Samuel Salzborn • Eldad Davidov
Jost Reinecke (Eds.)

Methods, Theories, and Empirical Applications in the Social Sciences

Festschrift for Peter Schmidt

Springer VS
2012

Multigroup and Multilevel Approaches to Measurement Equivalence

Joop J. Hox, Edith D. de Leeuw, Matthieu J.S. Brinkhuis, Jeroen Ooms

Comparative surveys have a number of characteristic analysis issues in common. When measurement instruments are used in different cultures or are translated into different languages, the first analysis questions concern measurement equivalence. May we assume that these instruments measure the same constructs? How can we assess whether we have measurement equivalence?

The classic approach to deal with these questions is structural equation modeling (SEM) using a multigroup analysis. However, when the number of groups (e.g., countries) becomes large, multigroup SEM becomes unwieldy. Multigroup SEM estimates a unique set of parameter values for each country, which results in a complex model. A random effects model, such as multilevel modeling (MLM), will treat the countries as a sample from a larger population. Instead of estimating different parameter values for each country, it assumes a distribution of parameter values and estimates its mean and variance. This makes MLM more parsimonious than SEM when a large number of countries is studied. At present, the larger comparative surveys involve enough countries to consider multilevel analysis (Hox, de Leeuw, & Brinkhuis, 2010; Hox, Maas, & Brinkhuis, 2010).

1 Comparing SEM and MLM

For many years, multigroup confirmatory factor analysis has been the analysis method of choice for analyzing data in international surveys (Jöreskog, 1971; for an overview see Davidov, Schmidt, & Billiet, 2011). If all factor loadings are invariant across all countries, we have a strong form of measurement equivalence (Vandenburg & Lance, 2000). Although the ideal is achieving complete measurement invariance, in practice a small amount of variation is accepted, which leads to partial measurement invariance (Byrne, Shavelson, & Muthén, 1989).

Multilevel models have been developed for the statistical analysis of data that have a hierarchical or clustered structure. As comparative surveys lead to clustered data with respondents clustered within countries or cultures, multilevel analysis of measurement equivalence is a promising approach. The most flexible method to date is multilevel structural equation modeling (MSEM, cf. Mehta & Neale, 2005). Including random slopes in the measurement model provides a

new approach to testing measurement equivalence. Equivalent measurement means that the same factor model fits in all groups with no factor loading having a coefficient that varies across groups. Thus, measurement equivalence can be established by testing if factor loadings have significant variation across groups. In this chapter, this approach is compared to the traditional SEM multigroup analysis in a simulation study.

2 Comparing SEM and MLM in a Simulation

The data were simulated, mimicking the structure of comparative studies, with a relatively small number of countries and a large sample of respondents within a country. This is a realistic setting in many international surveys. We also require the latent variable to be over-identified, which leads to four observed indicators for a single construct. There are two simulated conditions. In one condition, designated H_0, measurement equivalence holds. The goal of analyzing this condition is to investigate if the number of available countries permits accurate parameter estimates and standard errors and correct decisions about the equivalence of measurement. In the second condition, designated H_A, measurement equivalence does not hold. The goal of analyzing this condition is to investigate which chosen method of analysis leads to correct decisions about the equivalence of measurement. In essence, the first (H_0) condition investigates accuracy, and the second (H_A) condition investigates statistical power.

To represent the number of countries usually found in large-scale international studies, three different values have been chosen for the Number of Countries (NC = 20, NC = 30, and NC = 40). Within each country, 1,500 respondents are simulated. The H_0, under simulation, is presented in Figure X.1. It should be noted that means are fixed at 0 and that all simulations are performed 1,000 times in each condition.

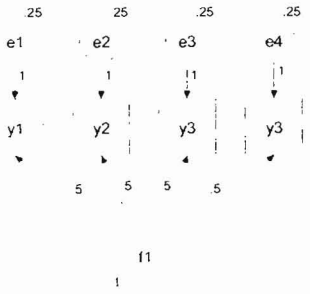

Figure 1: Path diagram for a factor model.

For the alternative hypothesis H_A, a model is simulated where the fourth regression weight is different from the others for half of the countries, namely, 0.3 instead of 0.5.

3 Simulation Results

Table 1 shows the results for the multigroup SEM analyses of the simulated data.

	H_0			H_A		
	n = 20	n = 30	n = 40	n = 20	n = 30	n = 40
Chi^2 $p < .05$	5.5%	5.9%	6,6%	100.0%	100.0%	100.0%
CFI > .90	100.0%	100.0%	100.0%	100.0%	100,0%	100,0%
TLI > .90	100.0%	100.0%	100.0%	100.0%	100,0%	100,0%
RMSEA < .05	100.0%	100.0%	100.0%	36.2%	38.1%	40.7%
L1 in 95% CI	95.2%	94.2%	94.9%	94.5%	94.2%	94.4%
L2 in 95% CI	94.2%	95.5%	95.7%	95.3%	96.1%	95.8%
L3 in 95% CI	96.1%	95.1%	95.7%	94.9%	94.5%	94.8%
L4 in 95% CI	93.5%	95.4%	94.4%	0.0%	0.0%	0.0%
Mean *p* value	0.498	0.486	0.481	<.00	<.00	<.00
Mean CFI	1.000	1.000	1.000	0.976	0.976	0.976
Mean TLI	1.000	1.000	1.000	0.985	0.985	0.985
Mean RMSEA	0.003	0.003	0.003	0.51	0.051	0.051
Mean L1	0.500	0.500	0.500	0.500	0.500	0.500
Mean L2	0.500	0.500	0.500	0.500	0.500	0.500
Mean L3	0.500	0.500	0.500	0.499	0.499	0.500
Mean L4	0.500	0.500	0.500	0.394	0.394	0.394

Table 1: Results for multigroup SEM H_0 and H_A data.

A remarkable result is that only the chi-square fit measure is able to detect model violations; the fit measures CFI, TLI, and RMSEA lack power to detect the violations introduced in the H_A data. Only the RMSEA provides some indication of a nonperfect fit in the majority of the simulations.

Table 2 shows the results of the multilevel analysis of the simulated data.

	H_0			H_A		
	n = 20	n = 30	n = 40	n = 20	n = 30	n = 40
chi² p > .05	0.071	0.051	0.064	0.053	0.061	0.05
CFI > .90	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
TLI > .90	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
RMSEA < .05	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
L1 in 95% CI	93.2%	93.6%	94.4%	91.4%	93.5%	93.6%
L2 in 95% CI	91.7%	94.2%	93.4%	93.8%	94.3%	95.2%
L3 in 95% CI	93.2%	93.7%	94.9%	93.4%	93.2%	93.5%
L4 in 95% CI	92.6%	92.8%	93.8%	0%	0%	0%
Mean *p* value	0.485	0.495	0.492	0.492	0.492	0.490
Mean CFI	1.000	1.000	1.000	1.000	1.000	1.000
Mean TLI	1.000	1.000	1.000	1.000	1.000	1.000
Mean RMSEA	0.002	0.002	0.002	0.002	0.002	0.001
Mean L1	0.500	0.500	0.500	0.500	0.500	0.500
Mean L2	0.500	0.500	0.500	0.500	0.500	0.500
Mean L3	0.500	0.500	0.500	0.500	0.500	0.500
Mean L4	0.500	0.500	0.500	0.400	0.400	0.400

Table 2: Results for multilevel SEM H_0 and H_A data.

The results of the multilevel analysis are very similar to those of the multigroup SEM analysis for the H_0 model. However, for the H_A model there is now no power for the global model fit test. As in the SEM model, the fit measures CFI, TLI, and RMSEA lack power to detect the violations introduced in the H_A.

4 Discussion

The most striking result is the lack of power if the model is incorrectly specified. Only in the classical multigroup analysis does the global chi-square test routinely reject the model when the H_A data are analyzed. But even in that case, the fit

indices would indicate a very good fit, and most analysts would probably argue that given the large total sample size, the chi-square test is overly powerful and the model rejection can therefore be ignored. Doing this, they would be ignoring a clear violation of measurement invariance.

The multilevel analysis almost never leads to a rejection of the model, even with rather large sample sizes, and violation of measurement invariance will not be detected. To explore the reason for this, some extra simulations were run based on a model that included extremely large violations of measurement equivalence. The model was again almost never rejected. Inspection of the parameter estimates makes clear what happens. Under H_A there are varying loadings for variable Y_4; when a model is estimated that specifies this loading as fixed, this leads to an inflated variance for Y_4, which is absorbed in the residual measurement error for Y_4. Thus, varying slopes can be detected only by making strong assumptions about the residual measurement errors. Multilevel SEM does not routinely incorporate these assumptions (as restrictions), and while multilevel regression does implicitly incorporate such assumptions (see Raudenbush, Rowan, & Kang, 1991, for details), it cannot test these.

The conclusion is that reliance on global fit indices is misleading when measurement equivalence is tested. It is advised to examine more specific indicators of lack of fit, such as modification indices and the corresponding estimated parameter change. In contrast to the chi-square test and associated fit indices, the modification indices are related to a specific parameter constraint. Therefore, when there is a specific fit problem in a model that generally fits well, the modification index has a better power to indicate the source of this problem. Furthermore, the estimated parameter change indicates how different the unconstrained parameter estimate is from the constrained estimate. Multigroup SEM is still the method of choice for testing measurement equivalence.

5 References

Byrne, Barbara M., Rich J. Shavelson, and Bengt O. Muthén. 1989. Testing for the Equivalence of Factor and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin* 105: 456-466.

Davidov. Eldad, Peter Schmidt, and Jaak Billiet, eds. 2011. *Cross-cultural analysis: Methods and applications*. New York: Psychology Press.

Hox, Joop J., Edith D. de Leeuw, and Matthieu J.S. Brinkhuis. 2010. Analysis Models for Comparative Surveys. In *Survey Methods in Multicultural, Multinational, and Multiregional Contexts*, eds. Janet Harkness, Michael Braun, Brian Edwards, Timothy Johnson, Lars Lyberg, Peter Mohler, Beth Ellen Pennell, and Tom W. Smith, 395-418 Hoboken, NJ: Wiley.

Hox, Joop J., Cora J.M. Maas, and Matthieu J.S. Brinkhuis. 2010. The Effect of Estimation Method and Sample Size in Multilevel SEM. *Statistica Neerlandica* 64: 157-170.

Jöreskog, Karl G. 1971. Simultaneous Factor Analysis in Several Populations. *Psychometrika* 36: 409-426.

Mehta, Paras D. and Michael C. Neale. 2005. People are Variables Too: Multilevel Structural Equations Modeling. *Psychological Methods* 10: 259-284.

Raudenbush, Steven W., Brian Rowan, and Sang Jin Kang. 1991. A Multilevel, Multivariate Model for Studying School Climate with Estimation via the EM Algorithm and Application to U.S. High-School Data. *Journal of Educational Statistics* 16(4): 295-330.

Vandenburg, Robert J. and Charles E. Lance. 2000. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods* 3(1): 4-70.