EUROPEAN JOURNAL OF DEVELOPMENTAL PSYCHOLOGY
2012, 9 (4), 486–492

Psychology Press
Taylor & Francis Group

Developmetrics

A checklist for testing measurement invariance

Rens van de Schoot[1,2], Peter Lugtig[1], and Joop Hox[1]

[1]Department of Methods and Statistics, Utrecht University, Utrecht, The Netherlands
[2]Optentia Research Program, North-West University, Vanderbijlpark, South Africa

The analysis of measurement invariance of latent constructs is important in research across groups, or across time. By establishing whether factor loadings, intercepts and residual variances are equivalent in a factor model that measures a latent concept, we can assure that comparisons that are made on the latent variable are valid across groups or time. Establishing measurement invariance involves running a set of increasingly constrained structural equation models, and testing whether differences between these models are significant. This paper provides a step-by-step guide to analysing measurement invariance.

Keywords: Confirmatory factor analysis; Validity; Measurement invariance.

In the social and behavioural sciences self-report questionnaires are often used to assess different aspects of human behaviour. These questionnaires consist of items that are developed to assess an underlying phenomenon with the goal to follow individuals over time or to compare groups. To be valid for such a comparison a questionnaire should measure identical constructs with the same structure across different groups. When this is the case, the questionnaire is called measurement invariant. If measurement invariance (MI) can be demonstrated then the participants across all groups

Correspondence should be addressed to Rens van de Schoot, Department of Methodology and Statistics, Utrecht University, PO Box 80.140, NL-3508 TC, Utrecht, The Netherlands. E-mail: a.g.j.vandeschoot@uu.nl

The first author received a grant from the Netherlands Organization for Scientific Research: NWO-VENI-451-11-008.

With many thanks to Marie Stievenart, Stefanos Mastrotheodoros, Leonard Vanbrabant and Esmee Verhulp for proofreading the manuscript.

http://www.psypress.com/edp http://dx.doi.org/10.1080/17405629.2012.686740

interpret the individual questions, as well as the underlying latent factor, in the same way. Having determined MI, future studies can compare the occurrence, determinants, and consequences of the latent factor scores. When MI does not hold, groups or subjects over time respond differently to the items and as a consequence factor means cannot reasonably be compared.

Jöreskog (1971) was the first author to write about the equivalence of factor structures. The concept of MI was introduced by Byrne, Shavelson, and Muthén (1989), after which the testing of MI took off. Recent review articles provided an overview of a multitude of substantive studies that tested MI (e.g., Vandenberg & Lance, 2000). However, a simple step-by-step checklist for testing MI is lacking and that is exactly the goal of the current paper.

SOFTWARE

MI can be tested using any structural equation modelling software program. LISREL (Jöreskog & Sörbom, 1996–2001) has long been the best option. It can handle categorical data, but it requires syntax and knowledge of matrix algebra. Amos (Arbuckle 2007) is very user friendly, but has limited capabilities for handling categorical data. Mplus (Muthén & Muthén, 2010) is currently the most flexible program, but requires knowledge of syntax. Lavaan (Rosseel, in press) and OpenMx (Boker et al., 2011) are both open-source R packages that are still being developed. We provide Mplus syntax at www.fss.uu.nl/mplus for all the analyses described in the current paper.

MODEL FIT AND MODEL COMPARISON

The most commonly used test to check global model fit is the chi-square test (Cochran, 1952), but it is dependent on the sample size: it rejects reasonable models if the sample is large and it fails to reject poor models if the sample is rather small. There are three other types of fit indices that can be used to assess the fit of a model. For details and references see Kline (2010).

First, the comparative indices that compare the fit of the model under consideration with fit of baseline model, for example the Tucker-Lewis Index (TLI) and Comparative Fit Index (CFI). Fit is considered adequate if the CFI and TLI values are $> .90$, and better if they are $> .95$. The TLI attempts to correct for complexity of the model but is somewhat sensitive to a small sample size. Also, it can become > 1.0, which can be interpreted as an indication of over fitting: making the model more complex than needed. If the $\chi^2 < df$, the CFI is set to 1.0, which makes it a normed fit index.

Second, there are absolute indices that examine closeness of fit, for example the Root Mean Square Error of Approximation (RMSEA). The cut-off value is RMSEA $< .08$, better is $< .05$. The RMSEA is insensitive to sample size, but sensitive to model complexity.

Third, there are information theoretic indices, for example the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Both can be used to compare competing models and make a trade-off between model fit (i.e., -2*log likelihood value) and model complexity (i.e., a computation of the number of parameters). A lower AIC/BIC value indicates a better trade-off between fit and complexity. There is no rule of thumb, the values depend on actual dataset and the model, simply chooses the model with the lowest IC value.

THE FACTOR MODEL

Consider Figure 1 which is a 1-item questionnaire, denoted by X. We assume there is an underlying mechanism causing the variance in X, denoted by the latent variable *ksi*.

The regression equation is:

$$X = b_0 + b_1 \times ksi + b_2 \times \text{error} \tag{1}$$

where b_0 is the intercept, b_1 is the regression coefficient (the factor loading in the standardized solution) between the latent variable and the item, and b_2 is the regression coefficient between the residual variance (i.e., error) and the manifest item. For model identification purposes this latter coefficient is fixed to equal 1. Note that if the means of *ksi* and the error are constrained at zero, the intercept of X is estimated. If, on the other hand, the intercept and the error mean are constrained at zero, then the mean of *ksi* is estimated.

As a result, there are two ways of parametrizing the CFA model. This is illustrated in Figure 2 where three items, $X1$–$X3$, are believed to measure the same underlying latent variable *ksi*. First, if the latent factor mean is constrained to equal 0 and the variance equal to 1, then all factor loadings and all intercepts are estimated, see Figure 2A. Second, if one factor loading is constrained to equal 1, and the corresponding intercept equal to zero, then

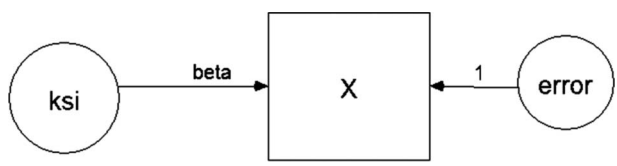

Figure 1. CFA with one item.

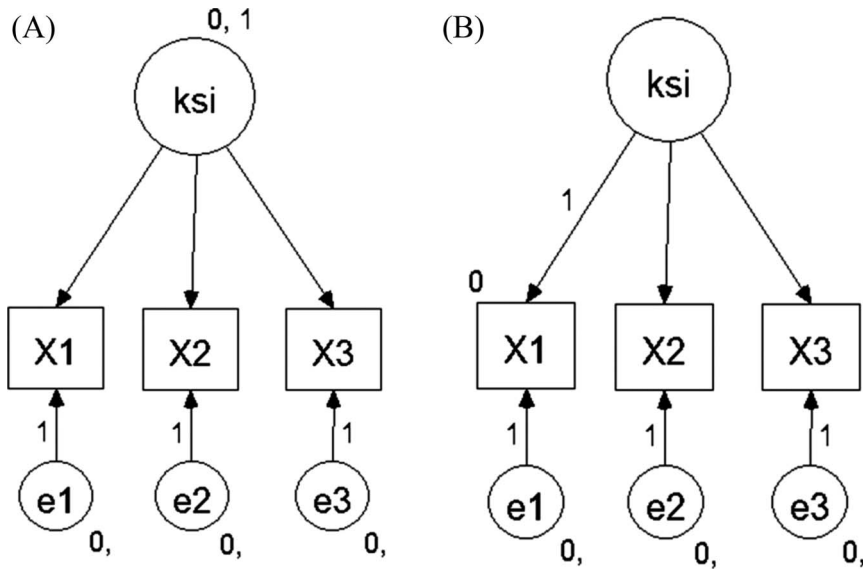

Figure 2. Two methods of parametrizing the CFA model.

the other factor loadings, the other intercepts, and the factor mean plus its variance are estimated. So, depending on what information you want to report either the parametrization in Figure 2A or the parametrization in Figure 2B should be applied. Basically the question boils down to:

- Do you want to compare the factor loadings across groups? Then, choose the parametrization in Figure 2A;
- Do you want to compare the latent means across groups? Then, choose the parametrization in Figure 2B. (Note that the parametrization of Figure 2B is the default in Amos, Lavaan and Mplus.)

Sometimes you have to switch between parametrizations within one paper to answer both questions.

TESTING FOR MEASUREMENT INVARIANCE

In this section we discuss all the steps necessary to evaluate MI. See the supplementary material on www.fss.uu.nl/mplus for Mplus syntax.

Before testing invariance, it is important that the data have been properly screened. For example, if one of the groups contains more (multivariate) outliers than the other group. MI studies rely on fitting the observed

covariance matrix (the data) to a model, so any bias in one of the groups due to outliers will affect factor loadings, intercepts and error variances.

Start with specifying a confirmatory factor analysis (CFA) that reflects how the construct is theoretically operationalized. This CFA model should be fitted for each group separately to test for configural invariance: whether the same CFA is valid in each group. Basically, this boils down to selecting each of the groups separately and running the CFA multiple times, or running a multiple group analysis without any equality constraints

To test for MI a set of models need to be estimated:

1. Run a model where only the factor loadings are equal across groups but the intercepts are allowed to differ between groups. This is called metric invariance and tests whether respondents across groups attribute the same meaning to the latent construct under study.
2. Run a model where only the intercepts are equal across groups, but the factor loadings are allowed to differ between groups. This tests whether the meaning of the levels of the underlying items (intercepts) are equal in both groups.
3. Run a model where the loadings and intercepts are constrained to be equal. This is called scalar invariance and implies that the meaning of the construct (the factor loadings), and the levels of the underlying items (intercepts) are equal in both groups. Consequently, groups can be compared on their scores on the latent variable.
4. Run a model where the residual variances are also fixed to be equal across groups. This is called full uniqueness MI and means that the explained variance for every item is the same across groups. Put more strongly, the latent construct is measured *identically* across groups. If error variances are not equal, groups can still be compared on the latent variable, but this is measured with different amounts of error between groups.

For straightforward interpretation of latent means and correlations across groups, both the factor loadings and intercepts should be the same across groups (scalar invariance). On the other hand, if the fit of Model 3 is significantly worse than Model 1 or 2, you can still try to establish partial MI (Steenkamp & Baumgartner 1998).

The goal of tests of partial MI is to find out which of the loadings or intercepts differ across groups. If only one of these is different across groups, we know that any differences on the latent variable can either be caused by a difference in this loading/intercept, or by the true latent variable group difference. As long as there are at least two loadings and intercepts that are constrained equal across groups, we can make valid inferences about the differences between latent factor means in the model (Byrne et al., 1989).

TABLE 1
Example table for presenting the results

	χ^2	*df*	*p*	*CFI*	*TLI*	*RMSEA*	*BIC*	*AIC*
Model 1								
Model 2								
Model 3								
Model 4								

However, to be able to compare the sum scores or comparable observed means, we must have full scalar equivalence (Steinmetz, in press). If it can be established which specific item is problematic, questionnaires can be altered in future (Lugtig, Boeije, & Lensvelt-Mulders, 2011).

To establish partial invariance, choose between Model 1 or 2. Study the size of the loadings and/or intercepts, and constrain all loadings and intercepts, except for the one loading/intercept with the largest unstandardized difference, which is released. Subsequently, compare this new model with the old Model 1 or 2. If $\Delta\chi^2$ is now insignificant, partial invariance is established. If $\Delta\chi^2$ is still significant release another item, and continue until the item that causes MI not to hold is identified.

REPORTING THE RESULTS

After testing the invariance of the measurement model, the next step is to test the equality of factor means and correlations between the latent variables, across groups. Remember from the section on the parametrization of the CFA, that if we are interested in comparing the latent means across groups, we need the parametrization in Figure 2B. Note that if you constrain the factor mean to be zero in one of the groups the estimated latent factor means in the other groups tests for significant differences between the groups.

Reporting on MI results can be cumbersome to applied researchers, as it involves testing many different models, and reporting both on the model results (the size of the factor loadings, intercepts, etc.) as well as the model fit. As a rule, first report on the model fit of every model, and use summary tables to give an overview of all models tested. Once it is established what level of MI holds, report the results only for the final model. Example text:

> The CFA model with the unconstrained factor loadings and intercepts is shown in Figure 1. Two CFAs were conducted for Group 1 ($\chi^2 =$; $p =$; $CFI =$; $TLI =$; $RMSEA =$), and Group 2 ($\chi^2 =$; $p =$; $CFI =$; $TLI =$; $RMSEA =$), separately. Next, we tested for measurement invariance, see Table 1 for the fit indices. Model *X* has the lowest AIC/BIC value and therefore the best trade-off between model fit

and model complexity. The other fit indices of Model *X* indicated a good fit. Compared to the Group 2, Group 1 appeared to have a significantly lower mean factor score ($\Delta M =$; $p =$).

Manuscript received 21 March 2012
Revised manuscript accepted 16 April 2012
First published online 14 May 2012

REFERENCES

Arbuckle, J. L. (2007). *Amos 16.0 User's guide*. Pennsylvania, PA: Spring House.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., & Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

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

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315–345.

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

Jöreskog, K. G., & Sörbom, D. (1996–2001). LISREL 8: User's reference guide (2nd ed.). Lincolnwood, Illinois: Scientific Software International.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Lugtig, P., Boeije, H., & Lensvelt-Mulders, G. J. L. M. (2011). Change, what change? *Methodology*. DOI: 10.1027/1614-2241/a000043.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Rosseel, Y. (in press). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*.

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

Steinmetz, H. (in press). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*.

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*, 4–70.