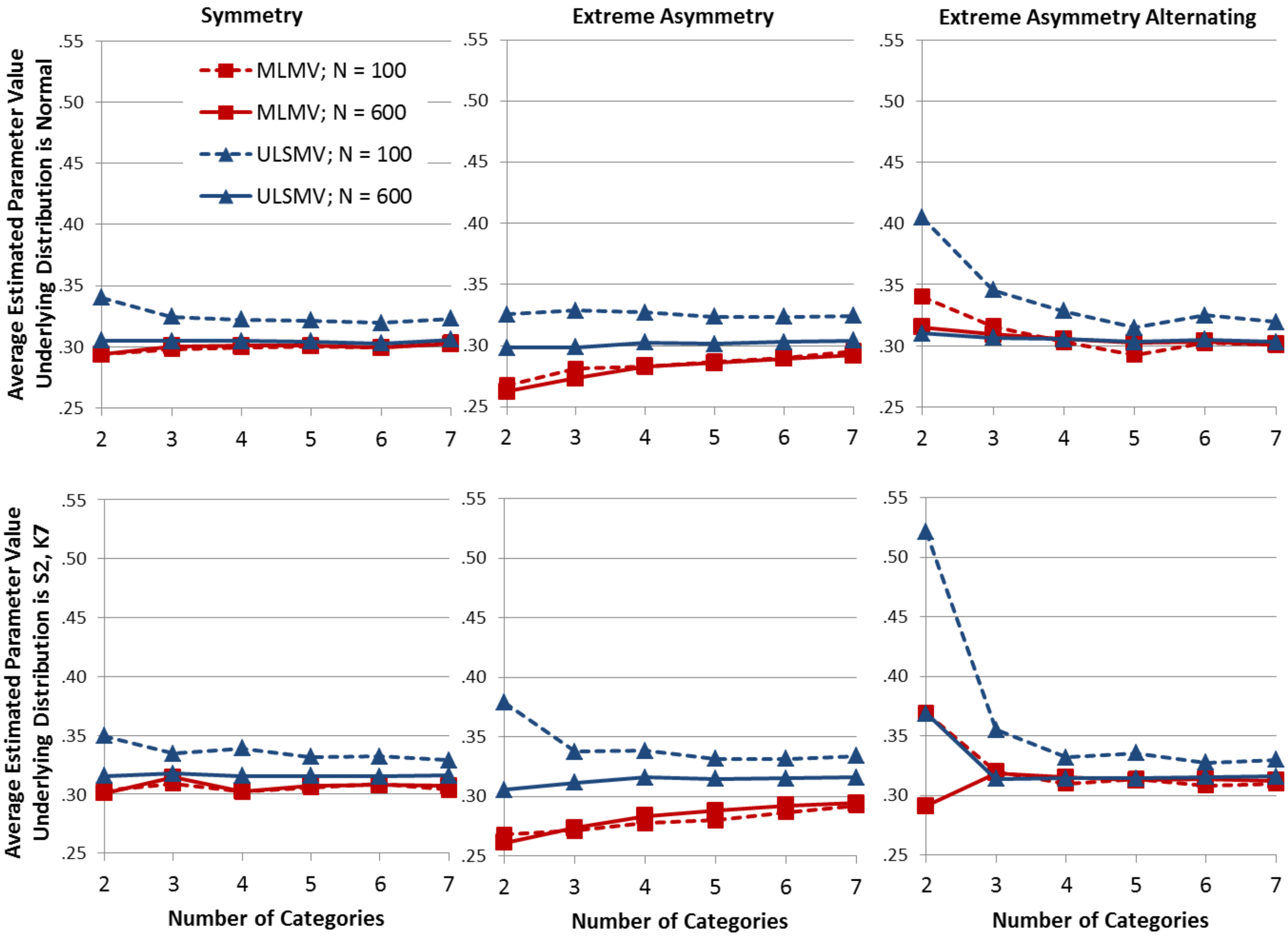

Figure 5. Parameter estimates (factor correlation, true value is .3). Values are averaged across model size. Lines represent different estimators and different sample sizes (see legend). The upper panel corresponds to conditions in which the underlying distribution is normal; the lower panel corresponds to conditions in which the underlying distribution is nonnormal (S2, K7). Vertical panels represent different levels of threshold symmetry.

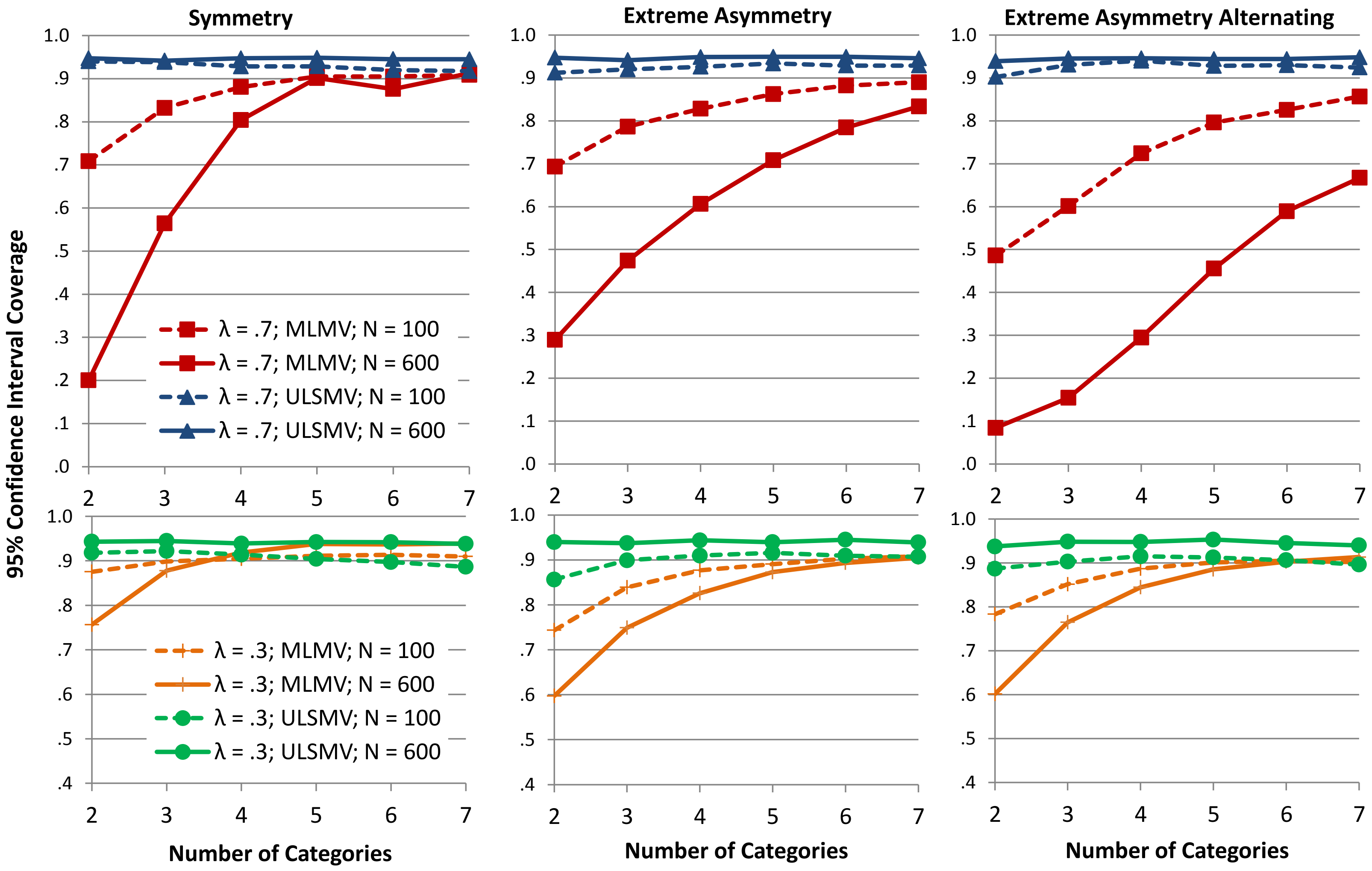

Figure 6. Coverage by number of categories (.7 and .3 factor loadings), underlying distribution is normal. Values presented correspond to conditions in which the underlying distribution is normal. Values are averaged across model size, and across all loadings for which the true parameter value was the same. The upper panel represents results for a true parameter value of .7. The lower panel represents results for a true parameter value of .3. Lines represent different estimators and different sample sizes (see legend). Vertical panels represent different levels of threshold symmetry.

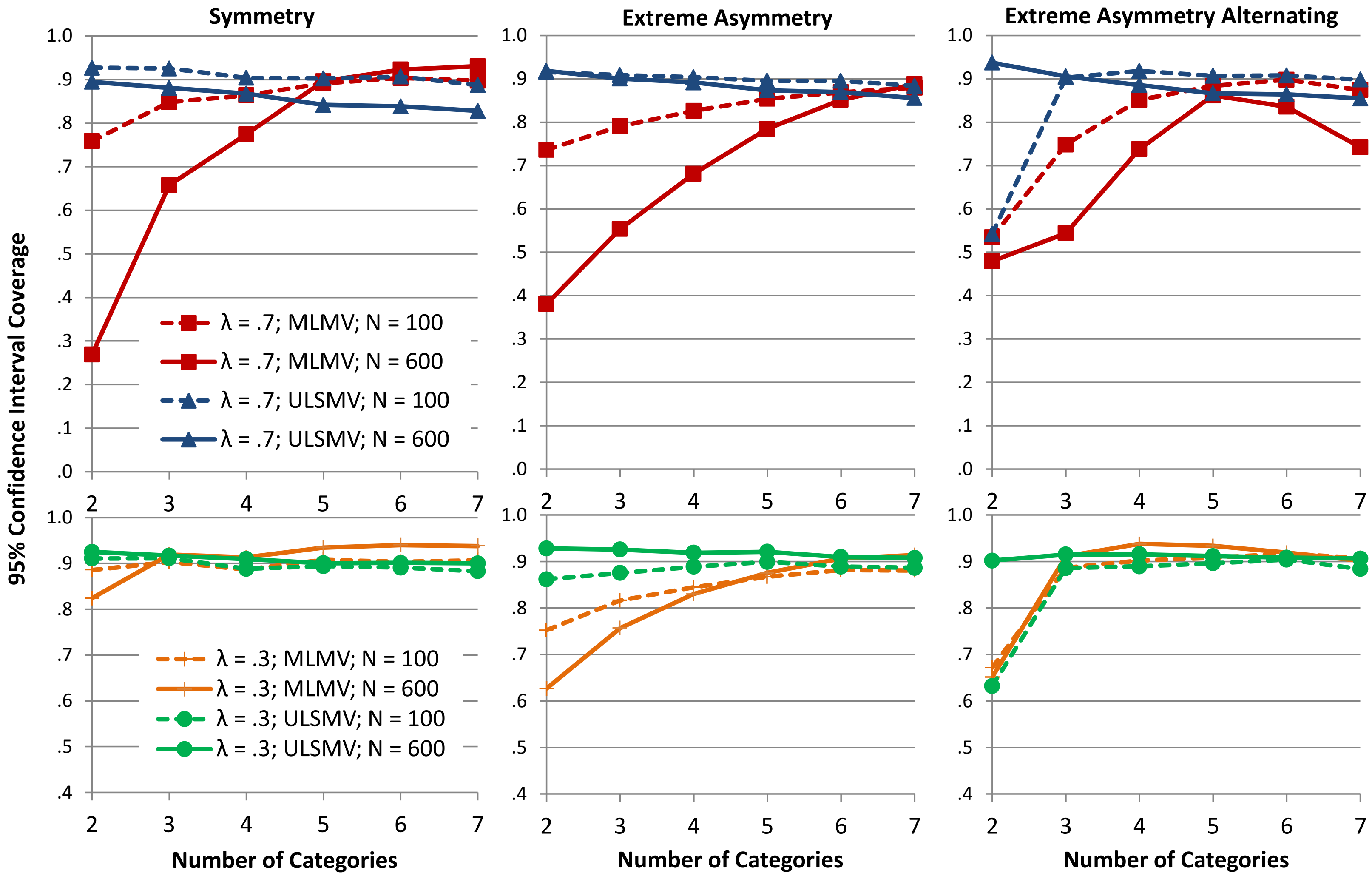

Figure 7. Coverage by number of categories (.7 and .3 factor loadings), underlying distribution is nonnormal (skew 2, kurtosis 7). Values presented correspond to conditions in which the underlying distribution is nonnormal (S2, K7). Values are averaged across model size, and across all loadings for which the true parameter value was the same. The upper panel represents results for a true parameter value of .7. The lower panel represents results for a true parameter value of .3. Lines represent different estimators and different sample sizes (see legend). Vertical panels represent different levels of threshold symmetry.

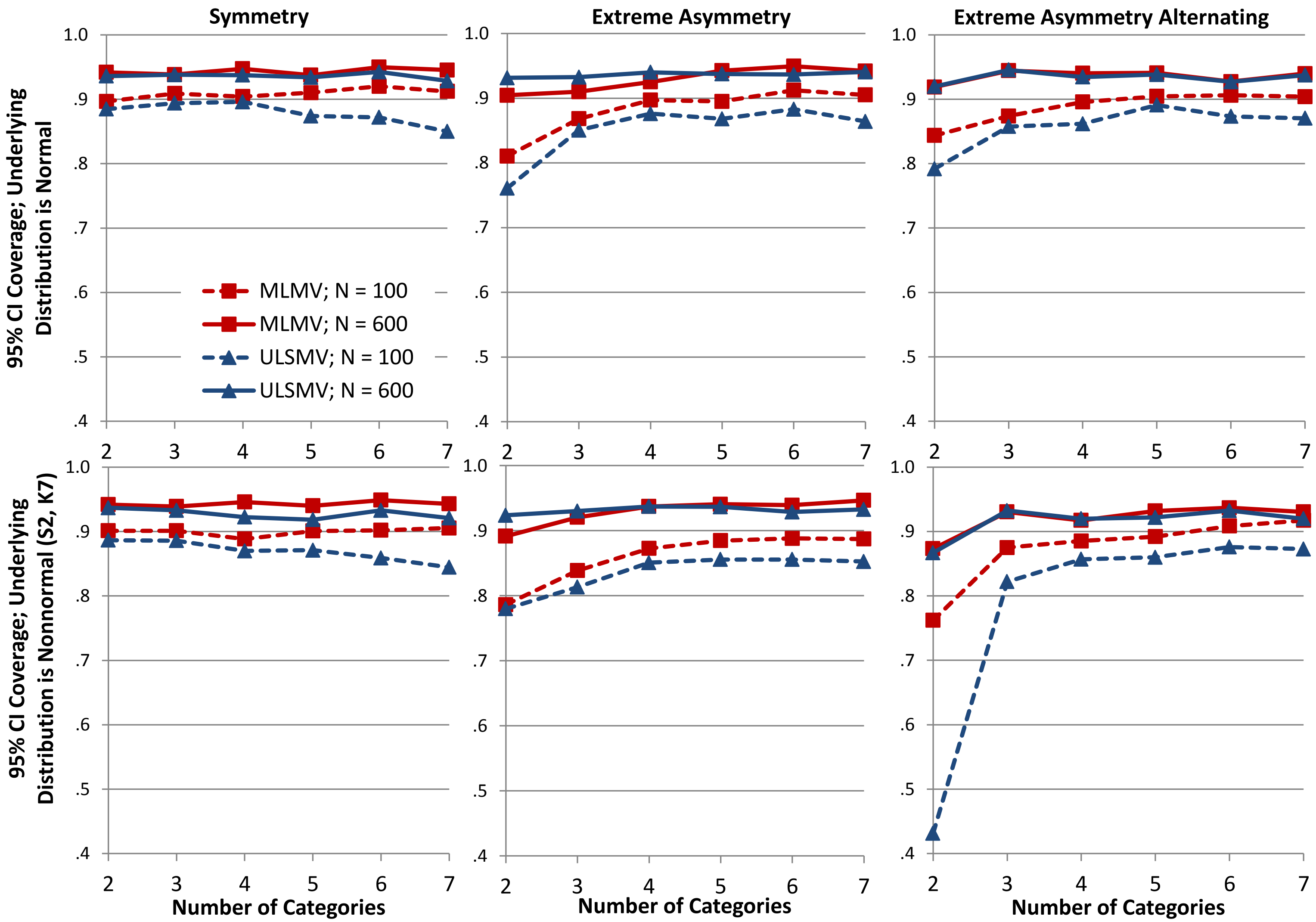

Figure 8. Coverage by number of categories (factor correlation). Values are averaged across model size. The upper panel corresponds to conditions in which the underlying distribution is normal; the lower panel corresponds to conditions in which the underlying distribution is nonnormal (S2, K7). Lines represent different estimators and different sample sizes (see legend). Vertical panels represent different levels of threshold symmetry.

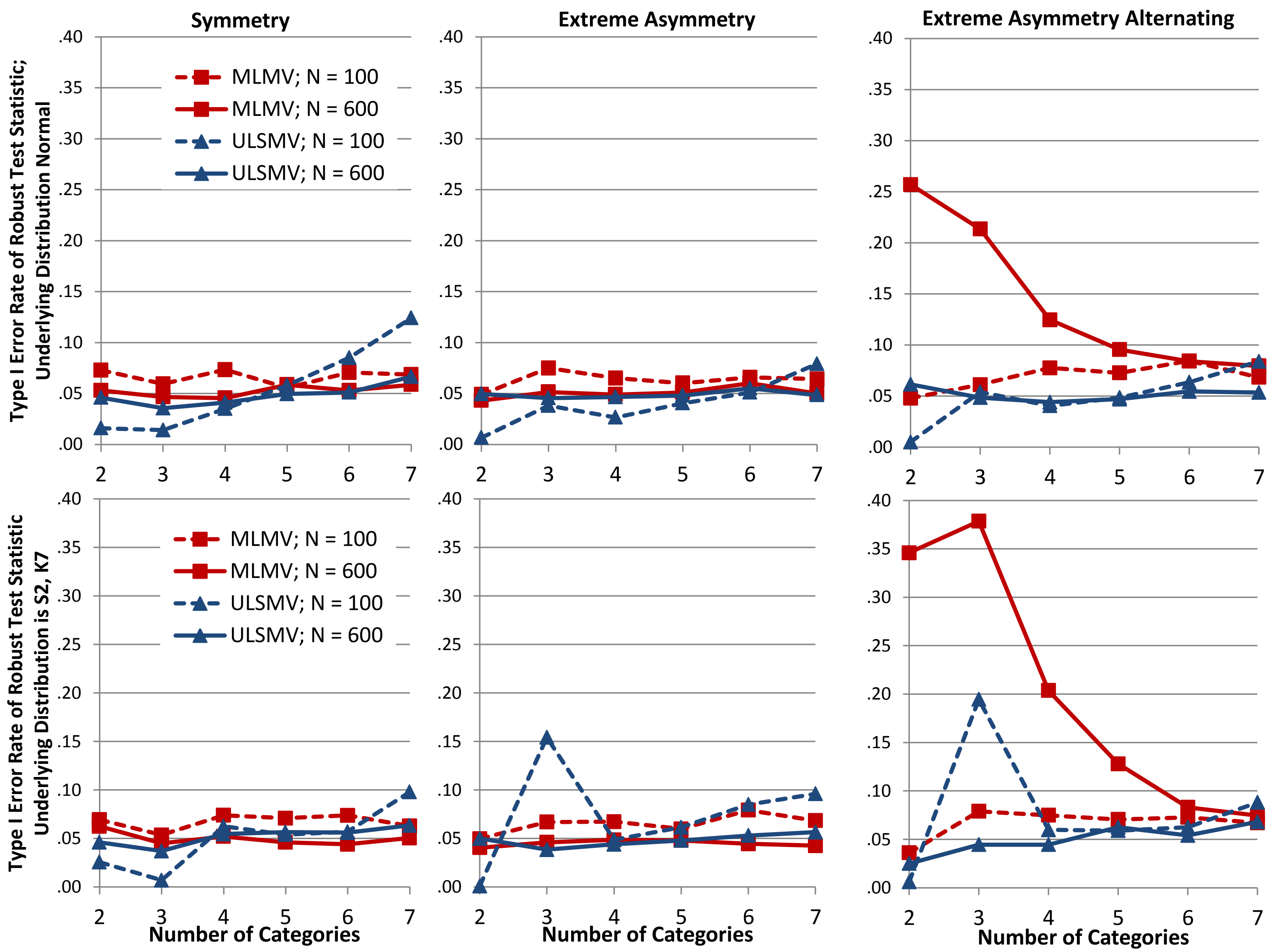

Figure 9. Type I error of mean-and-variance adjusted test statistic by number of categories. Values are averaged across model size. The upper panel corresponds to conditions in which the underlying distribution is normal; the lower panel corresponds to conditions in which the underlying distribution is nonnormal (S2, K7). Lines represent different estimators and different sample sizes (see legend). Vertical panels represent different levels of threshold symmetry.