Journal of Modern Applied Statistical Methods

Volume 7 | Issue 2 Article 12

11-1-2008

Multi-Group Confirmatory Factor Analysis for Testing Measurement Invariance in Mixed Item Format Data

Kim H. Koh
Nanyang Technological University, khkoh@nie.edu.sg

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Recommended Citation
Koh, Kim H. and Zumbo, Bruno D. (2008) "Multi-Group Confirmatory Factor Analysis for Testing Measurement Invariance in Mixed Item Format Data," *Journal of Modern Applied Statistical Methods*: Vol. 7 : Iss. 2 , Article 12.
DOI: 10.22237/jmasm/1225512660

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 471-477

1538 – 9472/08/$95.00

Multi-Group Confirmatory Factor Analysis for Testing Measurement Invariance in Mixed Item Format Data

Kim H. Koh
Nanyang Technological University
Singapore

Bruno D. Zumbo
University of British Columbia
Canada

This simulation study investigated the empirical Type I error rates of using the maximum likelihood estimation method and Pearson covariance matrix for multi-group confirmatory factor analysis (MGCFA) of full and strong measurement invariance hypotheses with mixed item format data that are ordinal in nature. The results indicate that mixed item formats and sample size combinations do not result in inflated empirical Type I error rates for rejecting the true measurement invariance hypotheses. Therefore, although the common methods are in a sense sub-optimal, they don't lead to researchers claiming that measures are functioning differently across groups – i.e., a lack of measurement invariance.

Key words: Multi-Group Confirmatory Factor Analysis, Measurement Invariance, Binary and Ordinal Items.

Introduction

Multi-group confirmatory maximum likelihood factor analysis has become the most commonly used scale-level technique to evaluate measurement invariance/ equivalence of a test across different groups (e.g., gender, language), over different mediums of administration (e.g., web-based versus paper-and-pencil testing), or across accommodated and non-accommodated conditions. Measurement invariance is tenable when the relations between observed variables and latent construct(s) are identical across relevant groups. In particular, individuals with the same standing on a latent variable but sampled from different subpopulations should have the same expected observed score on a test of that variable (Horn and McArdle, 1992). The common understanding in the research literature is that without measurement invariance, observed means (or latent means) are not directly comparable (Drasgow & Kanfer, 1985).

Mixed item format data are often found in educational measurement wherein many classroom and large-scale assessments in use today are blended instruments that include a mixture of multiple-choice and constructed-response items. Typically, multiple-choice items are dichotomously scored and constructed-response items are polytomously (partial-credit) scored. These two types of scores are on an ordinal scale. Two commonly encountered, and interrelated, problems associated with ordinal scale are measurement scale coarseness and multivariate nonnormality. Measurement scale coarseness is caused by a crude classification of the latent variables to ordinal scales with small numbers of response categories. Because of the discrete nature of ordinal scales, the distributions of the response data obtained from dichotomous and polytomous items are not conducive to multivariate normality.

Ideally, data derived from an ordinal scale should be analyzed using estimation methods that are designed for use with such data. Weighted Least Squares (WLS, Jöreskog

Kim H. Koh is Assistant Professor, Centre for Research in Pedagogy and Practice, National Institute of Education. Email: khkoh@nie.edu.sg. Bruno D. Zumbo is Professor of Measurement, Evaluation and Research Methodology, as well as member of the Department of Statistics and the Institute of Applied Mathematics. Email him at: bruno.zumbo@ubc.ca. An earlier version of this article was presented at the 2007 American Educational Research Association (AERA) conference.

& Sörbom, 1996), Asymptotic Distribution Free (ADF, Browne, 1984), or Robust Maximum Likelihood estimation of model parameters using the polychoric correlation and asymptotic covariance matrix is theoretically sound for MGCFA with ordinal and mixed item format data. Practitioners, however, seldom use these methods. The implicit reasoning appears to be two-fold: (a) there is lack of awareness of these relatively new methods, and (b) these new methods are understood to require large sample sizes; larger than ones found in many research settings, and are, generally, not computationally viable with tests or measures involving more than 25 items[1].

Consequently, the ordinal-scaled data are often treated as if they were continuous and analyzed with the normal theory Maximum Likelihood (ML) estimation method and Pearson covariance matrix. The purpose, therefore, of this study was to investigate the statistical properties of the maximum likelihood factor analysis of a Pearson covariance matrix for testing measurement invariance hypotheses in MGCFA with mixed item format data. Specifically, the study examined the effects of mixed item formats and sample size combinations on the Type I error rates of ML-based chi-square difference tests for two commonly investigated measurement invariance hypotheses, namely strong and full invariance.

To be clear, we are not advocating using a Pearson covariance matrix for testing measurement invariance with mixed item formats, but rather we are interested in investigating: (a) what happens to the Type I error rates for those researchers who continue to choose to use these sub-optimal methods, and (b) the empirical Type I error rate of the extant research literature that used these sub-optimal methods (before the more optimal ones were widely available) for measurement invariance. We are also not advocating for the exclusive use of hypothesis testing in this context. Our aim is to reflect common research and applied measurement practice (both in terms of the methods used and the type of data) and hence to document the Type I error rates that one would find in these applied settings. This matter of keeping an eye on everyday research practice will come up again in the Methods Section when we describe the various hypothesis tests we are investigating.

Theoretical Framework

The fundamental idea underlying the measurement models in MGCFA is the use of a set of observable variables (i.e., items) to represent the latent variable(s). When the ordinal-scaled items are used as proxies for the latent continuous variable(s), the assumptions of interval measurement scale and multivariate normality are violated. Measurement errors induced by a crude categorization of the latent continuous variables can lead to the violations of the covariance structure. Because the Pearson covariance is attenuated in the ordinal variables, the covariance structure model may not hold for the observed variables. Therefore, ML estimation based on the distorted sample covariance matrix is likely to be biased.

When ordinal data are used with the ML estimation method and Pearson covariance matrix in single-group confirmatory factor

[1] The WLS/ADF estimation method requires relatively large sample sizes (i.e., at least 2,000-5,000 observations per group, Browne, 1984) to alleviate problems due to convergence or improper solutions and is not a viable method for models with a large number of items. Also, diagonally weighted least squares with the corresponding asymptotic covariance matrix and the polychoric (or tetrachoric) covariance matrix is limited due to the fact that no more than 25 items can be used due to the excessive computer memory demands with the so-called weight matrix, i.e., asymptotic covariance matrix of the vectorized elements of the observed covariance matrix. With p variables there are L elements in the same covariance matrix, and the weight matrix is of order LxL, where L=(p(p+1))/2. Therefore, as an example, for a model that has 20 items, the weight matrix would contain 22,155 distinct elements and for 25 items the weight matrix would contain 52,975 distinct elements. Likewise, the Satorra-Bentler corrected chi-square in LISREL and Muthen's estimation method for ordered categorical data in the software Mplus are also limited by the large number of items that are found in large-scale educational measurement. Therefore, most applied research in MGCFA has ordinal or mixed item format data with small sample sizes and large numbers of items, therefore these computational and statistical restrictions prevent many applied researchers from using the WLS/ADF estimation method.

analysis, the chi-square goodness of fit statistic is inflated due to departures from multivariate normality in the observed variables, albeit negligible bias is found in the model parameter estimates (e.g., Hutchinson & Olmos, 1998; Muthén & Kaplan, 1992; Potthast 1993; Rigdon & Ferguson, 1991). Hence, using the ML chi-square statistic as a formal test statistic of model-data fit under the conditions of multivariate nonnormality leads to an inflated Type I error rate for rejecting a true model.

Methods

Simulation data focused on the situation wherein one has a test with a mixture of dichotomously and polytomously scored items. The design variables were three conditions of mixed item formats and six sample size combinations, resulting in a 3 × 6 factorial design with 18 cells in our simulation experimental design. Within each cell, 100 replications were generated.
A 30 item test was simulated with mixed item formats that were varied according to the proportions of dichotomous and polytomous items as follows:

A. 67% (20) dichotomous items and 33% (10) polytomous items (3 scale points),

B. 50% (15) dichotomous items and 50% (15) polytomous items (3 scale points), and

C. 33% (10) dichotomous items and 67% (20) polytomous items (3 scale points).

These item format proportions reflect the real achievement assessment data found in educational testing contexts such as the Trends in International Mathematics and Science Study (TIMSS) and the National Assessment of Educational Progress (NAEP). Given that most of the achievement data, when partial scores are allotted, use 3-category polytomous items, the polytomous items in the simulation were limited to item responses with 3 scale points.

The sample size combinations consisted of equal and unequal sample sizes for the two groups: 200 vs. 200; 500 vs. 500; 800 vs. 800; 200 vs. 500; 200 vs. 800; and 500 vs. 800. These were the typical sample sizes across two groups used with the ML estimation method and Pearson covariance matrix in MGCFA applied research.

Simulation Procedure

For unidimensional dichotomous items, the item responses were generated from the three-parameter logistic (3PL) item response theory model (Birnbaum, 1968),

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp[-1.7a_i(\theta - b_i)]},$$

where a_i, b_i and c_i are the item i discrimination, difficulty, and guessing parameters, respectively. The $P_i(\theta)$ denotes the probability of answering correctly to item i by a randomly selected examinee with ability θ. The 3PL item parameters a, b, and c of each of the 20 dichotomous items were real item parameter estimates taken from the 1999 *TIMSS Mathematics Achievement Test*.

Using a random number generator to produce numbers uniformly distributed on the interval [0,1], the probabilities were converted to either 0s or 1s to reflect examinee item scores. When the random number selected was less than or equal to $P_i(\theta)$, a 1 was assigned to an examinee for item i, and a 0 otherwise (Hambleton & Rovinelli, 1986).

For the polytomously scored items, the generalized partial credit model (GPCM)(Muraki, 1992) was used to generate unidimensional polytomous item responses, which were categorized into r_i+1 ordered score categories (0, 1, …, r_i) for i-th item. The model states that the probability of getting item score $U_j=q$ for a randomly sampled examinee with ability θ to the i-th item is given by

$$P_{i,q}(\theta) = \Pr ob(U_i = q|\theta) =$$

$$\frac{\exp[\Sigma_{v=0}^{q} 1.7a_i(\theta - b_i + d_{iv})]}{\Sigma_{j=0}^{ri} \exp[\Sigma_{v=0}^{j} 1.7a_i(\theta - b_i + d_{iv})]},$$

$$q = 0,1,\ldots, r_i,$$

where a_i is the slope parameter of item i; b_i is the location parameter of item i; and d_{iv} are a set of threshold parameters of item i with associated constrains $d_{i0}= 0$ and

$\Sigma_{v=1}^{r_i} d_{iv} = 0$ (Muraki, 1992). A total of 20 polytomous item parameters (*a*s, *b*s, *d*s) were obtained from the TIMSS data.

The approach described by González-Romá, Hernández & Gómez-Benito (2002) was used to generate ordered polytomous items. For each examinee, a latent trait estimate θ was generated from a standard normal distribution, *N*(0,1). The GPCM probabilities were summed across categories to create a cumulative probability for each score level, and then the probability of responding above category k $[P_k^*(\theta)]$ was computed. For each simulated item and examinee a single random number (*u*) was randomly sampled from a uniform distribution over the interval [0,1], and the item scores were assigned as follows:

k = 3 if $P_2^*(\theta) \geq u$

k = 2 if $P_2^*(\theta) < u \leq P_1^*(\theta)$

k = 1 if $P_1^*(\theta) < u$.

Two population data were simulated with equivalent parameters to represent measurement invariance. The population data consisted of 20 dichotomous and 20 polytomous items. Data sets with different proportions of dichotomous and polytomous items were then created by a random selection of the items from the first two population data. As can be seen in Table 1, the item response distributions across groups for each of the mixed item format conditions were only slightly negatively skewed.

Table 1: Mean Skewness of the Mixed Item Format Population Data

Mixtures of Item Formats	Mean Skewness
67% Dichotomous and 33% Polytomous Items	-0.39
50% Dichotomous and 50% Polytomous Items	-0.44
33% Dichotomous and 67% Polytomous Items	-0.40

Testing for Measurement Invariance Hypotheses

Three MGCFA nested models were used for the testing of the strong and full measurement invariance hypotheses. Model 1 served as a baseline model where no parameters were constrained between groups. The baseline model was properly specified and hence model misspecification was not a condition in the study. The first chi-square value was obtained from the baseline model for comparison with more constrained models. In Model 2 (i.e., strong measurement invariance model), the number of factors and factor loadings were constrained to be equal across groups. The number of factors, factor loadings, and error variances were constrained to equality across groups in Model 3 (i.e., full measurement invariance model). The tenability of an invariance hypothesis is determined by the statistical significance of the chi-square difference test between two nested models. A non-significant chi-square difference test statistic (e.g., baseline model versus full measurement invariance model) indicates that the full measurement invariance hypothesis is tenable.

It should be noted that, with an eye toward reflecting what goes on in research practice, we did not test for the equality of intercepts -- and hence we did not use a mean and covariance structure (MACS) model (Wu, Li, & Zumbo, 2007). That is, even though there has been periodic advocacy for testing for equality of intercepts it has been largely neglected in applied measurement practice. A thorough review of empirical tests of measurement invariance in applied psychology by Vandenberg and Lance (2000) revealed that although 99% of the studies that they had reviewed investigated loading invariance, only 12% investigated intercept equality and 49% investigated residual variance equality. Therefore by not using the MACS model and not testing intercepts we are not advocating that one ignore intercept equality but rather we are aiming to reflect common research practice. In short, we want our empirical Type I error rates from our simulation study to reflect those error rates in the research literature and in practice.

Estimation Method

The MGCFA was conducted by using the Pearson product moment covariance matrices along with the normal theory ML estimation method in the LISREL 8.53.

Dependent Variables

For each combination of the conditions, MGCFA was conducted for testing the two hypotheses of measurement invariance. Effects of mixed item formats and sample size combinations on the tests of hypotheses of measurement invariance were analyzed through the mean rejection rates of the true models (Type I error rates).

Results

A quality check on the simulated data was conducted by testing the full and strong measurement invariance hypotheses at the population level for each mixed item format combination. As can be seen in Table 2, the differences in chi-squares between models, that is, baseline vs. full invariance, and baseline vs. strong invariance are not statistically significant at the alpha level of .05. The results indicate that the factor structure of the artificial achievement test is invariant across groups. Thus, any sample data drawn from the population data are expected to yield equivalent factor structures for the two groups in the MGCFA framework.

The results in Table 3 show that the empirical rejection rates of the ML chi-square difference test have the nominal alpha (.05) that fall within their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) for the full and strong measurement invariance hypotheses across mixed item formats and sample size combinations. This indicates that mixed item formats and sample size combinations do not affect the empirical Type I error rates of the ML chi-square difference tests in the hypotheses testing of full and strong measurement invariance. Keep in mind that the item response distributions across groups are not very skewed.

Conclusion

The findings of the current study suggest that the practice of using multi-group confirmatory maximum likelihood factor analysis of a Pearson covariance matrix to test measurement invariance hypotheses with mixed item format data does not lead to inflated chi-square difference test statistics. These findings are certainly welcome news for someone reading and reviewing the extant research literature and research reports. However, although these are positive findings, we encourage researchers to use methods that treat the data as ordinal (e.g., polychoric matrices or perhaps full-information methods) and to test for the equality of intercepts. Our results lead us to conclude that although common practice is, in a sense, sub-optimal it at least is not leading to a tendency to over-claim differences in measurement scales across groups – i.e., an inflated Type I error rate.

[The reference list can be found after the subsequent tables.]

Table 2: Maximum Likelihood Chi-square Goodness-of-Fit Statistics between Models

Mixed Item Format	*Model*	*Chi-square Difference Statistic*	*P*
67% Dichotomous Items 33% Polytomous Items (20:10)	Baseline vs. Full Invariance	$\Delta\chi^2 = 32$, $\Delta df = 60$	1.00
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 21$, $\Delta df = 30$	.89
50% Dichotomous Items 50% Polytomous Items (15:15)	Baseline vs. Full Invariance	$\Delta\chi^2 = 38$, $\Delta df = 60$	.99
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 23$, $\Delta df = 30$	.82
33% Dichotomous Items 67% Polytomous Items (10:20)	Baseline vs. Full Invariance	$\Delta\chi^2 = 39$, $\Delta df = 60$	.98
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 23$, $\Delta df = 30$	.82

Note: Numbers of dichomotous and polytomous items are in parentheses.

Table 3: Empirical Type I Error Rates of ML Chi-square Difference Test for the Full and Strong Measurement Invariance Hypotheses Across Mixed Item Formats and Sample Size Combinations

Sample Sizes (*n1: n2*)	*Hypothesis*	*Mixed Item Formats*		
		67% Dichotomous 33% Polytomous	*50% Dichotomous 50% Polytomous*	*33% Dichotomous 67% Polytomous*
200 : 200	FI	.01	.02	.01
	SI	.00	.00	.00
500 : 500	FI	.00	.01	.00
	SI	.02	.01	.02
800 : 800	FI	.00	.01	.00
	SI	.01	.01	.00
200 : 500	FI	.00	.03	.00
	SI	.02	.00	.01
200 : 800	FI	.00	.03	.00
	SI	.00	.02	.00
500 : 800	FI	.00	.02	.02
	SI	.01	.01	.01

Note: Those empirical Type I error rates that have the nominal alpha (.05) outside of their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) would be in **bold font**. FI and SI denote Full and Strong Measurement Invariance Hypotheses, respectively.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*, 662-680.

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2002). *An evaluation of the multiple-group mean and covariance structure analysis model for detecting differential item functioning in graded response items*. Paper presented at the International Test Commission (ITC) Conference on Computer-Based Testing and the Internet. Winchester, UK.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analyses using ordered categorical data. *Structural Equation Modeling, 5,* 344-364.

Jöreskog, K. & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis for non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

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., & Ferguson, Jr. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, Vol. XXVIII*, 491-497.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Wu. A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1-26.