Published in final edited form as:

Struct Equ Modeling. 2016 ; 23(4): 479–490. doi:10.1080/10705511.2016.1141355.

Inference Based on the Best-Fitting Model can Contribute to the Replication Crisis: Assessing Model Selection Uncertainty Using a Bootstrap Approach

Gitta H. Lubke^{1,2} and Ian Campbell¹

¹University of Notre Dame

²VU Amsterdam

Abstract

Inference and conclusions drawn from model fitting analyses are commonly based on a single “best-fitting” model. If model selection and inference are carried out using the same data model selection uncertainty is ignored. We illustrate the Type I error inflation that can result from using the same data for model selection and inference, and we then propose a simple bootstrap based approach to quantify model selection uncertainty in terms of model selection rates. A selection rate can be interpreted as an estimate of the replication probability of a fitted model. The benefits of bootstrapping model selection uncertainty is demonstrated in a growth mixture analyses of data from the National Longitudinal Study of Youth, and a 2-group measurement invariance analysis of the Holzinger-Swineford data.

Keywords

model selection uncertainty; replication crisis; bootstrap procedure

Comparing multiple alternative models fitted to the same data has become one of the most popular approaches of behavioral data analysis. As models have become increasingly complex and more highly parameterized, the number of alternative models can be substantial. In empirical studies inference concerning model parameter estimates, or, more generally, inference concerning the structural relations in a model is often based on a single best-fitting model that has been selected using the same sample data. Such a strategy is problematic because sampling fluctuation might result in the selection of a different “best-fitting” model in another sample. This paper addresses the issue of model selection uncertainty and describes a simple bootstrap-based method to quantify selection uncertainty. We show that model selection uncertainty needs to be taken into account when interpreting the “best-fitting model” or assessing parameter significance.

Correspondence concerning this article should be addressed to Gitta Lubke, 110 Haggard Hall, University of Notre Dame, Notre Dame, IN 46556. Contact: glubke@nd.edu.

Gitta H. Lubke, Department of Psychology, University of Notre Dame, USA; Biological Psychology, VU Amsterdam, The Netherlands.

Ian Campbell, Department of Psychology, University of Notre Dame, USA.

The practice of using the same data set to select a best-fitting model and to assess the significance of model parameter estimates or interpret the model structure is based on the often implicit assumption that the selected model is the true model that generated the data (Wang, 2010). However, this assumption does not hold in general. The sampling error related to model selection is ignored if the same data are used for inference.

Conceptually, the problem of using the same data set for model selection and inference can be understood as follows. If a set of models including the true model were fitted to a data set comprising the entire population of interest, then the true model would be selected because (1) it exactly represents the data generating process and (2) there is no sampling fluctuation. However, the situation is quite different in finite samples from the population. If a set of models were fitted to a large number of finite samples, it would be unlikely that the same model would be selected as the best-fitting model in all samples because of sampling fluctuation and the fact that the true model is rarely included in the set of fitted models. Fitted models represent simplified approximations of the true data-generating process, and they are more or less misspecified. The models can differ not only in how well they represent the population structure, but also in how well they adapt to idiosyncrasies present in each of the samples (for examples, see for instance Efron, 2014, or Roberts and Marin, 2010). If model selection is done in a single sample and significance tests are carried out for the model parameters in the selected best-fitting model, then the significance tests do not account for the fact that a different model might be selected in another sample from the same population. Rather, the significance tests are based on the assumption that the selected model is the true model without accounting for the uncertainty of model selection. Accounting for model selection uncertainty is equally important when the focus is on structural relations between observed and/or latent variables rather than parameter significance. It is therefore of great interest to quantify model selection uncertainty to protect against overly strong inference based on the best-fitting model that might not be replicated in a different sample.

Although the issue of using the same data for model selection and inference has been discussed previously in the statistical literature (Efron, 2014; Hurvich & Tsai, 1990; Zhang, 1992), it does not seem to have received much notice recently in behavioral research. Previous research has in fact focused mainly on improving model selection criteria (Browne & Cudeck 1992; Cudeck & Browne, 1983; Bollen, Harden, Ray, & Zavisca, 2014; Bollen, Ray, Zavisca, & Harden, 2012). Obviously, improved criteria increase the probability of finding the most adequate representation of the population structure, and research concerning different model selection criteria has provided important insights into their similarities and dissimilarities as well as their performance in varying contexts (Whittaker & Stapleton, 2006). However, even a perfect selection criterion can only point to the model that most adequately represents the population structure to the extent that this structure is captured by the sample data. Cross validation as implemented in the Cross Validation Index (Cudeck & Browne, 1983) attempts to address the effect of sampling fluctuation on model selection but requires data splitting which leads to a decrease in power which in turn decreases the probability to detect the correct structure (e.g., finding a small class in mixture analyses, see Lubke, 2012, or Lubke & Spies, 2008). Most importantly, model selection criteria do not provide a quantification of the probability that a given model is selected as the best fitting model in a new sample. Quantification of model selection uncertainty is crucial

because it provides an indication of the likelihood of replicating model comparison results. In case of substantial model selection uncertainty parameter inference or the interpretation of the model structure of the best fitting model has to be more cautious.

Given the increasing popularity of research involving model comparisons, it is likely that studies in which the same data set is used for model selection and inference contribute to replication problems in behavioral research. Several factors relating to the replication crisis have recently been examined, including the arbitrary inclusion of covariates and selective adjustment of sample size (Simmons, Nelson, & Simonsohn, 2011), the different and at times improper ways of handling outliers (Bakker & Wicherts, 2014), and the pervasive “file-drawer” problem affecting the field as a whole (Howard et al., 2009). The importance of addressing the replication crisis is evident. Unreliable findings delay scientific progress, and also increase the general public’s distrust in scientific findings (Chwe, 2014). The current paper addresses the fact that ignoring model selection uncertainty in model comparisons can also contribute to difficulties in replicating published results.

Hurvich and Tsai (1990) provided a first quantification of the impact of model selection uncertainty on parameter inference in the context of multiple regression. The data in their simulation study were generated for three predictors with non-zero effects, and 4 additional predictors with zero effects. Nested models were fitted with up to seven predictors, and model selection was based on the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). Conditional on model selection, the confidence regions of the regression parameter estimates were too narrow and did not include zero for models with truly zero predictors. The sample sizes ($N=20, 30, 50$) and number of Monte Carlo replications (500) in the Hurvich and Tsai simulations were small, most likely reflecting the computational limitations at the time.

Hurvich and Tsai’s study is especially important for structural equation modeling applications that include a search for relevant covariates. Empirical data collections commonly include larger numbers of potentially interesting background variables, and research studies often involve a selection of covariates. The first part of the current study extends Hurvich and Tsai’s simulations for multiple regression with a focus on Type 1 error rates. The simulation design includes increasing numbers of fitted models, and evaluates several model selection criteria in addition to AIC and BIC (e.g., the likelihood ratio test, LRT). This part of our study aims at quantifying the Type 1 error inflation that can occur when models with different sets of covariates are compared and significance tests are carried out in the best-fitting model using the same sample data. Keeping Type I error under control is one of the most important safeguards to prevent replication problems. Although Type I error inflation in this simulation is quantified for regression parameter estimates, it is reasonable to expect that a similar inflation occurs for model parameters of more complex structural equation models. Through quantification of Type I error inflation, this part of our study seeks to provide a motivation to seriously consider the consequences of ignoring model selection uncertainty in studies involving model comparisons and parameter inference in the best fitting model.

Data splitting is often recommended when exploratory analyses are followed up by confirmatory ones (Hastie, Tibshirani, & Friedman, 2013; Hurvich & Tsai, 1990; Tukey, 1980). It can be done by carrying out the model selection in one part of the data and then fitting the selected model to the second part of the data in order to investigate significance of parameter estimates. The main advantage of data splitting is that Type I error rates are asymptotically correct because new data are used for inference. A second advantage of data splitting is that it has low computational requirements. However, data splitting has two important drawbacks. First, it causes a reduction in power due to using only part of the data for selection, and for inference. Second, and most relevant in the context of selection uncertainty, is the fact that the selected model might not have a high probability of being selected if the same set of models were fitted to additional samples from the same population. A solution to this drawback is the focus of the second part of our study, in which we propose a bootstrap approach to quantify model selection uncertainty.

Statistical power plays a crucial role in discriminating between fitted models. This is especially critical in latent variable modeling if the model-implied mean and covariance structures of the compared models are relatively similar (MacCullum, Browne, & Sugawara, 1996; Satorra, Saris, & Pijper, 1991). Several papers have therefore encouraged researchers to conduct simulation studies to assess the power to discriminate between models (Jiao, Neves, & Jones, 2008; Merkle, You, & Preacher, 2014; Preacher & Merkle, 2012; Simonoff, 2000). Such an assessment often takes the form of simulating data under a model A, and calculating the power to discriminate between model A and alternative models. Although informative, this approach has the disadvantage that data simulated under a model are usually much cleaner than empirical data. One of the reasons for this is that models are much simpler than real world data-generating processes, and the well-behaved distributions used to generate simulated data are at best approximations of the more messy real data distributions. Therefore, parametric bootstrap-based power calculations should be regarded as “best case scenario” power assessments. Furthermore, quantifying the power to discriminate between models only partially addresses the problems related to model selection uncertainty. While informative, it does not answer the question of how likely it would be that a best-fitting model would be preferred again in a different sample. That is, simulation based power assessments do not show directly how well a model would replicate.

In the second part of this study, we propose a simple way to quantify model selection uncertainty in an empirical study when only a single sample is available. It consists of drawing multiple bootstrap samples from the original sample data and then carrying out a model comparison and selecting the best-fitting model in each bootstrap sample. Model selection can be based on a single model selection criterion or a rule involving multiple indices of choice (e.g., AIC, BIC, or LRT for nested models). Aggregating the results of the model comparisons over bootstrap samples provides selection rates for each of the fitted models. A selection rate can be interpreted as an estimate of the probability of replicating the selection of this model in a different sample: if a model is selected as the best-fitting model in 50% of the bootstrap samples, then it is reasonable to assume that the replication probability in a new sample drawn from a similar population will be about 0.5.

The most important advantage of bootstrapping model selection uncertainty is that results from fitting the same set of models to the original data can then be contextualized in terms of replication probability. Suppose several models have substantial non-zero selection rates. Although apparently there is insufficient power to clearly discriminate between these models, all these models can be useful in the interpretation of results. Similarities in structure across the models would strengthen evidence, while dissimilarities would highlight the parts of the structure in the data that should be considered as still uncertain. Replicating model selection in a large number of bootstrap samples also provides valuable information concerning model instability and convergence problems. In addition to convergence rates, the individual model results provide information about the most frequent reasons for improper solutions.

The computational burden of bootstrapping model selection uncertainty can be large, especially for more complex models fitted to categorical data. However, we show that distributing the bootstrap on a cluster reduces computation time to acceptable levels.

It should be noted that the proposed bootstrap quantification of model uncertainty is different from what has been coined bootstrap model selection in the literature (Shao, 1996). Bootstrap model selection consists of drawing bootstrap samples from the original data, and selecting a single best-fitting model based on minimization of the prediction error across bootstrap samples. It is also different from model averaging, which can improve the precision of prediction (Buja & Stuetzle, 2006; Efron 2014; Parker & Hawkins, 2015), and from bootstrap methods designed to obtain more stable estimates of variance components (Marcoulides, 1989) or an assessment of sampling error (Marcoulides, 1990). Rather, the proposed bootstrap approach replicates the model comparison and selection process in all bootstrap samples, and thus provides selection rates for each model included in the set of fitted models. The selection rate for a given model directly quantifies the likelihood that the selection of this model as the best fitting model is replicated in a new sample.

The paper is structured as follows. The first part consists of a simulation showing Type I error inflation in multiple regression if the same data are used to select a model and to assess statistical significance of regression parameters. In the second part, the proposed bootstrap quantification of model uncertainty is illustrated in two different empirical data sets, namely the National Longitudinal Survey of Youth (NLSY) and the Holzinger-Swineford study. The NLSY data are used for a growth mixture analysis with covariates, whereas the Holzinger-Swineford data are used to investigate measurement invariance in two groups.

Methods

Regression Model Selection and Type 1 Error Inflation in Simulated Data

The first goal of this study was to confirm the basic proposition that using the same data for model selection and parameter inference leads to an increase in Type I Error rates. We sought to reproduce Hurvich and Tsai's (1990) findings and extend them to a broader context.

The data-generating model in this simulation was

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

where X_1 - X_3 as well as the error ε were standard normal random variables, and the β 's, the regression weights, were chosen such that the overall effect size for the true model was $R^2 = .10, .30$, or $.50$. Sample sizes were $N = 100, 200, 500$, or $1,000$, and either 2, 5, 10, 20, or 40 different alternative models were fit to each generated sample. Each combination of simulation conditions (R^2 , N , and number of models considered) was simulated 10,000 times.

In order to remove the confounding effect of increasing model complexity, all fitted alternative (misspecified) models had a single additional, truly zero predictor. Thus each alternative model was

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon,$$

where X_4 was a differently generated predictor variable in each alternative model. Specifically, X_4 was either a truly zero main effect, an interaction, or a transformation (log, square, square root, or inverse transformation). Model comparisons were made among 2, 5, 10, 20, or 40 alternative models, and models differed only with respect to the type of 4th covariate added to the true model). In pilot runs we ensured that the alternative models had very similar selection rates independent of the type of additional effect. This is necessary to investigate the increase of Type I error as a function of the number of fitted alternative models without the potential confounding influence of type of effect. The simulation reflects a real life situation of searching for relevant covariate effects.

The sets of models fitted to each simulated data set had increasing numbers of alternative models but always contained the true model with only 3 predictors. A single, best-fitting model was chosen based on either AIC, BIC, the LRT, R^2 , or adjusted R^2 . Although there is no formal justification for using R^2 or adjusted R^2 for model selection, a researcher might be tempted to use these criteria when comparing models to select the model with the largest effect size possible. The Type I error rate for the selected models across all the samples was then calculated by assessing how often the model's fourth parameter, if present, was significant at an $\alpha = .05$ level.

Data splitting, using different parts of the data for model selection and parameter inference, has long been suggested as the best way to prevent model selection uncertainty from inflating Type I error rates (Hastie et al., 2013; Hurvich & Tsai, 1990; Tukey, 1980). We included data splitting in the simulation by adding analyses using only half of the data for model selection and the other half for parameter estimation and significance testing.

Bootstrapping Model Selection Uncertainty: Growth Mixture Analysis of Data From the National Longitudinal Survey of Youth (NLSY)

The NLSY data have been previously used to illustrate growth mixture modeling, and the part of the data used here is available on the Mplus website at <http://www.statmodel.com/>

[examples/penn.shtml](#). We saved data from $N = 1,172$ individuals included in the example described in Muthén and Muthén (2000). The univariate outcome was alcohol consumption during the past 30 days (“how often did you have 6 or more drinks on a single occasion during the past 30 days?”), and was measured originally on a 6-point Likert scale. We decreased the response categories to 3 (never, 1–3 times, more than 3 times), and treated the outcome as categorical in all analyses. Similar to the analysis of these data described in Muthén and Shedden (1999), we utilized 5 time points: when participants were 18, 19, 20, 24, and 25 years old. We included early onset of drinking (yes/no), gender (male/female), and ethnicity (African American vs. rest) as covariates (denoted as X in the model descriptions below).

From these data, we obtained 100 bootstrap samples of the same N with replacement. We then fitted the same set of models to (1) the original data, and (2) to the 100 bootstrap samples.

The set of 20 different growth mixture models fitted to the data included models without random effects (Nagin, 2005; Nagin & Land, 1993), models with different types of random effects (e.g., random intercepts, random intercepts and slopes, etc.), and models with different covariate effects. By quantifying selection rates for the different models, this part of our study aims at demonstrating that conclusions of a model fitting analysis should not only be based on a single best-fitting model but should take model selection uncertainty into account.

Random effects (if included) were estimated as class-specific because the equality of variances across classes is unlikely in empirical data. All models were fitted with class-invariant thresholds because the distributions of the univariate outcome at each of the 5 time points were similar. Noting that less complex models may need additional classes to adapt to the structure in the data (Lubke & Muthén 2005), the set of fitted models had the following features:

Models 1–4: Quadratic GMMs without random intercepts and slopes (2, 3, 4, 5 classes), class on X

Models 5–7: Quadratic GMMs with random intercepts but no random slopes (2, 3, 4 classes), class on X

Models 8–10: Quadratic GMMs with random intercepts, random linear slopes, covariance between intercepts and linear slopes, no random quadratic slope (2, 3, 4 classes), class on X

Models 11–12: Quadratic GMMs with random intercepts, linear and quadratic slopes, and their covariances (2 and 3 classes), class on X

Models 13–15: as models 5–7, but with additional class specific effects of intercepts on X

Models 16–18: as models 8–10, but with additional class specific effects of intercepts on X

Models 19–20: as models 11–12, but with additional class specific effects of intercepts on X

In all analyses, if the likelihood was not replicated using 100 random starts with the 10 best solutions iterated until the default convergence criterion was met, we increased the number of random starts to 4,000 and iterated the 40 best solutions to the end. Models were deemed properly converged if the likelihood was replicated under those conditions, and the Fisher information matrix as well as the matrix of first derivatives were positive definite.

We used the Monte Carlo feature with external data provided in Mplus to fit the 20 models to the 100 bootstrap samples (Muthén & Muthén, 2015) and saved parameter estimates and their standard errors as well as fit measures. Properly converged models were compared for each bootstrap sample, and the best-fitting model was selected based on the lowest BIC, which is the most commonly used fit index in mixture model comparisons (Nylund, Asparouhov, & Muthén, 2007). Convergence rates were computed as the number of times a model converged divided by the number of bootstrap samples (i.e., 100). Selection rates were computed as the number of times a converged model was selected divided by the total number of bootstrap samples. Selection rates are therefore independent of how many times a given model converged.

Bootstrapping Model Selection Uncertainty: Analysis of Measurement Invariance Using the Holzinger-Swineford Data

The second illustration of quantifying model selection uncertainty used data from the memory and speed subscales of the Holzinger-Swineford data collected for boys and girls attending the suburban Grant-White school (Holzinger & Swineford, 1939). A set of 14 different 2-group models was fitted to the male and female memory scale (4 items) and speed scale (6 items) data. The set included different bifactor models as well as correlated 2-factor models that were increasingly constrained to correspond to different levels of measurement invariance (Meredith, 1993). The analysis therefore reflected a situation where the interest is in evaluating MI when the factor structure is not entirely known. In particular, for cognitive performance data the bifactor model is an interesting option as it represents a general factor. Previous analyses of the math and verbal subscales of the Holzinger-Swineford data have shown that a general factor could be equated to math performance because no separate specific math performance factor could be distinguished (Erosheva & Curtis, 2013; Gustafsson, 2001; Holzinger & Swineford, 1939). Our set of fitted models included bifactor models with and without specific factors for memory and speed, respectively, as well as models without a general factor. Comparing these models should permit investigating whether speed and memory were equally contributing to overall performance.

The conceptual interpretation of the fitted models differs, and the aim was to show that quantifying model selection uncertainty is especially useful to contextualize the conclusions of the analysis. As in the illustration with the NLSY data, the models outlined below were fitted to the original data that consisted of a total $N = 145$ (64 girls and 81 boys) as well as to 1,000 bootstrap samples of the same N drawn with replacement from these data.

The 14 models fitted to the Holzinger-Swineford data were:

- Model 1: bifactor model with specific memory and speed factors, all estimated parameters group specific
- Model 2: as model 1, loadings constrained to be group invariant
- Model 3: as model 2, item intercepts fixed, factor means estimated
- Model 4: bifactor model, as model 1 but no specific memory factor, all estimated parameters group specific
- Model 5: as model 4, loadings constrained to be group invariant
- Model 6: as model 4, item intercepts fixed, factor means estimated
- Model 7: bifactor model, as model 1 but no specific speed factor, all estimated parameters group specific
- Model 8: as model 7, loadings constrained to be group invariant
- Model 9: as model 7, item intercepts fixed, factor means estimated
- Model 10: correlated 2-factor model, all estimated parameters group specific
- Model 11: as model 10, loadings constrained to be group invariant
- Model 12: as model 11, memory item intercepts fixed and factor means estimated for the memory factor, speed item means group specific
- Model 13: as model 11, speed item intercepts fixed and factor means estimated for the speed factor, memory item means group specific
- Model 14: as model 11, both speed and memory item intercepts fixed, both factor means estimated

Note that some subsets of models are nested, thus permitting likelihood ratio tests (i.e., models 1–3, models 4–6, models 7–9, models 10, 11, 12, 14, and models 10, 11, 13, and 14). Across all models, BIC and AIC were used to quantify separate model selection probabilities.

Results

Multiple Regression Simulations

As expected, increasing the number of alternative regression models fitted to a data set inflates Type I error rates if model selection and inference is done using the same data. Also as expected, Type I error rates are correct when data are split, but splitting leads to a decrease of power (see Table 1).

The results of the simulation are averaged over the three effect sizes ($R^2 = .10, .30$, and $.50$) because effect size did not significantly alter the Type I error rates. As can be seen in Table 1, the pattern of increasing error rates as more models are considered was very similar regardless of sample size, and this same general pattern of results was also seen across the different methods of model selection tested (AIC, BIC, LRT, R^2 , and adjusted R^2). The

results displayed in Table 1 correspond to using AIC to determine the best-fitting model. The results for the LRT were within a few percentage points of the AIC results, the results from R^2 and adjusted R^2 had even slightly higher inflation, and the results when using BIC with a large sample size were slightly less inflated, but the general pattern of high inflation persisted across all methods of model selection.

As expected, applying the split-half procedure in these simulations protected against Type I error inflation, regardless of how many alternative models were considered. When an incorrect model is selected as the best-fitting model, Type I error rates remain at the desired nominal level of .05. An incorrect model can be chosen because it provides a better fit to sampling fluctuation-induced idiosyncrasies; however, this will not result in significance of the truly zero effect when this model is then fitted to new data.

The split-half procedure had the anticipated drawback of reduced power. In the split-half procedure only half of the data was used to test the model parameters for significance, and thus the effective sample size was $N/2$. Table 1 shows that for a given effect size, the split-half procedures with sample size N had, as expected, about the same power as the original, no data splitting, procedure with sample size $N/2$. Using a split-half procedure protects against an inflation of Type I error rates, but it effectively cuts the sample size in half.

Analysis of the NLSY data

Results from fitting growth mixture models to the original data—The key features of the 20 fitted growth mixture models are presented in Table 2 together with the number of classes, the number of estimated parameters, and commonly used fit indices. Using the lowest BIC as a selection criterion, the best-fitting model was model 16. This model is a 2-class model with random intercepts and slopes, and it includes regression of the intercept factor on the three covariates in addition to the direct effect of class on the covariates. The second best fitting model was model 13, a slightly more parsimonious 2-class model similar to model 16, but without random slopes. A slightly more complex 2-class model with full random effects had the third lowest BIC.

Bootstrap-based quantification of model selection uncertainty of the NLSY analysis—By distributing the estimation of the 20 models fitted to 100 data sets of size $N = 1,172$ over 5 nodes of a computer cluster with adequate specifications, the required time was just short of 3 days, and was equal to the time necessary for the most complex model (model 20 with three random effects). Increasing the number of random starts for models with non-replicated likelihoods added approximately 3 days to the computation time. If the bootstrap analysis had been done on a single computer with an up to date processor, the total computation time would have been around 11 days, which is still more than reasonable when compared to the time researchers often spend conducting mixture analyses.

The selection rates for the 20 models are shown in Table 3 together with the key features of the models: number of classes, number of estimated parameters, and convergence rates. The mean BIC across 100 bootstrap samples and the corresponding standard deviation are also shown, which illustrates the variability of likelihood based fit indices across bootstrap samples.

As can be seen in Table 3, five models had selection rates of at least 1%. The model selected as the best-fitting model in the analysis of the original data (model 16) had a selection rate just below 50%, thereby demonstrating that the best-fitting model cannot unequivocally be considered the best approximation of the structure underlying the data. Three other models had selection rates of 10% or above. The models differed with respect to the number of classes, model complexity, and convergence rates. Not surprisingly, models with more random effects had lower convergence rates.

The bootstrap model selection analysis can aid in properly contextualizing the results of the analysis of the original data in the following ways:

1. Growth models without random effects can be safely excluded because selection rates were zero. This is consistent with the higher BICs of these models in the original data analysis.
2. Two or three classes are most likely sufficient to describe the structure of the data.
3. Two and three class models with class specific regression of intercepts on covariates generally provide a better fit to the data compared to models that only regress class on covariates.
4. Models with full random effects (i.e., intercept, linear and quadratic slope) and more than two classes have very low convergence rates, indicating insufficient resolution in the data to support this level of complexity.
5. The most parsimonious model with a higher selection rate is model 13, which has 20 estimated parameters and features 2 classes with random intercepts and class specific regressions of intercepts on covariates.
6. The model with the highest selection rate was also the best-fitting model in the original analysis. The selection rate of ~48% implies that interpretation of this model as the best-fitting model should be cautious because this model would only be replicated in slightly less than 50% of replication attempts.
7. One can compare the estimates of the average growth curves of model 16 obtained in the original analysis to those of models 13, 14, and 19.

As can be seen in Table 4, the 2-class models generally distinguish between a larger class that starts lower, might slightly increase (not significant in model 13) and then levels off. The smaller class starts higher, and remains essentially flat in models 16 and 19 (combined selection rate 0.628). Models 13 and 14 give a slight increase in class 2. A third class is estimated only in model 14 (selection rate 0.101), and has an elevated start, followed by a linear drop superseded by a quadratic increase. The pattern of class specific covariate effects is very similar in models 16 and 19.

Analysis of the Holzinger-Swineford (HS) Data

Results from fitting 14 different factor models to the original data—The model fit indices of the 14 different factor models are shown in Table 5 together with the key features of the models. As can be seen, model 6 had the smallest BIC when comparing all

models. This model is a bifactor model with a general factor loading on all items, and a single specific factor for the speed item. No specific factor was specified for the memory factor, resulting in a conceptual interpretation in terms of the general factor being mainly a memory factor. This model imposed measurement invariance of loadings and item intercepts. The likelihood ratio test (LRT) comparing model 4 and 5 (loading invariance test) was not significant (p -value = 0.0603) and the LRT for model 5 and 6 (item intercept invariance) was also in favor of the more parsimonious model (model 6, p -value = 0.2051). The next best fitting models were correlated 2-factor models with item means invariant for only the speed factor (model 12), or both speed and memory factor (model 14).

Bootstrap-based quantification of model selection uncertainty of the HS analysis—Fitting 14 factor models to 1,000 bootstrap samples of sample size $N=145$ took less than 5 minutes on a laptop computer.

As can be seen in Table 6, multiple models had selection probabilities above 0.10. The difficulty to discriminate between alternative models is due in part to the small sample size ($N=145$) and the fact that at least some of the models have model implied mean and covariance structures that are not dramatically different. Both of these facts contribute to lowering the power to discriminate between models. The analysis shows that even under these conditions, useful conclusions can be drawn in the context of selection rates. For instance, when comparing the different bifactor models, one can see that models featuring a general factor and a specific speed factor have higher selection rates than models with a general factor and a memory factor. The general factor is therefore more likely influenced by memory performance than by speed. Furthermore, the bifactor models with a specific speed factor are somewhat more likely to capture the structure in the population than the 2-factor models. These conclusions are supported by selection using AIC or BIC.

With respect to MI, the use of AIC or BIC leads to different conclusions. Interestingly, bootstrapped likelihood ratio tests for the nested models (e.g., models 4 vs 5, and model 10 vs 11) are in line with AIC selection rates and reject loading invariance (91% of bootstrap samples had LRT's favoring model 4, and 86.5% of bootstrap samples had LRT's favoring model 10). Invariance including item intercepts was equally not consistently supported, although invariance of speed items is more likely than invariance of memory items (33.5% of tests vs. 80.1% of tests rejecting invariance). These results also show that the LRT's carried out in the original data are overly confident in accepting the MI constraints due to insufficient power.

Discussion

Selecting a best-fitting model and drawing inferences using the same sample capitalizes on chance, leading to a potentially large inflation of Type I error when assessing statistical significance of model parameter estimates. The first part of our study showed that Type I error inflation can be serious even if only a small number of models is compared. Although the simulation quantified Type I error inflation for regression coefficients, it should be expected that similar inflations can occur for parameters of more complex structural equation models. Data splitting where different parts of the data are used for model selection

and inference provides correct Type I error rates, but in addition to a decrease of power, it does not directly address model selection uncertainty.

Even if parameter significance is not the main interest, it is necessary to take model selection uncertainty into account when interpreting results from a model comparison. The proposed bootstrap quantification of model selection uncertainty is simple to perform, and it provides crucial information concerning the replicability of results. Selection rates are calculated for all fitted models by computing the proportion of bootstrap samples in which a given model was selected as the best-fitting model. Selection rates therefore give an estimate of the probability of replicating the selection of a given model while at the same time providing information concerning the power to discriminate between alternative models. As illustrated in the analyses of measurement invariance with a subset of the Holzinger-Swineford data, low-powered empirical analyses are especially vulnerable to overly confident inferences based on the best-fitting model.

Bootstrapping model selection uncertainty provides a much needed context for model fitting results because it directly clarifies the limitations regarding the difficulty to discriminate between models, thus protecting against conclusions that reach too far. Both illustrations showed how selection rates can be used to interpret the results of a model comparison by taking into account not only the structure of a single best-fitting model, but also that of models with substantial non-zero selection rates. If bootstrapping selection uncertainty results in a single model having a selection probability close to one, then there is tremendous support in favor of that model.

The Holzinger-Swineford analysis showed that tests used for model comparisons such as the LRT are affected by power (Hohensinn, Kubinger, & Reif, 2014, MacCallum et al., 1996, Satorra et al., 1991), which is often forgotten in practice. While the LRT between models with and without MI constraints fitted to the original data were in favor of the more parsimonious MI models, this was not supported by the bootstrap results. More generally, fitted models can differ in conceptual interpretation (e.g., bifactor vs. correlated factors in the Holzinger-Swineford analysis, class sizes in the mixture analysis). When bootstrapping model selection uncertainty, the decision between models is not only based on one comparison in a single data set, but cast in the context of how likely a model selection is to be replicated. This is especially important when attempting to decide whether tests or questionnaires are measurement invariant over groups, or when drawing conclusions about the existence of latent trajectory classes (e.g., classes representing high-risk groups). In addition to selection rates, the proposed bootstrap procedure also provides non-convergence rates, which is especially useful in mixture analyses. In a single analysis, a model either converges or not. However, in the context of comparisons made in bootstrap samples, the resulting convergence rates provide additional information about model stability.

The two analyses with empirical data demonstrated that model selection uncertainty is not limited to mixture analyses; rather, it depends on the power to discriminate between models. Power can be calculated using parametric bootstrap (i.e., generate data under one model, and fit different models to the simulated data). However, the resulting power estimate is based on very clean data that do not include the messiness usually present in empirical data. The

parametric bootstrap can help determine necessary sample sizes in the planning phase of a study, but it should be kept in mind that resulting power estimates are based on a best case scenario where the data generating process is much more simple than in real world settings. The advantage of bootstrapping the empirical sample data to calculate the probability of replicating model selection results is that it reflects empirical data analysis results. In addition, it can be done when data are already collected.

Quantifying model selection uncertainty is different from improving a model selection criterion because it shows how likely a given model selection would replicate in a new sample. Improving selection criteria focuses on distinguishing between alternative models that are fitted to a single sample. In our illustrations we used BIC or AIC as model selection criteria. It is of course possible to use other information criteria when selecting the best model in each of the bootstrap samples, or to use a rule that combines multiple criteria (Bollen, Harden, Ray, & Zavisca, 2014; Bollen, Ray, Zavisca, & Harden, 2012). The value of the proposed bootstrap method is not to improve model selection *per se*, but to provide a direct quantification of model selection uncertainty. It would be interesting to combine an evaluation of model selection criteria with the quantification of model selection uncertainty.

In sum, our study underlines the dangers of relying on a single best-fitting model without evaluating the replication probability of the model. As a most welcome side effect, the proposed bootstrap approach requires researchers to decide on the set of fitted models *a priori*, which precludes a strategy of sequentially fitting models based on the results of previously fitted models. Adding the bootstrap based quantification of selection uncertainty to the analysis of a single data set enhances the quality of the results as it addresses the issue of validation by quantifying replication probability.

Acknowledgments

The first author was supported by DA-018673 (NIDA) and FP7-602768 (European Union). The computational work was partially done on clusters acquired through BCS-1229450 (NSF).

References

- Bakker M, Wicherts JM. Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*. 2014; 19(3):409–427. [PubMed: 24773354]
- Bollen KA, Harden JJ, Ray S, Zavisca J. BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling*. 2014; 21(1):1–19.
- Bollen KA, Ray S, Zavisca J, Harden JJ. A comparison of Bayes factor approximation methods including two new methods. *Sociological Methods & Research*. 2012; 41(2):294–324.
- Browne M, Cudeck R. Alternative ways of assessing model fit. *Sociological Methods and Research*. 1992; 21(2):230–230.
- Buja A, Stuetzle W. Observations on bagging. *Statistica Sinica*. 2006; 16(2):323–351.
- Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*. 1989; 105(3):456–466.
- Chwe, MS. Scientific pride and prejudice. *The New York Times*. 2014 Jan 31. Retrieved from <http://www.nytimes.com/2014/02/02/opinion/sunday/scientific-pride-and-prejudice.html>

- Cudeck R, Browne MW. Cross-validation of covariance structures. *Multivariate Behavioral Research*. 1983; 18(2):147–167. [PubMed: 26781606]
- Efron B. Estimation and accuracy after model selection. *Journal of the American Statistical Association*. 2014; 109(507):991. [PubMed: 25346558]
- Erosheva, EA., Curtis, SM. Dealing with rotational invariance in Bayesian confirmatory factor analysis (Technical Report No. 589). Seattle, WA: Department of Statistics University of Washington; 2013.
- Gustafsson, J-E. Measurement from a hierarchical point of view. In: Braun, HJ, Jackson, DN., Wiley, DE., editors. *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Routledge; 2001. p. 77-101.
- Hastie, T., Tibshirani, R., Friedman, JH. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd. New York: Springer; 2009.
- Hohensinn C, Kubinger KD, Reif M. On robustness and power of the likelihood-ratio test as a model test of the linear logistic test model. *Journal of Applied Measurement*. 2014; 15(3):252–66. [PubMed: 24992249]
- Holzinger, KJ., Swineford, F. *A study in factor analysis: The stability of a bi-factor solution*. Chicago, IL: The University of Chicago; 1939.
- Howard GS, Lau MY, Maxwell SE, Venter A, Lundy R, Sweeny RM. Do research literatures give correct answers? *Review of General Psychology*. 2009; 13(2):116–121.
- Hurvich C, Tsai C-L. The impact of model selection on inference in linear regression. *The American Statistician*. 1990; 44(3):214.
- Jiao Y, Neves R, Jones J. Models and model selection uncertainty in estimating growth rates of endangered freshwater mussel populations. *Canadian Journal of Fisheries and Aquatic Sciences*. 2008; 65(11):2389–2398.
- Lubke GH, Muthén B. Investigating population heterogeneity with factor mixture models. *Psychological Methods*. 2005; 10(1):21–39. [PubMed: 15810867]
- Lubke GH. Old issues in a new jacket: Power and validation in the context of mixture modeling. *Measurement*. 2012; 10:212–216. [PubMed: 24098093]
- Lubke, GH., Spies, J. Choosing a ‘correct’ factor mixture model: Power, limitations, and graphical data exploration. In: Hancock, GR., Samuelsen, KM., editors. *Advances in latent variable mixture models*. Charlotte, NC: Information Age Publishing; 2008. p. 343-362.
- MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*. 1996; 1(2):130–149.
- Marcoulides GA. The estimation of variance components of in generalizability studies: A resampling approach. *Psychological Reports*. 1989; 65:883–889.
- Marcoulides GA. Evaluation of confirmatory factor analytic and structural equation models using goodness-of-fit indices. *Psychological Reports*. 1993; 67(2):669.
- Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*. 1993; 58(4):525–543.
- Merkle EC, You D, Preacher KJ. Testing nonnested structural equation models. *Psychological Methods*. 2015 Advance online publication. <http://doi.org/10.1037/met0000038>.
- Muthén B, Muthén LK. Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism, Clinical and Experimental Research*. 2000; 24(6):882–91.
- Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*. 1999; 55(2):463–469. [PubMed: 11318201]
- Muthén, LK., Muthén, BO. *Mplus User’s Guide*. Seventh. Los Angeles, CA: Muthén & Muthén; 1998–2015.
- Nagin, D. *Group-based modeling of development*. Cambridge, MA: Harvard University Press; 2005.
- Nagin DS, Land KC. Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology*. 1993; 31(3):327–362.

- Nylund K, Asparouhov T, Muthén B. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*. 2007; 14(4):535–569.
- Parker C, Hawkins N. The use of bootstrap model averaging when estimating survival curves. *Value in Health*. 2015; 18(3):A23.
- Preacher KJ, Merkle EC. The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*. 2012; 17(1):1–14. [PubMed: 22268762]
- Roberts S, Martin MA. Does ignoring model selection when assessing the effect of particulate matter air pollution on mortality make us too vigilant? *Annals of Epidemiology*. 2010; 20(10):772–778. [PubMed: 20627768]
- Satorra A, Saris WE, Pijper WM. A comparison of several approximations to the power function of the likelihood ratio test in covariance structure analysis. *Statistica Neerlandica*. 1991; 45(2):173–185.
- Shao J. Bootstrap model selection. *Journal of the American Statistical Association*. 1996; 91(434): 655–665.
- Simmons JP, Nelson LD, Simonsohn U. False-positive psychology. *Psychological Science*. 2011; 22(11):1359–1366. [PubMed: 22006061]
- Simonoff E. Extracting meaning from comorbidity: Genetic analyses that make sense. *Journal of Child Psychology and Psychiatry*. 2000; 41(5):667–674. [PubMed: 10946758]
- Tukey, JW. Methodological comments focused on opportunities. In: Jones, LV., Hill, M., editors. *The collected works of John W. Tukey*. NJ: Bell Telephone Laboratories; 1980. 1980.
- Wang, H-F. The impact of preliminary model selection on latent growth model parameter estimates. (Unpublished doctoral dissertation). University of Maryland: Department of Measurement, Statistics, and Evaluation; 2010.
- Whittaker TA, Stapleton LM. The performance of cross-validation indices used to select among competing covariance structure models under multivariate nonnormality conditions. *Multivariate Behavioral Research*. 2006; 41(3):295–335. [PubMed: 26750338]
- Zhang P. Inference after variable selection in linear regression models. *Biometrika*. 1992; 79(4):741–746.

Type I Error Rates and Power as More Models are Considered, Both With and Without Data Splitting

Table 1

Number of Models Considered	Type I Error Rate in Best-Fitting Model				
	No Data Splitting				Split-Half Procedure
	N = 100	N = 200	N = 500	N = 1,000	
2	.0521	.0501	.0498	.0489	.0485
5	.1839	.1871	.1874	.1841	.0509
10	.3488	.3526	.3473	.3472	.0494
20	.4431	.4520	.4485	.4402	.0513
40	.6515	.6604	.6582	.6599	.0495
	Power when comparing 20 models				Power Across Sample Sizes
20	.49	.77	.98	.99	.25 .46 .84 .97

Note: No Data Splitting means no data splitting technique is employed. The Type I error rates for the Split-Half Procedure are averaged over sample sizes. All Type I error rates are averaged over effect size and based on using AIC as the selection criterion. Power results are for comparing 20 models with an effect size of $R^2 = .10$. Power results for the Split-Half Procedure are for samples sizes of N = 100, 200, 500, and 1,000.

Table 2

Model Fitting Results of the Original NLSY Data

model	key features	n class	n par	Like- lihood	BIC	AIC	sa BIC
1	no random effects	2	12	-3989.06	8002.11	8062.91	8024.79
2		3	19	-3900.57	7935.41	7839.14	7875.06
3		4	26	-3861.29	7906.31	7774.59	7823.73
4		5	33	-3837.37	7907.94	7740.74	7803.12
5	random intercept	2	14	-3914.35	7927.64	7856.71	7883.17
6		3	22	-3861.41	7878.27	7766.82	7808.40
7		4	30	-3834.96	7881.91	7729.92	7786.62
8		2	18	-3872.75	7872.69	7781.50	7815.52
9	random intercept & linear slope	3	28	-3839.34	7876.54	7734.68	7787.61
10		4	38	-3814.37	7897.27	7704.75	7776.57
11		2	24	-3861.18	7891.97	7770.38	7815.74
12		3	37	-3822.95	7907.36	7719.90	7789.84
13	random intercept, i on X	2	20	-3849.83	7840.98	7739.66	7777.46
14		3	31	-3825.48	7870.02	7712.96	7771.55
15		4	42	-3806.02	7908.83	7696.04	7775.42
16		2	24	-3829.67	7828.94	7707.35	7752.71
17	random intercept & linear slope, i on X	3	37	-3804.52	7870.49	7683.03	7752.96
18		4	50	-3783.84	7921.01	7667.68	7762.19
19		2	30	-3820.81	7853.63	7701.63	7758.34
20		3	46	-3794.08	7913.21	7680.16	7767.10

Note: All models were quadratic growth mixture models. Random effects were always estimated with class-specific variance. Abbreviations: n class = number of classes, n par = number of estimated parameters, BIC = Bayesian Information Criterion, AIC = Akaike Information Criterion, sa BIC = sample-size adjusted BIC, I on X = class specific regression of the intercept factor on the covariates. The model with the lowest BIC is in bold and italic, models within a BIC difference of 15 of the best fitting model are presented in bold.

Table 3

Model Selection Rates of 20 GMMs Fitted to 100 Bootstrap Samples From a Subset of the NLSY

model	key features	n classes	n param	selection rate BIC	converg rate	mean BIC	sd BIC
1	no random effects	2	12	0.000	1.00	8032.703	159.7806
2		3	19	0.000	1.00	7894.484	158.7380
3		4	26	0.000	0.98	7845.655	157.3442
4		5	33	0.001	0.92	7830.026	154.1123
5	random intercept	2	14	0	0.86	7876.661	161.6511
6		3	22	0.002	0.92	7822.524	161.3376
7		4	30	0.009	0.83	7806.706	161.3829
8		2	18	0.002	0.77	7824.691	151.1699
9	random intercept & linear slope	3	28	0.008	0.55	7773.826	147.2956
10		4	38	0.003	0.52	7822.893	149.8097
11		2	24	0	0.77	7832.977	153.7616
12		3	37	0.002	0.47	7848.165	151.3798
13	random intercept, i on X	2	20	0.159	0.98	7785.352	154.8668
14		3	31	0.101	0.79	7812.968	157.4791
15		4	42	0.008	0.52	7823.341	148.0612
16		2	24	0.478	0.89	7761.509	154.7056
17	random intercept & linear slope, i on X	3	37	0.076	0.58	7791.234	162.0160
18		4	50	0	0.31	7792.152	143.6079
19		2	30	0.150	0.85	7794.638	153.0162
20		3	46	0	0.34	7852.469	147.1468

Note: All models were quadratic growth mixture models. Random effects were always estimated with class-specific variance. Abbreviations: n classes = number of classes, n param = number of estimated parameters, converg rate = convergence rate, mean BIC = Bayesian Information Criterion averaged over 100 bootstrap samples, sd BIC = the standard deviation of the BIC over 100 bootstrap samples, I on X = class specific regression of the intercept factor on the covariates.

Table 4

Average Estimated Growth Curve Parameters and Class Specific Covariate Effects of Models 13, 14, 16, and 19 in the Analysis of the Original Data

	model features	class	class size	mean of intercept	mean of linear slope	mean of quadr slope	intercept on X
model 13	random intercept, intercept on X	class 1	0.625	-5.308 (1.647)	0.113 (0.417)	-0.115 (0.054)	1.938 (0.443) -2.312 (1.803) 1.294 (1.211)
		class 2	0.375	-3.855 (0.352)	0.478 (0.206)	-0.065 (0.048)	0.965 (0.696) -1.345 (0.288) 1.095 (0.677)
model 14	random intercept, intercept on X	class 1	0.647	-2.849 (0.230)	0.400 (0.088)	-0.160 (0.029)	1.727 (0.261) -2.199 (0.273) 0.083 (0.471)
		class 2	0.262	-2.293 (0.392)	0.359 (0.098)	-0.016 (0.027)	1.041 (0.353) -1.212 (0.274) 1.338 (0.326)
		class 3	0.030	0.386 (1.499)	-1.583 (0.460)	0.197 (0.065)	0.628 (0.707) fixed -1.801 (1.242)
model 16	random intercept and lin slope	class 1	0.834	-1.648 (0.244)	0.521 (0.075)	-0.140 (0.018)	2.002 (0.230) -1.021 (0.329) 0.615 (0.465)
		class 2	0.166	-1.413 (1.199)	-0.553 (0.372)	0.168 (0.109)	1.489 (0.475) -4.155 (0.857) 1.388 (0.788)
model 19	full random effects	class 1	0.681	-1.674 (0.479)	0.490 (0.083)	-0.121 (0.026)	2.082 (0.313) -0.518 (0.814) -0.064 (0.628)
		class 2	0.320	0.102 (0.525)	-0.007 (0.378)	-0.016 (0.122)	2.087 (0.476) -4.107 (0.759) 1.095 (0.488)

Note: Parameter estimates that did not reach significance without controlling for multiple testing are in italic. The three covariate effects are gender (males=1), African American vs. rest (AA=1), early onset of drinking (early onset=1).

Table 5

Model Fitting Results of the Original Holzinger-Swineford Data

	key features	n param	Likelihood	AIC	BIC	chi- squared (df)	p-value chi- squared
model 1	bifactor, 2 spec factors, no MI	80	<i>no convergence</i>				
model 2	as m1, loading MI	60	-2824.63	5769.26	5947.86	94.64 (70)	0.027
model 3	as m2, means MI	53	<i>residual covariance matrix of females not positive definite, negative residual variance of one item</i>				
model 4	bifactor, specific speed factor, no MI	68	-2818.80	5773.60	5976.01	83.00 (62)	0.039
model 5	as m4, loading MI	54	-2830.30	5768.59	5929.33	105.97 (76)	0.013
model 6	as m5, means MI	46	-2835.77	5763.54	5900.47	116.9 (84)	0.010
model 7	bifactor, specific memory factor, no MI	72	-2819.68	5783.35	5997.68	84.735 (58)	0.013
model 8	as m7, loading MI	56	-2834.32	5780.63	5947.33	114.012 (74)	0.002
model 9	as m8 means MI	48	-2842.40	5780.80	5923.68	130.18 (82)	0.001
model 10	correlated 2- factor, no MI	62	-2826.95	5777.89	5962.45	99.272 (68)	0.008
model 11	as m10, loading MI	54	-2835.85	5779.72	5940.45	117.09 (76)	0.002
model 12	as m11, memory means MI	49	-2838.08	5774.15	5920.01	121.53 (81)	0.002
model 13	as m11, speed means MI	51	-2841.65	5785.29	5937.10	128.67 (79)	0.0004
model 14	as m11, speed and memory means MI	46	-2843.87	5779.74	5916.67	133.12 (84)	0.001

Note: Chi-squared stands for chi-squared test of model fit, and chi-squared p-value is the associated p-value. All other abbreviations are as in previous tables.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 6
Model Selection Rates of 14 Factor Models Fitted to 1000 Bootstrap Samples of the Holzinger-Swineford Data

	key features	n param	selection rate AIC	Selection rate BIC	converg rate
model 1	bifactor, 2 spec factors, no MI	80	0.009	0.000	0.012
model 2	as m1, loading MI	60	0.054	0.025	0.234
model 3	as m2, means MI	53	0.044	0.027	0.235
model 4	bifactor, specific speed factor, no MI	68	0.258	0.056	0.692
model 5	as m4, loading MI	54	0.095	0.139	0.870
model 6	as m5, means MI	46	0.062	0.129	0.726
model 7	bifactor, specific memory factor, no MI	72	0.189	0.054	0.706
model 8	as m7, loading MI	56	0.035	0.092	0.992
model 9	as m8 means MI	48	0.012	0.058	0.989
model 10	correlated 2-factor, no MI	62	0.150	0.083	0.956
Model 11	as m10, loading MI	54	0.031	0.110	0.996
model 12	as m11, memory means MI	49	0.041	0.122	0.998
model 13	as m11, speed means MI	51	0.007	0.050	0.995
model 14	as m11, speed and memory means MI	46	0.013	0.055	0.997

Note: Based on BIC, the two models that fitted best in the original data analysis (model 5 and 6) have a combined replication probability of 27%. Based on AIC, the most lenient bifactor model with a specific speed factor and no invariance is the most likely model with a 0.258 replication probability. Abbreviations are n param = number of estimated parameters, converg rate = convergence rate.