Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 275-282

1538 – 9472/05/$95.00

Manifestation Of Differences In Item-Level Characteristics In Scale-Level Measurement Invariance Tests Of Multi-Group Confirmatory Factor Analyses

Bruno D. Zumbo
University of British Columbia, Canada

Kim H Koh
Nanyang Technological University, Singapore

If a researcher applies the conventional tests of scale-level measurement invariance through multi-group confirmatory factor analysis of a PC matrix and MLE to test hypotheses of strong and full measurement invariance when the researcher has a rating scale response format wherein the item characteristics are different for the two groups of respondents, do these scale-level analyses reflect (or ignore) differences in item threshold characteristics? Results of the current study demonstrate the inadequacy of judging the suitability of a measurement instrument across groups by only investigating the factor structure of the measure for the different groups with a PC matrix and MLE. Evidence is provided that item level bias can still be present when a CFA of the two different groups reveals an equivalent factorial structure of rating scale items using a PC matrix and MLE.

Key words: multi-group confirmatory factor analysis, item response formats

Introduction

Broadly speaking, there are two general classes of statistical and psychometric techniques to examine measurement invariance across groups: (1) scale-level analyses, and (2) item-level analyses. The groups investigated for measurement invariance are typically formed by gender, ethnicity, or translated/adapted versions of a test. In scale-level analyses, the set of items comprising a test are often examined together using multi-group confirmatory factor analyses (Byrne, 1998; Jöreskog, 1971) that involve testing strong and full measurement invariance hypotheses. In the item-level analyses the focus is on the invariant characteristics of each item, one item at a time.

In setting the stage for this study, which involves a blending of ideas from scale- and item-level analyses (i.e., multi-group confirmatory factor analysis and item response theory), it is useful to compare and contrast overall frameworks for scale-level and item-level approaches to measurement invariance. Recent examples of this sort of comparison can be found in Raju, Laffitte, & Byrne (2002), Reise, Widaman, & Pugh (1993), and Zumbo (2003). In these studies, the impact of scaling on measurement invariance has not been examined. Hence, it is important for the current study to investigate to what extent the number of scale points effects the tests of measurement invariance hypotheses in multi-group confirmatory factor analysis.

Scale-level Analyses

There are several expositions and reviews of single-group and multi-group confirmatory factor analysis (e.g., Byrne, 1998; Steenkamp & Baumgartner, 1998; Vandenberg

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 at the University of British Columbia, Canada Email: bruno.zumbo@ubc.ca. Kim H. Koh is Assistant Professor, Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, Singapore. Email: khkoh@nie.edu.sg. An earlier version of this article was presented at the 2003 National Council on Measurement in Education (NCME) conference, Chicago Illinois. We would like to thank Professor Greg Hancock for his comments on an earlier draft of this article.

& Lance, 2000); therefore this review will be very brief. In describing multi-group confirmatory factor analysis, consider a one-factor model: one latent variable and ten items all loading on that one latent variable. There are two sets of parameters of interest in this model: (1) the factor loadings corresponding to the paths from the latent variable to each of the items, and (2) the error variances, one for each of the items. The purpose of the multi-group confirmatory factor analysis is to investigate to what extent each, or both; of the two sets of model parameters (factor loadings and error variances) are invariant in the two groups.

As Byrne (1998) noted, there are various hypotheses of measurement invariance that can be tested, from weak to strict invariance. That is, one can test whether the model in its entirety is completely invariant, i.e., the measurement model as specified in one group is completely reproduced in the other, including the magnitude of the loadings and error variances. At the other end of the extreme is an invariance in which the only thing shared between the groups is overall pattern of the model but neither the magnitudes of the loadings nor of the error variances are the same for the two groups, i.e., the test has the same dimensionality, or configuration, but not the same magnitudes for the parameters.

Item-level Analyses

In item-level analyses, the framework is different than at the scale-level. At the item level, measurement specialists typically consider (a) one item at a time, and (b) a unidimensional statistical model that incorporates one or more thresholds for an item response. That is, the response to an item is governed by referring the latent variable score to the threshold(s) and from this comparison the item response is determined.

Consider the following example of a four-point Likert item, "How much do you like learning about mathematics?" The item responses are scored on a 4-point scale such as (1) Dislike a lot, (2) Dislike, (3) Like, and (4) Like a lot. This item, along with other items, serve as a set of observed ordinal variables, *x*'s, to measure the latent continuous variable *x**, namely attitudes toward learning mathematics. For each observed ordinal variable *x*, there is an underlying continuous variable *x**. If *x* has *m* ordered categories, *x* is connected to *x** through the non-linear step function: $x = i$ if

$$\tau_{i-1} < x^* \leq \tau_{i,} \quad i = 1,2,3,...,m,$$

where

$$\tau_0 = -\infty, \tau_1 < \tau_2 < \tau_3 < ...\tau_{m-1},$$

and $\tau_m = +\infty$

are parameters called threshold values. For a variable *x* with *m* categories, there are *m-1* unknown thresholds. Given that the above item has four response categories, there are three thresholds with the latent continuous variable. If one approaches the item level analyses from a scale-level perspective, the item responding process is akin to the thresholds one invokes in computing a polychoric correlation matrix (Jöreskog & Sörbom, 1996).

In an item-level analysis measurement specialists often focus on differences in thresholds across the groups. That is, the focus is on determining if the thresholds are the same for the two groups. If studying an achievement or knowledge test, it should be asked if the items are equally difficult for the two groups, with the thresholds being used as measures of item difficulty (i.e., an item with a higher threshold is more difficult). These differences in thresholds are investigated by methods collectively called "methods for detecting differential item functioning (DIF)". In common measurement practice this sort of measurement invariance is examined, for each item, one item at a time, using a DIF detection method such as the Mantel-Haenszel (MH) test or logistic regression (conditioning on the observed scores), or methods based on item response theory (IRT).

The IRT methods investigate the thresholds directly whereas the non-IRT methods test the difference in thresholds indirectly by studying the observed response option proportions by using categorical data analysis methods such as the MH or logistic regression methods (see Zumbo & Hubley, 2003 for a review).

Although both item- and scale-level methods are becoming popular in educational and psychosocial measurement, many researchers are still recommending and using only scale-level methods such as multi-group confirmatory factor analysis (for example, see, Byrne, 1998; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). There are, of course, scale-level methods that allow one to incorporate and test for item threshold differences in multi-group confirmatory factor analysis; however, these methods are not yet widely used. Instead, the popular texts on structural equation modeling by Byrne as well as the widely cited articles by Steenkaump and Baumgartner, and Vandenberg and Lance focus on and instruct users of structural equation modeling on the use of Pearson covariance matrices and the Chi-squared tests for model comparison based on maximum likelihood estimation (For an example see Byrne, 1998, Chapter 8 on a description of multi-group methods and p. 239 of her text for a recommendation on using ML estimation with the type of data we are describing above).

The question that this article addresses is reflected in the title: Do Differences in Item-Level Characteristics Manifest Themselves in Scale-Level Measurement Invariance Tests of Multi-Group Confirmatory Factor Analyses? That is, if a researcher applies the conventional tests of scale-level measurement invariance through multi-group confirmatory factor analysis of a Pearson covariance matrix and maximum likelihood estimation to test hypotheses of strong and full measurement invariance when the researcher has the ordinal (often called Likert) response format described above, do these scale-level analyses reflect (or ignore) differences in item threshold characteristics? If one were a measurement specialist focusing on item-level analyses (e.g., an IRT specialist), another way of asking this question is: Does DIF, or other forms of lack of item parameter invariance such as item drift, manifest itself in construct comparability across groups?

The present study is an extension of Zumbo (2003). A limitation of his earlier work is that it focused on the population analogue and did not investigate, as in this, the pattern and characteristics of the statistical decisions over the long run; i.e., over many replications. We study the rejection rates for a test of the statistical hypotheses in multi-group confirmatory factor analysis.

Methodology

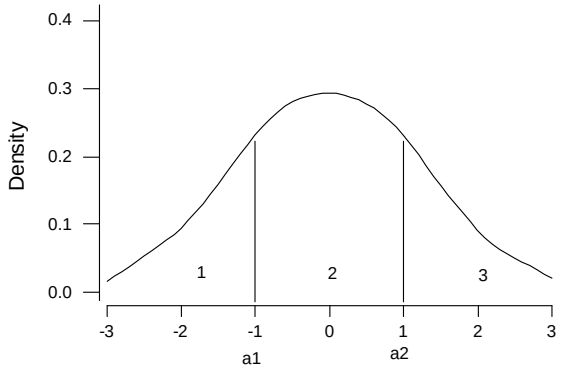
A computer simulation was conducted to investigate whether item-level differences in thresholds manifest themselves in the tests of strong and full measurement invariance hypotheses in multi-group CFA of a Pearson covariance matrix with maximum likelihood estimation.

Simulated was a one-factor model with 38 items. Obtained was a population covariance matrix based on the data reported in Zumbo (2000, 2003) that were based on the item characteristics of a sub-section of the TOEFL. Based on this covariance matrix, 100,000 simulees were generated on these 38 items with a multivariate normal distribution with marginal (univariate) means of zero and standard deviations of one. The simulation was restricted to a one-factor model because item-level methods (wherein differences in item thresholds, called DIF in that literature, is widely discussed) predominantly assume unidimensionality of their items, for example, IRT, MH, or logistic regression DIF methods.

The same item thresholds were used as those used by Bollen and Barb (1981) in their study of ordinal variables and Pearson correlation. In short, this method partitions the continuum ranging from –3 to +3. The thresholds are those values that divide the continuum into equal parts. The example in Figure 1 is a three-point scale using the notation described above for the x^* and x. Item thresholds were applied to these 38 normally distributed item vectors to obtain the ordinal item responses.

The simulation design involved two completed crossed factors: (i) number of scale points ranging from three to seven, and (ii) the percentage of items with different thresholds (i.e., percentage of DIF items) ranging from zero to 42.1 (1, 4, 8 and 16 items out of the total of 38).

Figure 1. A Three Category, Two Threshold x and its corresponding x*.



Note: Number of categories for x: 3 (values 1, 2, 3). Item thresholds for x*: a1, a2 (values of –1 and 1).

Three to seven item scale points were chosen because in order to only deal with those scale points for which Byrne (1998) and others suggest the use of Pearson covariance matrices with maximum likelihood estimation for ordinal item data. The resulting simulation design is a five by five completely crossed design.

The differences in thresholds were modeled based on suggestions from the item response theory (IRT) DIF literature for binary items. That is, the IRT DIF literature (e.g., Zumbo, 2003; Zwick & Ercikan, 1989) suggests that an item threshold difference of 0.50 standard deviations is a moderate DIF. This idea was extended and applied to each of the thresholds for the DIF item(s). For example, for a three-point item response scale group one would have thresholds of -1.0 and 1.0 whereas group two would have thresholds of –0.5 and 1.5. Note that for both groups the latent variables are simulated with a mean of zero and standard deviation of one. The same principle applies for the four to seven point scales.

Given that both groups have the same latent variable mean and standard deviation, the difference thresholds for the two groups (i.e., the DIF) would imply that the item(s) that is (are) performing differently across the two groups would have different item response distributions. It should be noted that the Bollen and Barb methodology results in symmetric Likert item responses that are normally distributed. The results in Table 1 allow one to compare the effect of having different thresholds in terms of the skewness and kurtosis.

The descriptive statistics reported in Table 1 were computed from a simulated sample of 100,000 continuous normal scores that were transformed with our methodology. For a continuous normal distribution the skewness and kurtosis statistics reported would both be zero. Focusing first on the skewness, it can be see in Table 1 that they range from -0.008 to 0.011 (with a common standard error of 0.008) indicating that, as expected, the Likert responses were originally near symmetrical. Applying the

Table 1. Descriptive Statistics of the Items without and without Different Thresholds.

# of Scale Points	Skewness		Kurtosis	
	Original	Different Thresholds	Original	Different Thresholds
3	-0.001	-0.004	0.144	-0.364
4	-0.008	0.125	-0.268	-0.294
5	0.011	0.105	-0.211	-0.277
6	-0.005	0.084	-0.185	-0.261
7	-0.003	0.082	-0.169	-0.238

Note: These statistics were computed from a sample of 100,000 responses using SPSS 11.5. In all cases, standard errors of the skewness and kurtosis were 0.008 and 0.015, respectively.

threshold difference, as described above, resulted in item responses that were nearly symmetrical for three, six, and seven scale points, and only small positive skew (0.125 and 0.105) for the four and five scale points. In terms of kurtosis, there is very little change with the different thresholds, except for the three-point scale that resulted in the response distribution being more platykurtic with the different thresholds.

The items on which the differences in thresholds were modeled were selected randomly. Thus in the four item condition, the item from the one-item condition was included and an additional three items were randomly selected. In the eight-item condition, the four items were included an additional four items were randomly selected, and so on.

The sample size for the multi-group CFA was three hundred per group, a sample size that is commonly see in practice. The number of replications for each cell in the simulation design was 100. The nominal alpha was set at .05 for each invariance hypothesis test. It is important to note that the rejection rates reported in this paper are, technically, Type I error rates only for the "no DIF" conditions. In the other cases, when DIF is present, the rejection rates represent the likelihood of rejecting the null hypothesis (for each of the full and strong measurement invariance hypotheses) when the null is true at the unobserved latent variable level, but not necessarily true in the manifest variables because the thresholds are different across the groups.

For each replication the strong and full measurement invariance hypotheses were tested. These hypotheses were tested by comparing the baseline model (with no between group constraints) to each of the strong and full measurement invariance models. That is, strong measurement invariance is the equality of item loadings – Lambda X, and the full measurement invariance is the equality of both item loadings and uniquenesses, Lambda X and Theta-Delta, across groups. For each cell, we searched the LISREL output for the 100 replications for warning or error messages.

A one-tailed 95% confidence interval was computed for each empirical error rate. The confidence interval is particularly useful in this context because we have only 100 replications so we want to take into account sampling variability of the empirical error rate. The upper confidence bound was compared to Bradley's (1978) criterion of liberal robustness of error. If the upper confidence interval was .075 or less it met the liberal criterion.

Table 2. Rejection Rates for the Full and Strong Measurement Invariance Hypotheses, with and without DIF Present.

Percentage of items having different thresholds across the two groups (% of DIF items)	**Number of scale points for the item response format**				
	3 pt.	**4pt.**	**5pt.**	**6pt.**	**7pt.**
0 (no DIF items)	FI .07 (.074) SI .03 (.033)	FI .01 (.012) SI .03 (.033)	FI .01 (.012) SI .04 (.043)	FI .05 (.054) SI .03 (.033)	FI .02 (.022) SI .06 (.064)
2.9 (1 item)	**FI .09 (.095) ⇑** SI .07 (.074)	FI .02 (.022) SI .02 (.022)	FI .01 (.012) SI .01 (.012)	FI .00 (.000) SI .03 (.033)	FI .02 (.022) SI .03 (.033)
10.5 (4 items)	FI .04 (.043) SI .06 (.064)	FI .03 (.033) SI .02 (.022)	FI .03 (.033) SI .04 (.043)	FI .03 (.033) SI .06 (.064)	FI .03 (.033) SI .07 (.074)
21.1 (8 items)	**FI .08 (.084) ⇑** SI .04 (.043)	FI .00 (.000) SI .00 (.000)	FI .04 (.043) SI .04 (.043)	FI .02 (.022) SI .01 (.012)	FI .02 (.022) SI .07 (.074)
42.1 (16 items)	FI .07 (.074) SI .04 (.043)	FI .02 (.022) SI .02 (.022)	FI .02 (.022) SI .06 (.064)	FI .02 (.022) SI .05 (.054)	FI .02 (.022) SI .02 (.022)

Note. The upper confidence bound is provided in parentheses next to the empirical error rate. The empirical error rates in the range of Bradley's liberal criterion are indicated in plain text type whereas empirical error rates that do not even satisfy the liberal criterion are identified with symbol ⇑ and in **bold font**.

Results

To determine whether the tests of strong and full measurement invariance (using the Chi-squared difference tests arising from using a Pearson Covariance matrix and maximum likelihood estimation in, for example, LISREL) are affected by differences in item thresholds we examined the level of error rates in each of the conditions of the simulation design. Table 2 lists the results of the simulation study. Each tabled valued is the empirical error rate over the 100 replications with 300 respondents per group (upon searching the output for errors and warnings produced by LISREL, one case was found of a non-positive definite theta-delta (TD) matrix for the study cells involving three scale points for the 2.9 and 21.1 percent of DIF items. The one replication with this warning was excluded from the calculation of the error rate and upper 95% bound for those two cells, therefore the cell statistics were calculated for 99 replications for those two cases). The values in the range of Bradley's liberal criterion are indicated in plain text type. Values that do not even satisfy the liberal criterion are identified with symbol ⇑.

The results show that almost all of the empirical error rates are within the range of Bradley's liberal criterion. Only two cells have empirical error rates that exceed the upper confidence interval of .075. These two cells are for the three-scale-point condition. This suggests that the differences of item thresholds may have an impact on the full measurement invariance hypotheses in some conditions for measures with a three-point item response format, although this finding is seen in only two of the four conditions involving differences in thresholds. For scale points ranging from four to seven, the empirical error rates are either at or near the nominal error. Interestingly, the empirical error rates of the three scale points are

slightly inflated when a measure has 10.5 and 21.1 percent (moderate amount) of DIF items.

Conclusion

The conclusion from this study is that when one is comparing groups' responses to items that have a rating scale format in a multi-group confirmatory factor analysis of measurement invariance by using maximum likelihood estimation and a Pearson correlation matrix, one should ensure measurement equivalence by investigating item-level differences in thresholds. In addition, giving consideration only to the results of scale-level methods as evidence may be misleading because item-level differences may not manifest themselves in scale-level analyses of this sort.

Of course, the conclusions of this study apply to any situation in which one is (a) using rating scale (sometimes called Likert) items, and comparing two or more groups of respondents in terms of their measurement equivalence, however, it also provides further empirical support for the recommendation found in the International Test Commission Guidelines for Adapting Educational and Psychological Tests that researchers carry out empirical studies to demonstrate factorial equivalence of their test across groups *and* to identify any item-level DIF that may be present (see Hambleton & Patsula, 1999; van de Vijver & Hambleton, 1996) and is an extension of previous studies by Zumbo (2000; 2003) comparing and item- and scale-level methods.

Overall, the results demonstrate the inadequacy of judging the suitability of a measurement instrument across groups by only investigating the factor structure of the measure for the different groups with a Pearson covariance matrix and maximum likelihood estimation. It has been common to assume that if the factor structure of a test remains the same in a second group, then the measure functions the same and measurement equivalence is achieved. Evidence is provided that item level bias can still be present when a CFA of the two different groups reveals an equivalent factorial structure of rating scale items using a Pearson covariance matrix and maximum likelihood estimation. Since it is the scores from a test or instrument that are ultimately used to achieve the intended purpose, the scores may be contaminated by item level bias and, ultimately, valid inferences from the test scores become problematic.

References

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*, 289-311.

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS*. Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adaptive tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*, 1-11.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G., & Sorbom, D. (1996). *LISREL 8: User's Reference Guide*. Chicago, IL.: Scientific Software International.

Luecht, R. (1996). MIRTGEN 1.0 [Computer Software].

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-529.

Reise, S. P., Widaman, K. F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

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

Zumbo, B. D. (2000, April). *The effect of DIF and impact on classical test statistics: undetected DIF and impact, and the reliability and interpretability of scores from a language proficiency test*. Paper presented at the *National Council on Measurement in Education (NCME)*, New Orleans, LA.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses?: Implications for translating language tests. *Language Testing, 20*, 136-147.

Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In Rocío Fernández-Ballesteros (Ed.). *Encyclopedia of Psychological Assessment* (p. 505-509). Sage Press, Thousand Oaks, CA.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.