The Effects of Collapsing Ordered Categorical Variables on Tests of Measurement Invariance

Leslie Rutkowski

Centre for Educational Measurement at the University of Oslo

Indiana University

Dubravka Svetina

Indiana University

Yuan-Ling Liaw

Centre for Educational Measurement at the University of Oslo

Author Note

This research was supported in part by a grant from the Norwegian Research Council under the FINNUT program (grant number 255246).

Correspondence should be addressed to Leslie Rutkowski, Department of Counseling and Educational Psychology, Indiana University, lrutkows@iu.edu; 812-856-8261.

Abstract

Cross-cultural comparisons of latent variable means demands equivalent loadings and intercepts or thresholds. Although equivalence generally emphasizes items as originally designed, researchers sometimes modify response options in categorical items. For example, substantive research interests drive decisions to reduce the number of item categories. Further, categorical multiple-group confirmatory factor analysis (MG-CFA) methods generally require that the number of indicator categories is equal across groups; however, categories with few observations in at least one group can cause challenges. In the current paper, we examine the impact of collapsing ordinal response categories in MG-CFA. An empirical analysis and a complementary simulation study suggests meaningful impacts on model fit due to collapsing categories. We also found reduced scale reliability, measured as a function of Fisher's information. Our findings further illustrate artifactual fit improvement, pointing to the possibility of data dredging for improved model-data consistency in challenging invariance contexts with large numbers of groups.

Keywords: multiple-groups models; ordinal variables; collapsing categories; model fit

The Effects of Collapsing Ordered Categorical Variables on Tests of Measurement Invariance

The operational work of demonstrating measurement equivalence (He & van de Vijver, 2013; Meredith, 1993; Steenkamp & Baumgartner, 1998) in large-scale, cross-national studies, such as the Programme for International Student Assessment (PISA) is complicated by the fact that many dozens of countries and non-national educational systems take part (OECD, 2017c, p. 2). Further, the categorical nature of most questionnaire items in such international studies – typically with two to five categories – adds further methodological complexity in that linear multiple group models are not well suited to the task (Lubke & Muthén, 2004). Rather, categorical models – that respect the scale of the indicator variables – are a preferable and increasingly used option. In this setting, discrete, ordinal points along a continuum represent response options. For example, *strongly agree* represents a higher magnitude of agreement than *agree*. Occasionally, however, response options are combined to create fewer categories. Two primary reasons for collapsing include substantive motives (e.g., to better align the definition of a scale to other research; OECD, 2017a) or when a category is infrequently or not used (Agresti, 2013, p. 77). The latter case poses challenges when using multiple group confirmatory factor analysis (MG-CFA) or item response theory (IRT) to establish measurement invariance since the number of categories for a given item should generally be the same across countries (e.g., a four category agreement item must have four categories across all measured groups). Exceptions to this requirement includes using the EQS software (Bentler, 2000-2008) or pattern mixture modeling with known classes (Widaman, Grimm, Early, Robins, & Conger, 2013).

Regardless of the motivation for combining categories, the impact of this is understood to a certain degree in the single-group context (Muraki, 1993; Strömberg, 1996). In particular, when categories are collapsed, research points to a loss of statistical information (Muraki), which

is directly related to item/scale reliability and measurement precision (Embretson & Reise, 2000, p. 184). Reduced power and incorrect inferences can also result from collapsing categories (Strömberg, 1996). Further, collapsing items from polytomous into dichotomous produced model convergence problems in at least one study (Savalei & Rhemtulla, 2013). In contrast, compared to modeling ordinal data as continuous, collapsing categories had negligible impact on model fit in one study (Tueller et al., 2016). Less understood is the impact of collapsing categories in a multiple-group setting. Furthermore, neither the potential loss of information (and related impact on reliability) nor change in model fit that arises out of substantive research decisions (e.g., collapsing an item that measures frequency of occurrence into *sometimes*/*never* vs. *frequently/always*) has been studied. To that end, we draw on categorical MG-CFA as well as equivalencies between CFA and item response theory (Takane & de Leeuw, 1987) to examine the effects of collapsing categories in multiple-groups contexts. We use an empirical example from a well-known international study (the 2011 Trends in International Mathematics and Science Study [TIMSS]) and a simulation to study the problem. As part of our analysis, we distinguish between collapsing *purposes* (substantively driven or due to infrequently used categories) by including small sample size components to the empirical example and simulation such that infrequently used categories are at issue.

Background

MG-CFA. The multiple-group extension of categorical measurement models is less widely discussed than the single-group case (Millsap, 2011, p. 126); nevertheless, the methods for investigating measurement invariance in the case of categorical observed variables is well-established (Millsap, 2011; B. O. Muthén & Asparouhov, 2002; B. O. Muthén & Christoffersson,

1981). Here, we briefly review the single group approach and extend this explanation to the multiple-group case subsequently. We begin with the assumption that a $p \times 1$ vector of observed variables, $\mathbf{X}$, take discrete ordered values $0, 1, 2, \ldots, C$. Further, it is assumed that for each $X_j, j = 1, 2, \ldots, p$ there is an underlying continuous latent response variable, X_j^*, the value of which determines the observed category of X_j. And X_j^* is related to X_j through a set of C threshold parameters, $\boldsymbol{\nu}_j = (\nu_{j0}, \nu_{j1}, \ldots, \nu_{jC+1})$ where $\nu_{j0} = -\infty$ and $\nu_{jC+1} = \infty$. Then, the probability that $X_j = c$ is given as

$$P(X_j = c) = P(\nu_{jc} \leq X_j^* \leq \nu_{jc+1}) \tag{1}$$

for $c = 0, 1, \ldots, C$. The model for the vector of latent response variables is given as

$$\mathbf{X}^* = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{2}$$

where, $\boldsymbol{\Lambda}$ represents the matrix of factor loadings that relate the vector of latent variables, $\boldsymbol{\xi}$, to the latent response variates and $\boldsymbol{\tau}$ is a vector of latent intercept parameters assumed to be zero for identification purposes. Finally, $\boldsymbol{\delta}$ is a vector of measurement errors in $\mathbf{X}^*$, with $E(\boldsymbol{\delta}) = \mathbf{0}$. The mean and covariance structure of this model are given as

$$\mathrm{E}(\mathbf{X}^*) = \boldsymbol{\mu}^* = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa}, \;\; \mathrm{Cov}(\mathbf{X}^*) = \boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{3}$$

with $\mathrm{E}(\boldsymbol{\xi}) = \boldsymbol{\kappa} = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Phi}$, and $\mathrm{Cov}(\boldsymbol{\delta}) = \boldsymbol{\Theta}$. By assumption, $\mathrm{E}(\mathbf{X}^*) = \mathbf{0}$. A further identification restriction is that the latent response variables have unit variances, which implies that $\boldsymbol{\Theta}$ is estimated as a remainder (Millsap, 2011, p. 128; Muthén & Asparouhov, 2002). When $C > 1$, which we assume here, the correlation between the latent response variables is a tetrachoric correlation. The parameters to be estimated in this single group model include $\boldsymbol{\nu}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}$, and $\boldsymbol{\Theta}$.

This discrete factor model can be easily extended to the multiple-group case by allowing for separate thresholds and covariance matrices of the latent response variables for each population, that is $\boldsymbol{\nu}^{(k)}$and $\boldsymbol{\Sigma}^{*(k)}$, with $k = 1, 2, \ldots, K$ (note that $\boldsymbol{\tau}^{(k)} = \mathbf{0}$ for all k). Then, a common practical approach is to start with the null hypotheses of equivalent thresholds and covariances, H_0: $\boldsymbol{\Sigma}^{*(1)} = \boldsymbol{\Sigma}^{*(2)} = \cdots = \boldsymbol{\Sigma}^{*(K)}$, $\boldsymbol{\nu}^{*(1)} = \boldsymbol{\nu}^{*(2)} = \cdots = \boldsymbol{\nu}^{*(K)}$. If these hypotheses are rejected, then, typically, a series of hierarchical tests are conducted, proceeding from least to most restrictive. Similar to invariance investigations that assume normally distributed observed variables (e.g., Bollen, 1989; Horn & McArdle, 1992), when observed variables are ordinal, the first test, usually referred to as the "baseline test" evaluates the plausibility of the same form of the models (the same number of latent variables and the same pattern of factor loadings, thresholds, and measurement errors across populations). The second test in the hierarchy examines whether the pattern and value of the *salient* factor loadings (Horn, McArdle, & Mason, 1983) are statistically equal across populations (H_0: $\boldsymbol{\Lambda}^{(1)} = \boldsymbol{\Lambda}^{(2)} = \cdots = \boldsymbol{\Lambda}^{(K)}$). The traditional test is a chi-square difference test with degrees of freedom equal to the number of imposed parameter constraints. The third test is one of equality of thresholds, H_0: $\boldsymbol{\nu}^{(1)} = \boldsymbol{\nu}^{(2)} = \cdots = \boldsymbol{\nu}^{(K)}$. Again, a chi-square difference test with degrees of freedom equal to the number of additional parameter constraints is used to evaluate the tenability of this hypothesis. In addition, a set of fit indices, discussed in further detail in the *Methods* section, can supplement the chi-square difference test.

Implicit in the model specification above is the assumption that the number of categories for a particular item must be the same across compared populations. That is, for each $X_j^{(k)}$, $C^{(k)} = C^{(l)}$ for all k and l. This assumption allows the dimension of $\boldsymbol{\nu}^{*(k)}$ to be equal for all k

which, in turn, allows for a statistical comparison of the equality of thresholds across groups. In practice, this assumption is challenged when one or more groups have a small number of observations in one or more categories for at least one item, leading to problems with model estimation and convergence (Savalei & Rhemtulla, 2013; Tueller et al., 2016). Consequently, the typical choice is to collapse the offending category with an adjacent category for all groups. In a similar vein, researchers sometimes choose to collapse categories in substantively meaningful ways (OECD, 2017a, p. 253). For example, a four point scale that measures agreement from *strongly disagree* to *strongly agree* might – based on a researcher's interests – be collapsed into a dichotomous measure of agreement vs. disagreement.

Scale information. In the categorical indicator case, it is straightforward to show that equivalencies exist between CFA and IRT model parameters (Kamata & Bauer, 2008; Meade & Lautenschlager, 2004; Muthén & Asparouhov, 2002; Takane & de Leeuw, 1987). As such, IRT traditions are useful in this context, where the concept of test or scale *information* is useful. In the single-factor (e.g., unidimensional) dichotomous case, a general formula for an item information curve is:

$$I(\xi) = \frac{P_i^*(\xi)^2}{P_i(\xi)(1 - P_i(\xi))}$$

where $P_i(\xi)$ – or item response curve – gives the conditional probability of endorsing item *i* and $P_i^*(\xi)$ is the first derivative of the item response curve with respect to ξ. For binary items, the shape of the information function depends on the factor loading (related to discrimination), the residual variance, and the conditional variance at each ξ level. The greater the discriminating power and the smaller the variance, the greater the information, and, hence, the smaller the standard error of the measurement. For polytomous items, each category provides information

and the item information curve can have multiple peaks (Samejima, 1997). Each item gives maximal information near its location (difficulty) parameter or thresholds.

Due to the local independence assumption of IRT models (Hambleton & Swaminathan, 1985) individual item information can be summed to form the test or scale information function (Lord, 1980). The amount of information provided by a set of items at a given trait level is inversely and directly related to the standard error of measurement at ξ: $SE(\xi) = 1/\sqrt{TI(\xi)}$, where TI indicates *test or scale information*. Plotting the amount of scale information against ξ yields a scale information curve and indicates the range of ξ where a scale is best at discriminating among individuals with more or less of the latent trait and where measurement is most precise.

Considering a single-group case under an IRT framework, collapsing over categories is known to impact measurement precision due to a decrease in scale information and an associated increase in the standard error of estimation (Muraki, 1993). The effect of reduced scale information is that respondents are not well (e.g., precisely) measured at points along the latent variable continuum where information is low (Embretson & Reise, 2000). In an international context with dozens of populations, it is reasonable that the effect could be more severe for some groups.

Empirical Methods

To begin to answer our research question, we used a categorical MG-CFA approach for a single-factor model with ordinal indicators. Model identification was achieved using established methods (Millsap & Yun-Tein, 2004). We supplemented this by fitting equivalent graded response models in an IRT framework to produce information functions under each collapsing strategy. MG-CFA models were fit in M*plus* 7.31 (Muthén & Muthén, 1998) and information

curves were estimated using unconditional maximum likelihood via an expectation-maximization algorithm within the *mirt* package (version 1.20.1, Chalmers, 2012) in *R* (*R* Development Core Team, 2016).

Data

To demonstrate the effect of collapsing adjacent categories for substantive purposes (termed *substantive collapsing*), we use the fourth grade data from the 2011 TIMSS. TIMSS, administered every four years in fourth and eighth grades, is an internationally comparable assessment of mathematics and science. The TIMSS design features a two-stage stratified sampling design where schools are drawn with probability proportional to size and one or two intact classrooms are chosen to participate. TIMSS also collects a wealth of background data from students, teachers, and principals of participating schools. For the current analysis, we use a six-item bullying scale that asks students about the frequency of various bullying victimization behaviors. All relevant items are situated under a single stem that asks "During this year, how often have any of the following things happened to you at school?" The individual items include (1) "I was made fun of or called names;" (2) "I was left out of games or activities by other students;" (3) "Someone spread lies about me; (4) "Something was stolen from me"; (5) "I was hit or hurt by other students(s) (e.g. shoving, hitting, kicking);" and (6) "I was made to do things I didn't want to do by other students." Students respond to one of four item category options: *never*, *a few times a year*, *once or twice a month*, and *at least once a week*, coded 0 to 3, respectively. The 2011 data include 48 educational systems with an overall sample size of 247,338, and individual educational system sample sizes that range from 3,056 (Norway) to 14,394 (United Arab Emirates). Interested readers will find complete details for this scale and the full study in the technical documentation (Martin & Mullis, 2012).

To demonstrate the effect of collapsing due to infrequently used categories (termed *infrequent collapsing*), we randomly sampled 10% of observations within each educational system. This reduced sample size, more representative of situations observed in field trials (OECD, 2014, 2017b), created conditions where infrequently used categories made collapsing desirable. Under this scheme, educational system sample sizes range from 305 to 1443. Supplementary materials [1] contain cell counts for each educational system, item, and category. Here, we can see several instances of infrequently used categories (e.g., item R09D option 2 and 3 in Denmark; item R09F option 1 in Georgia).

Analysis

For both empirical examples, we follow a typical MG-CFA strategy, first fitting a baseline model, followed by a model of equal loadings, and a model of equal loadings and thresholds. For the purposes of this paper, we do not take into account the sampling scheme. Overall model-data consistency was evaluated by the chi-square test statistic; however, the sensitivity of the chi-square difference test to sample size is well-known (Bagozzi, 1977; Bentler & Bonett, 1980). As such, we also considered the root-mean squared error of approximation (RMSEA; Steiger & Lind, 1980). Models with non-significant chi-square statistics and RMSEA $\leq .05$ are considered to be well fitting. Although we also inspect the comparative fit index (CFI; Bentler, 1990) and the Tucker-Lewis index (TLI), recent research uncovered serious deficiencies associated with these measures as overall measures of fit in this setting (Rutkowski & Svetina, 2014, 2017).

[1] Appendices (as PDF) and supplementary materials (as .zip), including simulation and analysis files, can be found at http://hdl.handle.net/2022/22509. The readme file (READ_ME.docx) within the All_Supplementary_Materials.zip folder includes detailed information for each folder and file.

We use an incremental fit strategy to determine whether imposing additional constraints on the models are tenable. In particular, we use a non-significant adjusted chi-square difference test (Satorra & Bentler, 2001); ΔCFI no less than -.020 for tests of equal loadings and -.010 for tests of equal loadings and thresholds (Cheung & Rensvold, 2002). We rely on ΔRMSEA no larger than .030 for slope equality (Rutkowski & Svetina, 2017) and no larger than .010 for slope and threshold equality (Chen, 2007). In addition to the multiple-group analysis, we also fit graded response IRT models (Samejima, 1997) to each group as a straightforward means of producing information functions. We only present scale information for the baseline model.

For the *substantive collapsing* example, across all 48 educational systems, we refit the series of nested models of invariance under four different collapsing strategies: original coding with four categories and collapsing all possible pairs of adjacent categories (e.g., 0 with 1, 1 with 2, or 2 with 3). Adjacent category collapsing is similar to an approach used to model bullying data in another international assessment (OECD, 2017a). We report average overall and incremental fit indices across collapsing strategies. To summarize the large number of results in a digestible way, we fit one-way repeated measures ANOVA models[2] to (a) the total area under the scale information curve; (b) the maximum of the information function (i.e., location where trait is most precisely measured); and (c) the value of the latent trait at which the information function is maximal. The collapsing strategy is used as a factor. We use Tukey's honest significant difference (HSD) test for multiple comparisons across levels of the factor.

For the *infrequent collapsing* example, we only collapse variables where observations per category are fewer than 10 in at least one country, recognizing that this threshold is arbitrary.

[2] To account for the correlation induced by analyzing the same groups (countries) across several different collapsing strategies (conditions).

Based on our 10% sampling strategy, this produced small cell counts for variable R09D in Denmark, Finland, and the Netherlands and for variable R09F in Croatia, Germany, Georgia, and Sweden. In particular, it was necessary to combine categories 2 and 3 for R09D and categories 1, 2, and 3 for R09F.

Empirical Results

For the *substantive collapsing* analysis, Table 1 includes a summary of model fit. With respect to tests of invariance, we found that collapsing categories 1 and 0 produced smaller chi-square tests of model fit and smaller RMSEA values at each level of invariance than not collapsing and all other collapsing strategies. Similarly, chi-square difference test values and absolute values of ΔCFI and ΔRMSEA were smaller in this condition. Indeed, taking a relatively liberal view in terms of fit indices and discounting the chi-square as sensitive to sample size, it could be argued that this strategy met reasonable criteria for assuming equal loadings and thresholds. It is important to note, however, that this improvement in fit does not reflect the measurement properties of the original scale. Rather, it is an artifact of combining these adjacent categories.

Insert Table 1 about here

Table 2 contains ANOVA results for scale information for the *substantive collapsing* analysis. In particular, the total information, as measured by the area under the information

curve, was considerably larger when categories were left intact compared to all adjacent category collapsing strategies. And the maximum value of the information function differed across all collapsing strategies compared with not collapsing. Specifically, combining category 3 and 2 or category 2 and 1 produced lower maximum scale information (*diff* column). Taken together, these findings suggest that precision about a particular range of the latent trait decreased and that overall measurement precision was reduced when compared to not collapsing. Although overall information was less when category 1 and 0 were combined, the average maximum value of the scale information function compared increased. These findings indicate that although overall precision fell, the ability of the scale to precisely measure at a particular area along the latent trait continuum – the location of which is elaborated next – increased. Again, this result is an artifact of collapsing. Finally, the trait level associated with the information maximum was significantly different for all collapsing strategies compared to not collapsing. In particular, combining category 1 and 0 or 2 and 1 shifted the trait value higher on the continuum than not collapsing while combining category 3 and 2 shifted the trait value lower on the latent continuum.

Insert Table 2 about here

Results of the *infrequent collapsing* analysis, based on a random sample of 10% of cases within each system, are located in Table 3. In the interest of space and because the pattern of changes in information is similar to the full data example (especially, information is lost and the location where the maximum is located changes), we do not present results for scale information. With respect to measurement invariance findings, a few results are worth noting. First, based on

overall and incremental fit indices, the baseline model and model of equal slopes and thresholds exhibited similar fit to the full dataset. This was reasonable, given that a random sample was drawn. In contrast to the full dataset, the model of equal thresholds and slopes did not converge in the non-collapsed example, likely due to small cell counts. By comparison, the collapsed baseline model fit similar to the non-collapsed baseline model. Because we collapsed variable 6 into just two categories, we could not fit a model that assumed equal loadings. The model of equal loadings and thresholds exhibited a meaningful deterioration in fit over the baseline model based on the overall and incremental fit measures; however, in contrast to the non-collapsed condition, this model converged normally. Nevertheless, the fit of the collapsed equal loadings and thresholds model was poor based on overall and incremental fit measures.

Insert Table 3 about here

Simulation Study Methods

To derive generating parameters, we first fit individual unidimensional categorical CFA models to the 47 educational systems from the empirical analysis (due to convergence issues with simulated data, Botswana was removed from the analysis). Based on the empirical results, we then calculated the mean and variance of each of the parameter estimates (i.e., slopes, thresholds, residual variance, latent variable means and variance) across countries. These means and variances served as the generating distributions, from which we sampled for each of the 47 populations for our simulation. Specifically, for the baseline condition, we drew parameters at

random for each of the 47 simulated populations. We then drew slope parameters for each item that were set as equal across the 47 simulated populations. Finally, we drew threshold parameters for each item that were also set equal across our simulated populations. Generating parameter distributions are located in Appendix A, Table A1. Based on these generating parameters, we generated data stemming from three conditions: same form (baseline), equal slopes, equal slopes and thresholds with sample sizes that matched our empirical conditions. Under each generating condition, data were simulated 500 times to evaluate the stability of our results. These procedures were followed for both the *substantive collapsing* and *infrequent collapsing* simulations.

For the *substantive collapsing* simulations, variables were collapsed into all possible adjacent combinations of three (e.g., 1 with 0; 2 with 1; or 3 with 2) and two categories (e.g., 0 vs. 3, 2, and 1; 0 and 1 vs. 2 and 3; 0, 1, and 2 vs. 3). Under each collapsing strategy condition and across replications, we refit the series of nested invariance models, first fitting a baseline model, followed by a model of equal slopes, and a model of equal slopes and thresholds. Data were simulated and models were fit to simulated data with *Mplus 7.4* (Muthén & Muthén, 1998). In addition to the multiple-group analysis, we also fit graded response IRT model to each group using the *mirt* package (Version 1.25; Chalmers et al., 2017) in *R*. Analyses, across 500 replications, are similar in fashion to the empirical analysis: we summarize the model fit results overall and for tests of invariance. We also produce repeated measures ANOVAs for scale information with multiple comparisons across collapsing categories as a factor.

Simulation and analysis procedures for the *infrequent collapsing* simulation were identical to the *substantive collapsing* simulation. We made this decision for the following reasons. In contrast to the small sample empirical analysis, where we only collapsed across item

categories with small counts, the occurrence of small cell counts could differ in any group-item-category combination across replications. This could create a situation where replication-specific collapsing decisions might number as high as 47 groups x 6 items x 4 categories (with 6 possible ways of collapsing) x 500 replications = 564,000. Besides the large numbers of possible collapsing decisions, this strategy would also prevent detecting patterns of fit or information changes.

Simulation Results

Substantive Collapsing

We present results that are similar in content to the empirical analysis; however, in this case, we report for each generating model (baseline, equal slopes, and equal slopes and thresholds). Results are averaged across replications. Tables 4 and 5 include a summary of model fit and Table 6 contains ANOVA results for scale information that is relevant to the *substantive collapsing* simulation. We focus on collapsing strategies that are different from not collapsing at all. Appendix B contains plots of average fit values across the simulation conditions and invariance tests. Appendix C contains plots of changes in information by generating model, collapsing strategy, and country. In the interest of space and because the results are qualitatively similar, we do not include these figures for the small simulation study.

When data were generated under a baseline assumption (Table 4), we found that conditions 3a, 3c, 2b, and 2c produced better fit in terms of the overall and incremental CFI and RMSEA. These changes, however, were not enough to result in a different decision than not collapsing (e.g., that an assumption of same form invariance is the most stringent supported by the data). In all other conditions, the fit measures produced expected results – all hypotheses

except that of same form were rejected. When data were generated with equal loadings (Table 5), collapsing conditions 3a, 2a, and 2c produced better fit. Further, condition 2a resulted in evidence of more stringent invariance. In particular, collapsing resulted in accepting the hypothesis of equal loadings and thresholds, with overall and incremental RMSEA and CFI falling within accepted cutoffs. In contrast, conditions 3b and 3c produced results that would lead to rejecting an equal loading assumption, with, especially, incremental fit indices outside of normal cut-offs. As in the empirical analysis, these changes in conclusion are a direct artifact of collapsing. Finally, when data were generated to have equal slopes and thresholds, in all conditions, the models fit the data well and equal loadings and thresholds were consistently justified. As such, we did not provide a table of results in the interest of space.

In Tables 4 and 5 we note one particularly interesting finding – across several conditions, including when response categories were not collapsed, the number of replications that did not converge was higher for invariance tests that were far from the generating model. For instance, when data were generated under a model of same form but tested for equal loadings and thresholds, only 360 of 500 replications converged when the data were not collapsed and just 117 replications converged in binary condition 2c. Condition 2c produced just 112 convergent solutions when data were generated to have equal slopes and tested for equal loadings and thresholds. In contrast, we did not observe the same issues with convergence when the generating model assumed equal loadings and thresholds, no matter the collapsing strategy.

Tables 4 and 5 about here

Table 6 summarizes the simulation findings that pertain to scale information for graded IRT models. In all cases, the total scale information and information at the maximum was significantly higher under the no collapsing strategy compared to all other strategies, with the largest differences in scale information between not collapsing and the binary conditions. In a complementary fashion, these findings suggested that measurement precision was reduced when categories were collapsed, with reductions in precision most severe for the binary conditions. Finally, the average trait level shifted predictably when data were collapsed. That is, when the highest categories were collapsed, peak information and measurement precision shifted toward lower latent trait levels and vice versa. Strikingly, the latent trait value that maximized information was shifted by 81% of a standard deviation when categories 1 through 3 were combined. This suggested that a very different point along the latent trait continuum was being measured with the most precision under such a strategy.

Table 6 about here

Infrequent Collapsing

In general, findings for infrequent collapsing were in line with the substantive collapsing simulation. In the baseline generating condition (Table 7), not collapsing produced average overall and incremental fit statistics that supported the hypothesis of same form but rejected the hypothesis of equal slopes or equal slopes and thresholds when the data were not collapsed. In contrast, conditions 3a, 3b, and 2c produced higher overall CFI/TLI, lower overall RMSEA, and

smaller absolute values of incremental fit indices than not collapsing. We note in many conditions, high percentages of replications did not converge. For example, when not collapsing, only 166 replications converged for the test of equal loadings and thresholds. At the extreme, just 43 replications converged in condition 2c.

Table 8 summarizes the results of the slope equality-generating model. Although most results were highly similar, we note a few consequential differences. For example, in condition 3b, overall and incremental fit indices for the assumption of equal slopes were larger than condition 4; however, they fell within acceptable cut-offs, implying the same substantive conclusions about model fit. This change, in the form of a slightly smaller ΔRMSEA, leads to support for equal loadings and thresholds – again, an artifact of collapsing and not a reflection of actual measurement properties. As in the substantive collapsing results, the equal slopes and thresholds generating model produced results in support of equal slope and threshold parameters, regardless of collapsing decision. Again, we do not present those results in the interest of space.

Finally, changes to scale information, located in Table 9, were nearly identical to the substantive collapsing results. In particular, collapsing generally reduced information and predictably shifted the location of highest precision.

Discussion

Using empirical data from TIMSS, a well-known international survey of educational achievement, and simulated data, we probed the impact of collapsing adjacent categories on tests of measurement invariance and scale information. We approached our investigation by borrowing from two closely related measurement traditions: CFA and IRT. By capitalizing on equivalencies between these two methods, we borrowed concepts from both that helped elucidate

our research question of the impact of collapsing categories in establishing measurement invariance. To that end, we used a categorical MG-CFA approach and we supplemented this analysis by drawing on scale information curves, a typical tool used within the IRT tradition. Under several typical conditions IRT and categorical CFA are essentially reparameterizations of the same model; however, the two paradigms have much to offer independently and tools from each can be borrowed to provide richer information about the adequacy and performance of a given model.

In terms of between-group tests of parameter equality, we found that some collapsing strategies produced different invariance results and that overall conclusions about measurement invariance differed – combining category 1 with 0 in the empirical example and in the substantive collapsing simulation condition, 2a, where categories 1 through 3 were combined, led to accepting the hypothesis of equal loadings and thresholds when the generating data came from a model of equal loadings. In contrast, collapsing also led to rejecting the assumption of slope equality in spite of a commensurate data-generating model. Importantly, in situations where changes in fit supported more stringent conclusions about invariance, this impression was merely an artifact of collapsing categories. In the case of collapsing on substantive grounds, these findings pointed to the possibility of "p-hacking" (Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014) or data dredging improved invariance results (whether intentional or not).

Although the possibility of more stringent invariance conclusions make collapsing an inviting option, especially when the analyst is searching for evidence of invariance across dozens of countries or populations, we advise against this practice, as any conclusions and associated claims do not extend to the generating model. Rather, artifactual improvements are limited to the

collapsed data, raising questions about construct validity and generalizability of results. We also note that in several situations, we encountered convergence issues that were the result of fitting models that were not supported by the generating data. This was also true under some empirical collapsing strategies. As such, our findings point to convergence failure as an additional indicator of poor model-data consistency, beyond typical fit measures. Such convergence failures could reasonably be called "catastrophic misfit."

Collapsing categories had consistent impacts on scale information (reliability) across all data generating models. As expected, overall information decreased when any collapsing decision was made, and the level and location of precision were affected by collapsing decisions. In particular, leaving data intact generally produced maximum information, in terms of area and height of the information curve. The trait level at which the maximum information occurred also depended on collapsing strategies, according to our simulation and empirical results.

We highlight condition 2a in the substantive collapsing simulation as one example. In this condition, categories 1, 2, and 3 are combined, which produced less overall information for the scale and a slightly lower level of information at the maximum point, which was lower on the trait continuum. Put in the context of the empirical example on bullying, such a finding suggests that combining self-reported levels of less frequent bullying produces overall less certainty about the trait along the whole continuum, with the greatest certainty at some lower level of bullying, where the information maximum occurs. Under an equal slopes generating model, condition 2a also produced better invariance results than the no-collapsing condition. Based on results from the MG-CFA approach, it would appear that pursuing such a collapsing strategy would have little negative consequence in terms tests of invariance.

In some respects, we agree with this statement. However, it is important to first consider what is intended by measuring and reporting on any latent trait – in this example, bullying. Although a more stringent level of invariance is achieved by combining particular categories, the consequence is that the point at which measurement is more precise is, on average, at lower levels of bullying. If the research or reporting emphasis is on average levels of bullying, inferences about individuals with average or higher levels of the latent trait can be compromised. Further, since the shape of the information function will necessarily be more peaked – based on less total information and but a similar maximum height – the interval where the latent trait is precisely measured will be narrower. Finally, these findings are confined to the collapsed data and do not reflect measurement properties of the original scale, as designed and administered.

In terms of simulation results where collapsing could ostensibly be justified as a way to treat small cell counts and the potential convergence issues that can attend, we note that such decisions can impact conclusions about the level of tenable invariance. Treating small cell counts in this way is common practice, especially at the field trial stage in international assessments, where sample sizes are relatively small. However, it is important to note that conclusions might not necessarily reflect the actual measurement properties of an instrument. Rather, any conclusions that result from treating small cell counts in this way should be carefully validated once larger sample sizes – typically available from the main survey – are available.

An additional finding worth emphasizing is that convergence failures frequently resulted when the assumed model did not fit the generating data. In fact, in a number of conditions across both simulations, fitted models that were more stringent than, especially, a baseline generating data model, produced high rates of non-convergent models. We observed this in both simulations; however, the effects were more extreme in smaller samples and there were

occurrences in even the slope equality generating model. As one example, we point to condition 2c in the substantively collapsing simulation under a baseline generating model, where 383 replications did not converge for the test of equal slopes and thresholds. This points to one additional means of detecting a poor fitting model, especially given that no model convergence issues presented in situations where the generating and estimated model matched. Certainly, however, many false negatives slip through by this criterion – condition 2a in Table 5 exhibited full convergence *and* support for equal slopes and thresholds in spite of an equal slopes generating model. Nevertheless, non-convergence, when it occurs, suggests "catastrophic misfit."

Given the growth in international surveys and an increased emphasis on and awareness of the importance of establishing measurement invariance in these settings (OECD, 2014, 2017a), researchers are often required to engage in formal tests of parameter equivalence with large numbers of groups. Further, as surveys such as PISA, TIMSS, and others beyond education feature items that are ordinal in nature, categorical multiple-groups analysis is gaining in popularity. But the very nature of the data place constraints on these methods, including a desire to combine categories for substantive reasons. A recent example includes the OECD's PISA 2015 report on student well-being, where a decision was taken to collapse categories across all items on the bullying scale (OECD, 2017a). Importantly, the authors note that this substantively-driven decision "improved the international invariance of the scale" (p. 253). We again point to issues around construct validity – the choice was made to combine response options "a few times a month" with "once a week or more." Although Solberg and Olweus (2003) argued that a few times a month can serve as a lower bound for identifying students that meet some definition of bullying, the impacts on inferences of losing the finer-grained distinction of a few times a month

versus at least once a week has not been studied, to our knowledge. Further, the effect of these sorts of choices has not been studied in the multiple-groups context, more broadly.

As a concluding point, we note that we did not consider the effect of collapsing categories on a relatively new method of invariance testing, the alignment method (Asparouhov & Muthén, 2014). The reason for this omission is that, currently, operational procedures in large-scale international surveys are limited to traditional measurement invariance testing methods (e.g., MG-CFA). Finally, as noted previously, we did not include the study design (stratification and sampling weights) in the analysis or simulation. As such, future research might consider the extent to which the sampling design impacts the generalizability of our findings.

References

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bagozzi, R. P. (1977). Structural equation models in experimental research. *Journal of Marketing Research*, *14*(2), 209–226. https://doi.org/10.2307/3150471

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (2000-2008). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., & Meade, A. (2017). *mirt: multidimensional item response theory (Version 1.25) [Computer software]*.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahway, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

He, J., & van de Vijver, F. J. (2013). Methodological issues in cross-cultural studies in educational psychology. In *Advancing cross-cultural perspectives on educational psychology, Information Age Publishing, Charlotte, NC* (Vol. Advancing cross-cultural perspectives on educational psychology). Charlotte, NC: Information Age Publishing Inc. Retrieved from http://www.researchgate.net/profile/Fons_Van_de_Vijver/publication/259176988_Methodological_Issues_in_Cross-Cultural_Studies_in_Educational_Psychology/links/02e7e52a2175522d1d000000.pdf

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. https://doi.org/10.1080/03610739208253916

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*, *1*(4), 179–188.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136–153. https://doi.org/10.1080/10705510701758406

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Routledge.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(4), 514–534. https://doi.org/10.1207/s15328007sem1104_2

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. https://doi.org/10.1207/S15327906MBR3903_4

Muraki, E. (1993). *Information functions of the generalized partial credit model* (ETS Research Report Series) (pp. i–12). Princeton, NJ: Educational Testing Service. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1993.tb01538.x/abstract

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Mplus Web Notes No. 4). Los Angeles, CA: University of California, Los Angeles.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*(4), 407–419. https://doi.org/10.1007/BF02293798

Muthén, L., & Muthén, B. O. (1998). *Mplus user's guide* (Seventh edition). Los Angeles, CA: Muthén & Muthén.

Muthén, Linda, & Muthén, B. O. (1998). *Mplus user's guide.* (Seventh edition). Los Angeles, CA: Muthén & Muthén.

OECD. (2014). *TALIS 2013 technical report*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/edu/school/TALIS-technical-report-2013.pdf

OECD. (2017a). *PISA 2015 Results (Volume III): Students' Well Being*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264273856-en

OECD. (2017b). *PISA 2015 technical report*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

OECD. (2017c). Scaling procedures and construct validation of context questionnaire data. In *PISA 2015 technical report*. Paris: OECD Publishing.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*, *30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243540

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_5

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201–223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Solberg, M. E., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior*, *29*(3), 239–268. https://doi.org/10.1002/ab.10047

Steenkamp, J.-B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107.

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. Presented at the Meeting of the Psychometric Society, Iowa City, IA.

Strömberg, U. (1996). Collapsing ordered outcome categories: A note of concern. *American Journal of Epidemiology*, *144*(4), 421–424.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Tueller, S. J., Johnson, K. L., Grimm, K. J., Desmarais, S. L., Sellers, B. G., & Van Dorn, R. A. (2016). Effects of sample size and distributional assumptions on competing models of the factor structure of the PANSS and BPRS. *International Journal of Methods in Psychiatric Research*, *26*(4), 1–10. https://doi.org/10.1002/mpr.1549

Widaman, K. F., Grimm, K. J., Early, D. R., Robins, R. W., & Conger, R. D. (2013). Investigating factorial invariance of latent variables across populations when manifest variables are missing completely. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 384–408. https://doi.org/10.1080/10705511.2013.797819