Survey Research Methods (2012)
Vol.6, No.2, pp. 87-93
ISSN 1864-3361
http://www.surveymethods.org

How few countries will do? Comparative survey analysis from a Bayesian perspective

Joop Hox
Utrecht University

Rens van de Schoot
Utrecht University

Suzette Matthijsse
Utrecht University and Dept of Public Health,
University Medical Center, Rotterdam

Meuleman and Billiet (2009) have carried out a simulation study aimed at the question how many countries are needed for accurate multilevel SEM estimation in comparative studies. The authors concluded that a sample of 50 to 100 countries is needed for accurate estimation. Recently, Bayesian estimation methods have been introduced in structural equation modeling which should work well with much lower sample sizes. The current study reanalyzes the simulation of Meuleman and Billiet using Bayesian estimation to find the lowest number of countries needed when conducting multilevel SEM. The main result of our simulations is that a sample of about 20 countries is sufficient for accurate Bayesian estimation, which makes multilevel SEM practicable for the number of countries commonly available in large scale comparative surveys.

Keywords: Multilevel SEM, sample size, cross-national research, Bayesian estimation

1 Introduction

International cross-cultural and other comparative surveys involve a number of analysis issues. Measurement instruments must often be translated into different languages, which raises the issue of measurement equivalence. Can we assume that these instruments measure the same constructs in the same way? We need to assess whether we have measurement equivalence, and if not we need to investigate how we may correct measures in order to achieve measurement equivalence. Next, the analysis focuses on examining relationships within and between countries (or other contexts). That is, relationships can be established at the individual level within each country, but in comparative research the central issue is often the question whether such relationships are the same or different across countries. Finally, if we establish differences between countries, the question is whether country characteristics can explain such differences.

The classic approach to deal with these questions is structural equation modeling (SEM) using a multi-group analysis. This analysis method makes it possible to test equivalence of measurement models; special procedures for categorical data enable SEM to be used to estimate and test Item Response (IRT) models. Criteria for measurement equivalence were already formulated by Jöreskog (1971), for a review we refer to Vandenberg and Lance (2000), while for a discussion in the context of comparative surveys we refer to Harkness et al. (2010). If measurement equivalence may be assumed, multigroup SEM can be used to investigate the degree of equivalence of structural (substantive) models across countries.

When the number of countries is large, multi-group SEM becomes unwieldy. The setups become complicated, especially if subtle differences in measurement properties must be included. The statistical model for the structural differences also becomes complicated. Multi-group SEM is a fixed effects model, which means that it takes each group or country as given and the set of countries as the complete universe to generalize to. Unless many equality constraints are imposed, SEM estimates a unique set of parameter values for each different country, which results in a large model. Multilevel modeling (MLM) offers a different approach. Multilevel modeling treats the countries as a sample from a larger population. Instead of estimating a different parameter value for each country, it assumes a (normal) distribution of parameter values and estimates its mean and variance. This makes MLM much more parsimonious than SEM when the number of countries increases. In addition, differences between countries can be modeled formally using country-level variables. For a general introduction to multilevel modeling, we refer to Goldstein (2011), Raudenbush and Bryk (2002) and Hox (2010). Multilevel modeling for comparative surveys has been discussed by Hox, de Leeuw and Brinkhuis (2010) and Van de Vijver, van Hemert and Poortinga (2008). We mention in passing that multilevel modeling of comparative survey data not only poses statistical questions, but also methodological questions about the design. The statistical model assumes random sampling at all levels, while the survey design in fact does not use sampling at the country level. We can still use multilevel modeling, but its use is based on

Contact information: Joop Hox, Utrecht University, the Netherlands, e-mail: j.hox@uu.nl

the advantages of a model based approach where we can explicitly include country level explanatory variables and country level residual variation in the model, rather than a sample design based argumentation. We refer to Groves (1989) for a discussion of these two perspectives.

When multigroup SEM is used, the number of countries is not a principled issue. Multigroup SEM can be used to compare any number of groups. If the number of groups is huge, there may be practical analysis issues, such as the capacity of the software or the computer (or even the interpretational capacity of the analyst), but there is no formal lower or upper limit on the number of groups. In multilevel analysis, the second level sample size (in comparative surveys generally the number of countries) is an issue. The second level sample size must be large enough to permit accurate parameter estimates and associated standard errors.

Simulations have shown that multilevel regression modeling can be used with second-level samples as low as 20, provided that the interpretation focuses on the regression coefficients (Maas and Hox 2005). However, accurate estimation and testing of variances requires much larger sample sizes, Maas and Hox (2005) suggest 50 groups as a lower limit when variances are important. Structural equation modeling with latent variables relies on (co)variances, which suggests that for multilevel SEM even larger samples are needed for accurate estimation. Indeed, a simulation involving a two-level confirmative factor model shows that with fewer than 50 groups, the group level model parameters and their corresponding standard errors are not estimated with acceptable accuracy (Hox, Maas and Brinkhuis 2010). These simulations suggest that for accurate estimation at least 50 groups should be available.

Meuleman and Billiet (2009) have carried out a simulation study directly aimed at the question how many countries are needed for accurate multilevel SEM estimation in comparative surveys. They specified within country sample sizes to follow the sample sizes typically achieved in the European Social Survey. The number of countries was varied from 20 to 100. The simulation model at both the individual and the country level is a confirmative one-factor model for four indicator variables, plus a structural effect predicting the factor from an exogenous observed variable. Meuleman and Billiet (2009) conclude that a sample of 20 countries is simply not enough for accurate estimation. They do not suggest a specific lower limit for the country level sample size; instead, they discuss how model complexity and goal of the analysis affect the country level sample size requirements. However, their simulation results indicate that if we require that the 95% confidence interval for country level factor loadings lies in fact between 90 and 99 percent, which corresponds to a bias of about 5%, we require at least 60 countries. For 60 countries, the empirical alpha level for a test that the structural effect equals zero is 0.083, which is acceptable. With 40 countries, the empirical alpha level is 0.103, which is not acceptable (cf. Boomsma and Hoogland 2001). The power for a medium size structural effect at the country level is 0.523 with 60 countries, well below the value of 0.80 that Cohen (1988) recommends as a worth pursuing. In conclusion, Meuleman and Billiet confirm the suggestion that about 50 countries is the minimum sample size at the second level for accurate estimation in multilevel SEM.

The sample size requirements suggested by the simulation studies reviewed above imply that for most comparative surveys the country level sample sizes are problematic. For instance, the European Social Survey round four (2008) includes 30 countries (http://www.europeansocialsurvey.org), the third wave of SHARE (2008-2009) includes 13 countries (http://www.share-project.org), the 2007 wave of the mathematics survey TIMMS includes 36-48 countries (http://nces.ed.gov/timss), and the 2009 large scale educational assessment PISA sponsored by the OECD includes 65 countries (http://www.opisa.oecd.org). These country level sample sizes suggest that only the larger collaborative comparative surveys involve enough countries to consider employing multilevel SEM, but the majority appears too small to employ multilevel structural equation modeling.

Recently, Bayesian estimation methods have been introduced in structural equation modeling (Lee 2007). Bayesian estimation works well with lower sample sizes, and will not produce inadmissible parameter estimates such as negative variances. Bayesian methods generally imply prior information in the analysis, but when uninformative priors are used this has only a small effect on the resulting parameter estimates.

The goal of the current paper is to examine how well Bayesian estimation deals with the problem of estimating parameters in a multilevel SEM model with a small sample size at the country level. The paper starts with an introduction of Bayesian estimation methods and the issues involved in a Bayesian multilevel SEM analysis. Next, it describes the simulation design which is patterned after Meuleman and Billiet (2009). Our simulation design explicitly studies the accuracy of the estimation method with very small numbers of countries. The results and their implications for the analysis of comparative surveys are discussed in detail.

We provide a basic introduction of Bayesian statistics, but interested researchers could further refer to Lynch (2007) for an introduction to Bayesian estimation, and for technical details to Gelman, Carlin, Stern, and Rubin (2004). Bayesian structural equation modeling is discussed by Lee (2007) and Bayesian multilevel modeling by Hox (2010). In this paper we use the software Mplus (Muthén and Muthén 1998-2010) because it is often used by applied researchers. For the technical implementation of Bayesian statistics in Mplus, see Asparouhov and Muthén (2010).

2. Estimation methods in multilevel SEM

In this section we describe briefly different estimation methods for multilevel SEM, including Bayesian estimation. For a more elaborate accessible introduction we refer to Hox (2010), and for a statistical treatment we refer to Kaplan (2009).

Multilevel SEM assumes sampling at both individual and country levels. The individual data are collected in a

p-variate vector Y_{ij} (subscript i for individuals, j for groups). The data Y_{ij} are decomposed into a between groups (Group level) component $Y_B = \overline{Y}_j$, and a within groups (individual level) component $Y_W = Y_{ij} - \overline{Y}_j$. These two components are orthogonal and additive, thus $Y_T = Y_B + Y_W$. The population covariance matrices are also orthogonal and additive, thus $\sum_T = \sum_B + \sum_W$. Multilevel structural equation modeling assumes that the population covariance matrices $\sum_B$ and $\sum_W$ are described by distinct models for the between groups and within groups structure. Several approaches have been proposed to estimate the parameters of the multilevel SEM. Muthén (1989) suggests to approximate the full maximum likelihood solution by assuming equal group sizes, which leads to a limited information estimation method called MUML (for Muthén's Maximum Likelihood). A more accurate way to estimate a model for $\sum_B$ and $\sum_W$ is a Weighted Least Squares (WLS) method implemented in Mplus. Full maximum likelihood estimation for multilevel structural equation modeling requires to model the raw data. This minimizes the fit function given by

$$F = \sum_{i=1}^{N} log|\textstyle\sum_i| + \sum_{i=1}^{N} log(x_i - \mu_i)' \textstyle\sum_i^{-1}(x_i - \mu_i), \qquad (1)$$

where the subscript i refers to the observed cases, x_i to those variables observed for case i, and μ_i and $\sum_i$ contain the population means and covariances of the variables observed for case i. Mehta and Neale (2005) show that models for multilevel data, with individuals nested within groups, can be expressed as a structural equation model. The fit function (1) applies, with clusters as units of observation, and individuals within clusters as variables. Unbalanced data, here unequal numbers of individuals within clusters, are included the same way as incomplete data in standard SEM. The two-stage approaches that model $\sum_B$ and $\sum_W$ separately (MUML and WLS) include only random intercepts in the between groups model, the full ML representation can incorporate random slopes as well (Mehta and Neale 2005). Maximum likelihood estimation assumes large samples, and relies on numerical methods to integrate out random effects. In comparison, Bayesian methods are reliable in small samples, and are better able to deal with complex models. The Bayesian approach is fundamentally different from classical statistics (Barnett 2008). In classical statistics, the population parameter has one specific value, only we happen to not know it. In Bayesian statistics, we express the uncertainty about the population value of a model parameter by assigning to it a probability distribution of possible values. This probability distribution is called the *prior* distribution, because it is specified independently from the data. After we have collected our data, this distribution is combined with the Likelihood of the data to produce a *posterior* distribution, which describes our uncertainty about the population values after observing our data. Typically, the variance of the posterior distribution is smaller than the variance of the prior distribution, which means that observing the data has reduced our uncertainty about the possible population values.

More formally, let M be a statistical model with a vector of unknown parameters θ, for example regression parameters and correlations, and let Y be the observed data set with sample size n. In Bayesian estimation, θ is considered to be random and the behavior of θ under Y in such a Bayesian model can be described by

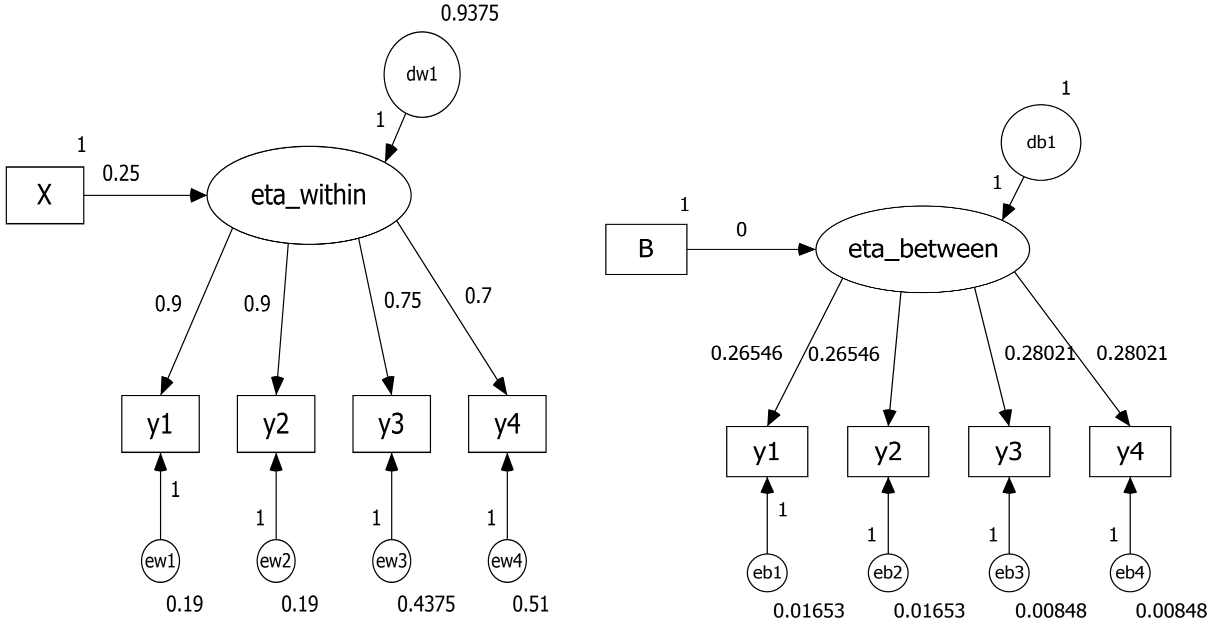
$$p(\theta|Y, M) \propto p(\theta|M) \times p(Y|\theta, M) \qquad (2)$$

where $p(Y|\theta, M)$ is the likelihood function, the information about the parameters in the data, $p(\theta|M)$ is the prior distribution, the information about the parameters before observing the data, and $p = (\theta|Y, M)$ is the posterior distribution, the information about the parameters after observing the data and taking the prior information into account.

For the prior distribution, we have a fundamental choice between using an informative prior or an uninformative prior. An informative prior is a peaked distribution with a small variance, which expresses a strong belief about the unknown population parameter, and has a substantial effect on the posterior distribution. In contrast, an uninformative or diffuse prior serves to produce the posterior, but has very little influence. An example of an uninformative prior is the uniform distribution, which simply states that all possible values for the unknown parameter are equally likely. Another example of an uninformative prior is a very flat normal distribution specified with an enormous variance. Sometimes such a prior is called an ignorance prior, to indicate that we know nothing about the unknown parameter. However, this is not accurate, since total ignorance does not exist. All priors add some information to the data, but diffuse priors add very little information, and therefore do not have much influence on the posterior. For our analyses we used the default prior specifications of Mplus which uses uninformative priors.

If the posterior distribution has a mathematically simple form, the known characteristics of the distribution can be used to produce point estimates and confidence intervals. However, in complex models the posterior is generally a complicated multivariate distribution, which is often mathematically intractable. Therefore, simulation techniques are used to generate random draws from the multivariate posterior distribution. These simulation procedures are known as Markov Chain Monte Carlo (MCMC) simulation. MCMC simulation is used to produce a large number of random draws from the posterior distribution, which is then used to compute a point estimate and a confidence interval (for an introduction to Bayesian estimation including MCMC methods see Lynch 2007). Typically, the marginal (univariate) distribution of each parameter is used.

Given a set of initial values from a specific multivariate distribution, MCMC procedures generate a new random draw from the same distribution. Suppose that $Z^{(1)}$ is a draw from a target distribution $f(Z)$. Using MCMC methods, we generate a series of new draws: $Z^{(1)} \rightarrow Z^{(2)} \rightarrow \ldots \rightarrow Z^{(t)}$. MCMC methods are attractive because, even if $Z^{(1)}$ is not from the target distribution $f(Z)$, if t is sufficiently large, in the end $Z^{(t)}$ is a draw from the target distribution $f(Z)$. Having good initial values for $Z^{(1)}$ helps, because it speeds up the conver-



(a) within (individual) level (b) between (country) level

Figure 1. Path diagram for within (individual) and between (country) level

gence on the target distribution, so the classical maximum likelihood estimates are often used as initial values for $Z^{(1)}$.

The number of iterations *t* needed before the target distribution is reached is referred to as the 'burn in' period of the MCMC algorithm. It is important that the burn in is complete. To check if enough iterations of the algorithm have passed to converge on the target distribution, several diagnostics are used. A useful diagnostic is a graph of the successive values produced by the algorithm. A different procedure is to start the MCMC procedure several times with widely different initial values. If essentially identical distributions are obtained after *t* iterations, we decide that *t* has been large enough to converge on the target distribution (Gelman and Rubin 1992).

An additional issue in MCMC methods is that successive draws are dependent. Depending on the distribution and the amount of information in the data, they can be strongly correlated. Logically, we would prefer independent draws to use as simulated draws from the posterior distribution. One way to reach independence is to omit a number of successive estimates before a new draw is used for estimation. This process is called *thinning*. To decide how many iterations must be deleted between two successive draws, it is useful to inspect the autocorrelations between successive draws. If the autocorrelations are high, we must delete many estimates. Alternatively, since each draw still gives some information, we may keep all draws, but use an extremely large number of draws.

The mode of the marginal posterior distribution is an attractive point estimate of the unknown parameter, because it is the most likely value, and therefore the Bayesian equivalent of the maximum likelihood estimator. Since the mode is more difficult to determine than the mean, the mean of the posterior distribution is also often used. In skewed posterior distributions, the median is an attractive choice. In Bayesian estimation, the standard deviation of the posterior distribution is comparable to the standard error in classical statistics. However, the confidence interval generally is based on the $1/2\ \alpha$ and $100 - 1/2\ \alpha$ percentiles around the point estimate. In the Bayesian terminology, this is referred to as the $100 - \alpha\%$ *credibility interval.* Mplus by default uses the median of the posterior distribution for the point estimate, and the percentile-based 95% credibility interval, which we have followed in our simulations. Bayesian methods have some advantages over classical methods. To begin, in contrast to the asymptotic maximum likelihood method, they are valid in small samples. Given the correct probability distribution, the estimates are always proper, which solves the problem of negative variance estimates. Finally, since the random draws are taken from the correct distribution, there is no assumption of normality when variances are estimated. In this study, we examine if Bayesian estimation will help in drawing correct inferences in multilevel SEM if the number of groups (countries) is relatively small. The simulation studies cited in the introduction typically find that at smaller country level sample sizes the parameter estimates themselves are unbiased, but that the standard errors are underestimated, which leads to poor control of the alpha level and undercoverage for the confidence intervals. We expect that the credibility intervals in our Bayesian estimation will perform better at the country level sample sizes usually encountered in comparative survey research.

3. Simulation design

The simulation design in this study closely follows Meuleman and Billiet (2009). The model at both the individual and the country level is a one-factor model with four indicators. There is one structural effect from an observed exogenous variable on the factor. Figure 1 shows the path diagram with the population parameter values.

The simulated data were generated from a population that has the same characteristics as used in Meuleman and Billiet (2009:48):

- The observed variables have a multivariate distribution.
- The intraclass correlation of the observed indicators is 0.08.
- The within level unstandardized factor loadings are 0.90, 0.90, 0.75 and 0.70.
- The between level factor unstandardized loadings are 0.27, 0.27, 0.28 and 0.28.
- The within level independent variable has an unstandardized effect of 0.25.
- The between level independent variable has an effect that is manipulated. One condition has an effect size of 0.00. The other effect sizes were manipulated to be 0.10 (small), 0.25 (medium), 0.50 (large) and 0.75 (very large), following Cohen's (1988) suggestions for effect sizes.
- The within level sample size is 1755.

Meuleman and Billiet generate data for five different numbers of countries: 20, 40, 60, 80 and 100. We have generated data for 10, 15 and 20 countries, with 1000 replications for each condition in our simulation design.

We have used Mplus 6.1 for our simulation. Mplus has a set of commands that can be used to tweak the Bayesian estimation process. Assuming that most users will use the default settings, we have not attempted to modify the default settings. The major issue here is to let Mplus automatically decide how long the burn-in must be. Mplus uses the Gelman-Rubin potential scale reduction (PSR; Gelman and Rubin 1992) to decide when the chain has converged. By default, two independent MCMC chains are produced, and the between and within chain variation is compared. When the between chain variance is smaller than 0.05, convergence is assumed. Lee (2007) discusses this and other Bayesian model checks, we will come back to this issue in the discussion.[1]

4. Results

The simulation results are summarized in Table 1, which also reports a selection of the results obtained by Meuleman and Billiet (2009).

Table 2: Statistical power for detecting the country level structural effect, for various effect sizes and country level sample sizes

	Number of countries					
	Bayesian estimation			ML estimation[1]		
	10	15	20	20	40	60
Effect size						
None (0.00)	0.03	0.05	0.05	0.16	0.10	0.08
Small (0.10)	0.04	0.06	0.06	0.18	0.15	0.16
Medium (0.25)	0.08	0.13	0.15	0.31	0.41	0.53
Large (0.50)	0.26	0.43	0.58	0.75	0.94	0.99
Very large (0.75)	0.67	0.89	0.97	1.00	1.00	1.00

[1] Parameters estimated by Meuleman and Billiet 2009.

Table 1 shows that, compared to ML estimation, Bayesian estimation tends to result in a much larger bias for the country level residual variance estimates, but to less bias for the country level factor loadings and the structural effect. The 95% credibility intervals show a much better coverage in Bayesian estimation than their maximum likelihood based counterparts. For example, with 20 countries the between level factor loadings have a mean absolute bias of 0.03 in Bayesian estimation, and -0.07 in Maximum Likelihood estimation. The actual coverage of the nominal 95% interval is 0.94 in Bayesian estimation, and 0.84 with Maximum Likelihood estimation, which is woefully inadequate.

Table 2 shows the proportion of p-values below 0.05, for various effect sizes. For an effect size of zero, the table shows the operating alpha level, which indicates the prevalence of the type I error. It is clear that ML estimation does not control the alpha level well, with an operating alpha level of 16% with twenty countries. Thus, if the nominal alpha level is set at the common value of 0.05, the prevalence of type I errors is actually 0.16. The alpha level is much better controlled in Bayesian estimation, where even at 10 countries the operating alpha level is 0.03, which is reasonably close to the nominal alpha level of 0.05.

Table 2 also shows that with a small number of countries the power in both Bayesian and Maximum Likelihood to detect anything but the largest effects is low. When the effect size is not zero, ML estimation does reject the null hypothesis more often than Bayesian estimation. For example, with 20 countries the power to detect a large effect is 0.58 in Bayesian estimation and 0.75 in Maximum Likelihood estimation. As we showed above, this increased power is at the expense of a very poorly controlled alpha level.

5. Discussion

The results of the simulation show that Bayesian estimation indeed can get away with far fewer countries than Maximum Likelihood estimation. Both the parameter estimates and the coverage of the 95% interval are surprisingly good. However, the between level residual error variances are estimated very poorly. We come back to this issue later in the

[1] One simulation run encountered convergence problems, which were solved by setting this convergence criterion to 0.01.

Table 1: Mean absolute bias for various country level sample sizes

	Number of countries					
	Bayesian estimation			ML estimation[1]		
	10	15	20	20	40	60
Parameter bias						
Within factor loadings	0.00	0.00	0.00	0.00	0.00	0.00
Within error variances	0.00	0.00	0.00	0.00	0.00	0.00
Within structural effect	0.00	0.00	0.00	0.00	0.00	0.00
Between factor loadings	0.50	0.04	0.03	-.07	-.03	-.02
Between error variances	0.59	0.33	0.24	-.10	-.05	-.04
Between structural effect	-.05	-.05	-.04	0.11	0.05	0.04
Coverage						
Within factor loadings	0.95	0.96	0.95	0.93	0.94	0.94
Within error variances	0.93	0.94	0.93	0.93	0.94	0.94
Within structural effect	0.95	0.95	0.96	0.93	0.94	0.94
Between factor loadings	0.96	0.96	0.94	0.84	0.89	0.91
Between error variances	0.95	0.95	0.94	0.81	0.88	0.90
Between structural effect	0.96	0.94	0.95	0.85	0.90	0.92

[1] Parameters estimated by Meuleman and Billiet 2009.

discussion, when we discuss convergence problems in the Bayesian context. With respect to statistical power, it is clear that Bayesian estimation does not solve the problem of small sample, only very large country level effects can be discovered when the number of countries is small.

The results also show that Bayesian estimation is not magic. With ten countries, problems start to show in the summary tables, but they are clearer when the simulation output is studied in more detail. For the condition with ten countries, each simulation run contains some outliers for the estimates of the error variances and corresponding standard errors, with estimates up to twenty times the population values. Such outliers would be recognized as such in a real analysis. The between model contains a total of 10 parameters, so it is not surprising that problems arise when the number of countries approaches the number of parameters in the between model. Simplifying the model, for instance by using the mean of the observed indicators instead of a latent variable would make estimation easier.

We briefly mentioned convergence problems and outlying estimates. In MCMC estimation, convergence means convergence of the chain to the correct distribution. In our simulation, we have decided to emulate a relatively nave user and therefore to follow all defaults implemented in the software (Mplus 6.1). We also used an automatic cut-off criterion to decide whether convergence had been reached. In one simulation run, we needed to change the default criterion to a more strict value. Textbooks introducing Bayesian statistics caution users to always use diagnostic tools such as plots of the iteration history (trace plots, c.f. Gelman, Carlin, Stern and Rubin 2004; Lynch 2007), and we completely agree with such recommendations. Obviously, in a simulation, visually inspecting trace plots for 15,000 replications times 20 parameters is not possible. In applied Bayesian analysis, we consider such inspection mandatory. In addition, especially in modeling situations as extreme as having as many parameters as we have countries, we recommend inspection of autocorrelations and setting much stricter criteria for convergence. In fact, if we deviate from the software defaults and set the convergence criteria much stricter, the bias in the residual variances at the country level becomes much smaller, at the cost of a much increased computation time.

Softwarewise, we have simply specified a different estimation method. From a principled standpoint, we have chosen a different kind of statistics. As a result, the 95% credibility interval now may correctly be interpreted as the interval that contains the population parameter with 95% probability. In our power table, we presented *p*-values. In the Bayesian case, this is not the normal *p*-value, but the so-called posterior predictive *p*-value. This is roughly interpreted as a standard p-value, but it is actually a different entity. Bayesian modeling in general prefers that decisions about parameters are based on credibility intervals, and that decisions about models are based on comparative evidence, such as information criteria or Bayes factors. A discussion of these issues is beyond the scope of this paper, but we believe that applied researchers should be aware that doing a Bayesian analysis is not just choosing a different estimation method.

In our analysis, we have chosen the default uninformative priors provided by Mplus. Other choices are possible. One interesting option is using an informative prior. For example, the default prior for a factor loading in Mplus is a normal distribution with a mean of zero and a very large variance (10^{10}). We have more prior knowledge than that. If we model seven-point answer scales with an underlying factor, using standard identifying constraints, we know that the (absolute) factor loadings will not exceed, say, the value ten. Why not use a prior distribution that reflects this knowledge? In doing so, we would become real subjectivist statisticians, a position that is far away from mainstream statistics. If we impose priors that describe only realistic parameter values, the convergence problem discussed above will disappear. But in small samples, such prior information could easily dominate the information in the data. In this paper, we have taken the position that this is undesirable, and prefer to work with uninformative priors.

Acknowledgements

Rens van de Schoot received a grant from the Netherlands Organisation for Scientific Research: NWO-VINI-451-11-008.

References

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using Mplus. Version 4.* Unpublished manuscript, accessed October 13, 2011 on http://www.statmodel.com/download/BayesAdvantages18.pdf.

Barnett, V. (2008). *Comparative statistical inference*. Chicester, UK: Wiley.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (p. 139-168). Chicago: Scientific Software International.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511.

Goldstein, H. (2011). *Multilevel statistical models*. Chicester, UK: Wiley.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., et al. (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. Chicester, UK: Wiley.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications*. NY: Routledge.

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

Kaplan, D. (2009). *Structural equation modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Lee, S. (2007). *Structural equation modeling: a Bayesian approach*. Chicester, UK: Wiley.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Berlin: Springer.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 85-91.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychological Methods*, *10*, 259-284.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study : How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*, 45-58.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Van de Vijver, F. R., van Hemert, D. A., & Poortinga, Y. H. (Eds.). (2008). *Multilevel analysis of individuals and cultures*. NY: Taylor & Francis.