Reference:

Jobst, L. J., Bader, M., & Moshagen, M. (in press). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods.* https://doi.org/10.1037/met0000423

A Tutorial on Assessing Statistical Power and Determining Sample Size for Structural Equation Models

Lisa J. Jobst, Martina Bader, and Morten Moshagen
Institute of Psychology and Education, Ulm University

Author Note

Lisa J. Jobst https://orcid.org/0000-0002-4088-5451

Martina Bader https://orcid.org/0000-0002-5706-8933

Morten Moshagen https://orcid.org/0000-0002-2929-7288

We have no known conflict of interest to disclose.

Some of the ideas elaborated in the present manuscript are published at https://github.com/moshagen/semPower to illustrate functions of the R package semPower.

We thank Luisa Horsten and David Scholz for their thoughtful comments on the manuscript.

Correspondence concerning this article should be addressed to Lisa J. Jobst, Research Methods, Institute of Psychology and Education, Ulm University, 89069 Ulm, Germany. Email: lisa.jobst@uni-ulm.de

Abstract

Structural equation modeling (SEM) is a widespread approach to test substantive hypotheses in psychology and other social sciences. However, most studies involving structural equation models neither report statistical power analysis as a criterion for sample size planning nor evaluate the achieved power of the performed tests. In this tutorial, we provide a step-by-step illustration of how *a priori*, *post hoc*, and *compromise power analyses* can be conducted for a range of different SEM applications. Using illustrative examples and the R package semPower, we demonstrate power analyses for hypotheses regarding overall model fit, global model comparisons, particular individual model parameters, and differences in multi-group contexts (such as in tests of measurement invariance). We encourage researchers to yield reliable – and thus more replicable – results based on thoughtful sample size planning, especially if small or medium-sized effects are expected.

Keywords: structural equation modeling, statistical power, sample size planning, goodness of fit

A Tutorial on Assessing Statistical Power and Determining Sample Size for Structural Equation Models

Structural equation modeling (SEM) is a powerful modeling technique that can be applied to many research questions in and beyond psychology (for reviews, see Fan et al., 2016; MacCallum & Austin, 2000). The main purpose of SEM is to model relations between manifest indicator variables and latent constructs as well as among latent constructs by combining path analytic and factor analytic techniques. In substantive research, SEM is commonly used as a tool to investigate various types of hypotheses. One type of hypothesis pertains to whether or not the model actually provides an adequate description of the underlying population, that is, whether the model fits the data. Other types of hypotheses may involve the comparison of competing models imposing different structures on the observed variables (such as models assuming a different number of factors) or may refer to a particular parameter in the model (such as whether a latent regression slope differs from zero).

Regardless of the particular hypothesis under scrutiny, any hypothesis test is associated with two decision errors that may occur, the α error of incorrectly rejecting a true null hypothesis and the β error of incorrectly retaining a false null hypothesis. Statistical power is defined as the complement of the β error $(1 - \beta)$ and thus gives the probability to correctly reject a false null hypothesis. Whereas the probability of committing an α error is always known and usually fixed in advance (typically $\alpha = .05$ or $\alpha = .01$), the β error – and thus statistical power – additionally depends on the sample size and the extent to which the null hypothesis is factually wrong (the magnitude of effect). Therefore, whenever a certain effect is expected (such as that a regression slope is hypothesized to differ from zero), researchers should not only control the α error probability, but should also ensure that statistical power is sufficiently high to be able to actually detect the hypothesized effect.

Despite its importance, the concept of statistical power has often been neglected in the past, leading to underpowered studies detecting an actual effect only on chance level (e.g., Cohen, 1962, 1992; Sedlmeier & Gigerenzer, 1989). Cohen (1962) as well as Sedlmeier and Gigerenzer (1989) assessed the actually achieved power of a test after the study was performed (so that the sample size is fixed), which is called *post hoc power analysis*. This procedure is useful to retrospectively determine the probability to find an actual effect based on a given study design. An arguably more important type of power analysis is called *a priori power analysis*, which serves to determine the required sample size to achieve the desired power to detect a certain effect in advance of a study. Although it seems rather obvious to aim at a sufficiently large sample to reliably identify a hypothesized effect, sample size planning in the context of SEM often relies on rules of thumb (see Kyriazos, 2018) rather than on power considerations (e.g., MacCallum & Austin, 2000). Improper sample size planning can have severe consequences, however: On the one hand, small samples may lead to actual effects not being detected. On the other hand, it is sometimes more likely to yield at least one significant result by performing several underpowered studies compared to one study with a large sample size (Bakker et al., 2012). Both issues can result in inconsistencies and artifacts in the scientific literature and, as a consequence, to biased assumptions about effect sizes in the population (Maxwell, 2004). Hence, it is not surprising that underpowered studies are considered as one major factor contributing to the so-called replication crisis (for details, see Anderson & Maxwell, 2017; Maxwell et al., 2015). Because both errors – rejecting a true null hypothesis and incorrectly retaining a false null hypothesis – can have severe consequences, a third type of power analysis (*compromise power analysis*; Faul et al., 2007) aims at providing a decision rule to balance both error probabilities.

As statistical power refers to the probability to detect a certain effect, it is important to be familiar with various types of effects that regularly occur in SEM. In general, the magnitude of an effect is given by the discrepancy between two competing, nested[1] models. Depending on the models that are compared, different types of hypotheses can be tested, which we call *global* versus *local hypothesis tests*. Global hypothesis tests assess the overall structure of a model. If the hypothesized model is compared with the saturated model[2], the test evaluates whether the hypothesized model perfectly represents the data. If the hypothesized model is compared with another nested and overidentified (i.e., not saturated) model, the test evaluates whether the more restrictive model involving less free parameters fits the data as well as the more general model involving more free parameters. This type of hypothesis test can, for instance, be applied to investigate whether a more restrictive one-factor model describes the data equally well as a more general two-factor model.

In contrast to global hypothesis tests, the evaluated hypothesis in *local hypothesis tests* refers to one or a few particular model parameters, but not to the overall model structure. A typical example of local hypothesis testing is the comparison of a more general model that freely estimates a parameter of interest (e.g., a slope parameter in a latent regression model) with a more restrictive alternative model constraining the same parameter to a certain value such as zero.

Both global and local hypothesis tests also arise in the context of multi-group models. An example of global hypothesis testing in such a context are tests of measurement invariance across groups, such as comparing a model freely estimating all factor loadings against a restricted model constraining all loadings to equality across groups. A typical scenario for local hypothesis testing involving multiple groups is a test evaluating whether the value of a latent regression slope differs across groups.

[1] Nestedness means that the parameter space of the more restrictive model represents a subspace of the parameter space of the more general model.

[2] A saturated model is a model in which the number of free parameters matches the number of observed values (such as observed means and covariances), so that the degrees of freedom of the model are zero.

In this tutorial, we aim at providing an accessible introduction to power analysis in SEM, addressing all three types of power analyses (a priori, post hoc, and compromise). We illustrate step by step – considering local as well as global hypothesis testing – how researchers can determine the required sample size to detect an effect of interest, how to compute the probability to detect a certain effect for a given study design, and how to make a balanced decision between the two errors of rejecting a true null hypothesis and retaining a false null hypothesis. The remainder of this tutorial is structured as follows: after providing a brief description of the statistical background underlying power analyses, we explain global hypothesis testing based on differences in a measure of fit between the saturated and a hypothesized model illustrating a priori, post hoc, and compromise power analyses. Then, we describe a priori power analyses for global hypothesis testing in not explicitly specified models (i.e., the involved models are not directly specified, but only their extent of misfit is defined) as well as global and local hypothesis testing involving two explicitly specified models. The final part of the tutorial covers hypothesis testing in a multi-group context illustrating a priori power analyses based on differences in fit indices as well as in model parameters. The article closes with a discussion of the implications and limitations regarding the herein described approach to determine power.

Theoretical Background

Parameter Estimation and Model Test

The parameters $(\boldsymbol{\theta})$ of a structural equation model (such as loadings) are estimated by minimizing a certain fit function. The minimum of the fit function quantifies the degree of discrepancy between the model and the data. A very general fit function is weighted least squares (WLS), which is in the population defined as

$$F_0 = \left(\boldsymbol{\sigma_0} - \boldsymbol{\sigma}(\boldsymbol{\theta})\right)' \boldsymbol{W}^{-1} \left(\boldsymbol{\sigma_0} - \boldsymbol{\sigma}(\boldsymbol{\theta})\right) \tag{1}$$

with $\boldsymbol{\sigma_0}$ and $\boldsymbol{\sigma}(\boldsymbol{\theta})$ representing vectors containing the unique elements of the population variance-covariance matrix $\boldsymbol{\Sigma_0}$ and the model-implied variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, respectively, and $\boldsymbol{W}$ denoting a weight matrix (for details, see Bollen, 1989; Browne, 1974, 1984). The minimized population value of this function, F_0, is a scalar quantifying the degree of discrepancy between the predictions by the model, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, and the data, $\boldsymbol{\Sigma_0}$. As the population variance-covariance matrix is unknown in practice, a sample estimate $\hat{F}$ is obtained by replacing $\boldsymbol{\sigma_0}$ by a vector $\boldsymbol{s}$ containing the unique elements of the empirically observed sample variance-covariance matrix, $\boldsymbol{S}$. The particular choice of the weight matrix determines the particular estimator. Maximum likelihood (ML) estimates, for example, can be obtained by using the unique elements of the model-implied variance-covariance matrix based on parameter estimates, $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, as weights (Browne, 1974). Throughout this tutorial, we rely on the ML fit function, as ML is the most popular estimation method in the context of SEM. However, the herein described routine to determine power is valid for any fit function, as long as the respective test statistic can be assumed to follow asymptotically a central chi-square distribution when the null hypothesis is true and a noncentral chi-square distribution when the null hypothesis is false (MacCallum et al., 1996; Olsson et al., 2004; Steiger et al., 1985).

The population minimum of the fit function F_0 ranges from zero to infinity. If the model-implied variance-covariance matrix equals the variance-covariance matrix in the population, F_0 turns zero (i.e., the model fits perfectly, so that the effect equals zero). In contrast, non-zero values of F_0 indicate a discrepancy between both matrices (i.e., the model does not fit exactly, so that the effect is larger than zero). Correspondingly, the sample estimate $\hat{F}$ can be used to measure the magnitude of misfit of the model to the sample data, with larger values of $\hat{F}$ indicating larger misfit. Based on $\hat{F}$, a likelihood-ratio model test (LRT) statistic can be constructed to evaluate the null hypothesis that the model-implied variance-covariance matrix equals the variance-covariance matrix in the population (i.e., $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_0$). The resulting (Wishart) LRT statistic is

$$T = (N - 1) \cdot \hat{F} \tag{2}$$

with *N* indicating the sample size. If the null hypothesis is true and certain assumptions hold, as for instance multivariate normal data in the case of ML estimation (for details, see Bollen, 1989; Browne, 1974), the LRT statistic asymptotically follows a central chi-square distribution. This is why the LRT is often called "chi-square model test" and why we refer to *T* as the "chi-square value". The degrees of freedom (*df*) are given by $df = 0.5 \cdot p \cdot (p + 1) - q$, where p denotes the number of manifest variables and q represents the number of freely estimated parameters. When performing an LRT for two explicitly specified, nested models, inferences are drawn from the difference between the chi-square values of both models and the *df* are given by the difference between the *df* of both models (Bollen, 1989).

Statistical Power

Statistical hypothesis tests are generally associated with two types of potential errors. The α error (type I error) indicates the probability that a true null hypothesis (i.e., the more restrictive model does not differ in terms of fit from the more general model) is rejected. The β error represents the probability that a false null hypothesis (i.e., the more restrictive model actually differs in terms of fit from the more general model) is retained. Statistical power is defined as the probability to reject a false null hypothesis and is given by $1 - \beta$.

To perform a goodness of fit test by means of the LRT, a critical value of the central chi-square distribution T_c is determined corresponding to a certain α error probability (e.g., α = .05). If the observed chi-square value exceeds the critical value T_c, the null hypothesis is rejected. Alternatively, the cumulative probability of obtaining a test statistic at least as large as the observed chi-square value is obtained from the central chi-square distribution. The associated p-value is then compared to a certain threshold defined by the type I error probability. Thus, this procedure rejects a true null hypothesis with a probability of α.

If the null hypothesis is actually false and certain conditions are met (for details, see MacCallum et al., 1996; Olsson et al., 2004; Steiger et al., 1985), the LRT statistic follows a noncentral chi-square distribution $\chi^2(df, \lambda)$ with λ denoting the noncentrality parameter. The expected value of the noncentral chi-square distribution is $df + \lambda$, so that λ leads to a shift to the right compared to the central chi-square distribution (see Figure 1). The noncentrality parameter λ is a function of the sample size and the magnitude of effect (i.e., the actual population discrepancy)

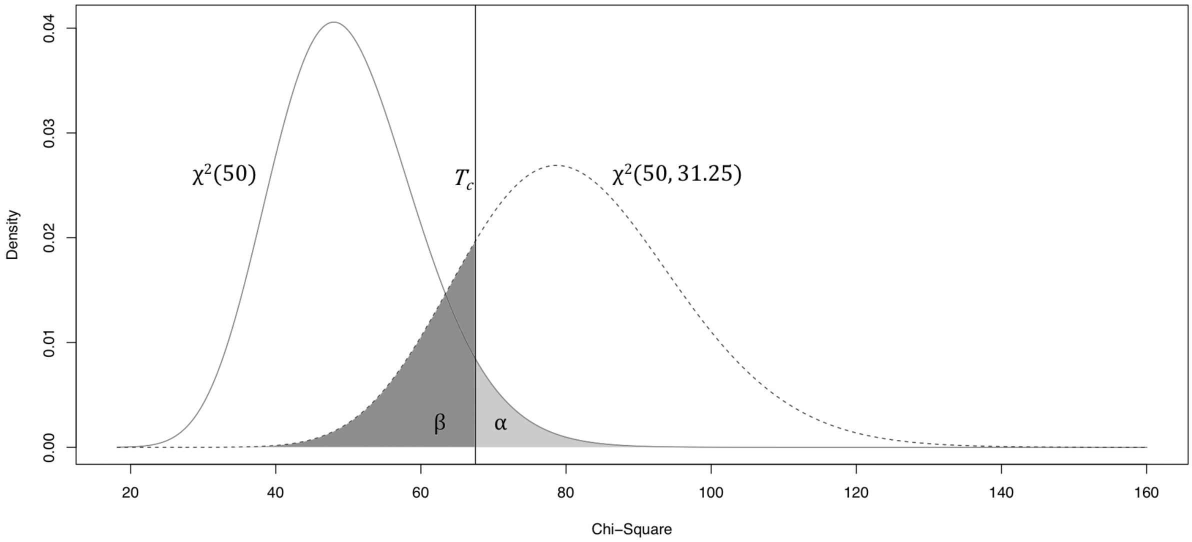
$$\lambda = (N - 1) \cdot F_0. \tag{3}$$

The distance between the central and the noncentral distribution therefore increases with the sample size and the magnitude of effect.

When the null hypothesis is actually false, the desired behavior of any statistical test obviously is to indicate rejection of the null hypothesis. As is evident from Figure 1, a false null hypothesis is incorrectly retained when the observed value of T is smaller than the critical value T_c. Thus, the β error probability is given by the probability of values smaller than T_c in the noncentral chi-square distribution. The statistical power to reject a false null hypothesis (i.e., $1 - \beta$), in turn, is given by the probability of values equal or larger than T_c in the noncentral chi-square distribution. As illustrated in Figure 1, α and β errors are directly related because increasing the α error probability reduces the β error probability, and vice versa. Because the noncentrality parameter λ determines the extent of shift in the noncentral chi-square distribution, the overlap between the distributions is reduced as λ increases. Thus, when α is held constant at a particular value such as .05, statistical power will increase with the sample size N and the magnitude of effect in terms of the degree of discrepancy indicated by F_0.

Figure 1

Central (Solid Line) and Noncentral (Dashed Line) Chi-Square Distributions



Note. df = 50 and $\lambda = 31.25$. The vertical line indicates the critical value T_c. The light gray area represents the α error probability and the dark gray area represents the β error probability.

Expressing the Degree of Discrepancy via Effect Size Measures

Statistical power is defined as the probability to detect a certain, true effect. In the context of SEM, the magnitude of effect refers to the degree of discrepancy between two models in terms of model misfit associated with the more restrictive model in comparison to the more general model, such as the discrepancy between the hypothesized model and the perfectly fitting saturated model. If the more restrictive model does not differ in terms of fit from the more general model, there is no discrepancy in the population and the effect is zero. If the specified model differs in terms of fit from the more general model, the effect is larger than zero.

In SEM, there are a number of different indices that can be used to quantify the effect (for an overview, see West et al., 2012). For the purpose of power analyses, the choice of the particular index measuring the effect is rather arbitrary. Indeed, any fit index can be used as an effect size measure, provided that it allows for a (straightforward) computation of the noncentrality parameter. One index meeting this requirement is the minimum of the fit function, F_0, itself, because it immediately determines the noncentrality parameter (see Equation 3). However, F_0 has a quite unintuitive scaling and is therefore rarely used as an effect size measure. To arrive at a scaling that is easier to interpret, McDonald (1989) proposed to transform F_0 by

$$Mc_0 = exp(-0.5 \cdot F_0), \tag{4}$$

so that the resulting McDonald's noncentrality index (Mc) ranges from 0 to 1. Unlike F_0, lower values for the Mc indicate a larger discrepancy and thus a larger effect, whereas a value of 1 indicates a null effect. Using Mc, the noncentrality parameter can then be computed as $\lambda = (N - 1) \cdot (-2\ln(Mc))$.

Another approach to rescale F_0 is the root mean square error of approximation (RMSEA; Steiger, 2016) adjusting F_0 by the *df* of the model leading to

$$RMSEA_0 = \sqrt{F_0/df}. \tag{5}$$

The RMSEA thus measures misfit as discrepancy per *df*, with larger values indicating a larger discrepancy. If two models exhibit the same extent of misspecification in terms of F_0, the RMSEA would favor the more parsimonious model, because this model is associated with more *df* and thus lower RMSEA values. Based on the RMSEA, the noncentrality parameter is given by $\lambda = (N - 1) \cdot RMSEA^2 \cdot df$.

Having chosen a certain measure of effect, the next decision is to define a magnitude of effect that is deemed relevant for a particular research question. Usually, the magnitude of effect is defined in a way that reflects a meaningful discrepancy between two models. One way to determine a meaningful magnitude of an effect is to rely on prior research or to base this decision on theoretical assumptions. Concerning the comparison of a hypothesized against the perfectly fitting saturated model, certain rules of thumb are also often applied (such as an RMSEA of about .050; e.g., Browne & Cudeck, 1992; Hu & Bentler, 1999). However, it is important to note that these guidelines are not universally valid, because effect size measures are also influenced by factors unrelated to the effect of interest, such as loading magnitude (Moshagen & Auerswald, 2018). For this reason, such rules of thumb should be applied with caution and alternative approaches to determine a meaningful effect should also be considered. For example, McNeish and Wolf (2020) suggested to determine the relevant magnitude of effect dynamically depending on the properties of the model under scrutiny and the underlying data. This approach uses Monte Carlo simulations estimating a model that slightly differs from the hypothesized model as well as a model that equals the hypothesized model to obtain two distributions of a particular fit index, one for the slightly differing model and one for the model that equals the hypothesized model. The percentiles of both distributions are then used to determine a meaningful effect. In this context, it should also be noted that discrepancy can originate from the measurement model (i.e., the part of the model that relates the latent factors to the manifest indicator variables) or from the structural model (i.e., the part of the model that defines relations among the latent factors), and that these sources of misfit are usually conflated in any effect size measure (but see, Moshagen & Auerswald, 2018; Williams & O'Boyle, 2011). A further way to define a meaningful discrepancy thus involves expressing the discrepancy in terms of particular values for the model parameters (e.g., Muthén & Muthén, 2002), as we will illustrate further below.

Performing Power Analyses

After the effect of interest is specified – for instance, in terms of F_0, Mc_0, or RMSEA_0 – various types of power analyses depending on the particular research question can be performed. *A priori power analyses* are used to answer the question which sample size is required to detect a certain effect with a certain power. The required sample size is computed based on the specified α and β error probabilities, the magnitude of effect, as well as the model *df*. A priori power analyses indicate how many observations are needed to reject a factually false null hypothesis – as specified by the chosen effect – with a particular probability defined by $1 - \beta$.

In contrast, *post hoc power analyses* are performed to quantify the actually achieved power of a test. The power is computed based on the specified α error probability, the sample size, the magnitude of effect, and the model *df*. Thus, post hoc power analyses indicate the probability $1 - \beta$ to reject a false null hypothesis (as specified by the chosen effect).

A third type of power analyses are *compromise power* analyses (Faul et al., 2007; Moshagen & Erdfelder, 2016) answering the question which T_c value should be used to balance α and β error probabilities. This type of analysis uses the sample size, the magnitude of effect, and the desired ratio between α and β error probabilities (e.g., $\alpha/\beta = 1$ for equal error probabilities). Compromise power analyses determine the critical value T_c that leads to the desired ratio between α and β error probabilities, thereby providing a decision rule concerning whether the model is rather aligned with the null hypothesis (i.e., the model fits perfectly) or the alternative hypothesis (i.e., unacceptable extent of misfit as specified by the effect) based on proportional α and β error probabilities.

Tutorial

In the following, we demonstrate how different types of power analyses for various types of hypotheses in SEM can be performed using the R package semPower (Moshagen & Erdfelder, 2016). A subset of the illustrated analyses can also be performed using the web frontend of semPower available online at https://sempower.shinyapps.io/sempower. However, some analyses require the specification of the effect in terms of model-implied variance-covariance matrices, so we describe how to use semPower within the R environment (R Core

Team, 2020) in conjunction with a suitable SEM package such as lavaan (Rosseel, 2012).[3] Both packages are available from CRAN or from github at https://github.com/moshagen/semPower and https://github.com/yrosseel/lavaan, respectively. Before starting, the user should ensure that both packages are installed and activated via the `install.packages` and `library` commands, respectively. All R scripts for this tutorial are available in the Open Science Framework (OSF) at https://osf.io/4dwmv/?view_only=4c2c72e4003b4d9584635abdcb0785ed.

Based on an illustrative example model, we show how to perform power analyses for (a) global hypothesis testing regarding the comparison of a hypothesized model against the saturated model, (b) global hypothesis testing regarding the comparison of a hypothesized model against an incorrect but not explicitly specified model, (c) global hypothesis testing regarding the comparison of two explicitly specified models, (d) local hypothesis testing regarding a particular model parameter, and (e) both global and local hypothesis testing in the context of multi-group analyses. Concerning (a), the three different types of power analyses – a priori, post hoc, and compromise power analyses – are explained in detail. To avoid redundancies, we only illustrate a priori power analyses for the remaining applications, as the examples can be easily transferred to conduct post hoc and compromise power analyses. In what follows, we arbitrarily defined the relevant magnitude of effect to exemplify how to perform power analyses. Substantive researchers, however, should choose a magnitude of effect that represents a meaningful effect based on their particular research question.

Example Model

Throughout this tutorial, we rely on a substantive example to illustrate the different types and applications of power analyses.[4] The example model is based on a study by

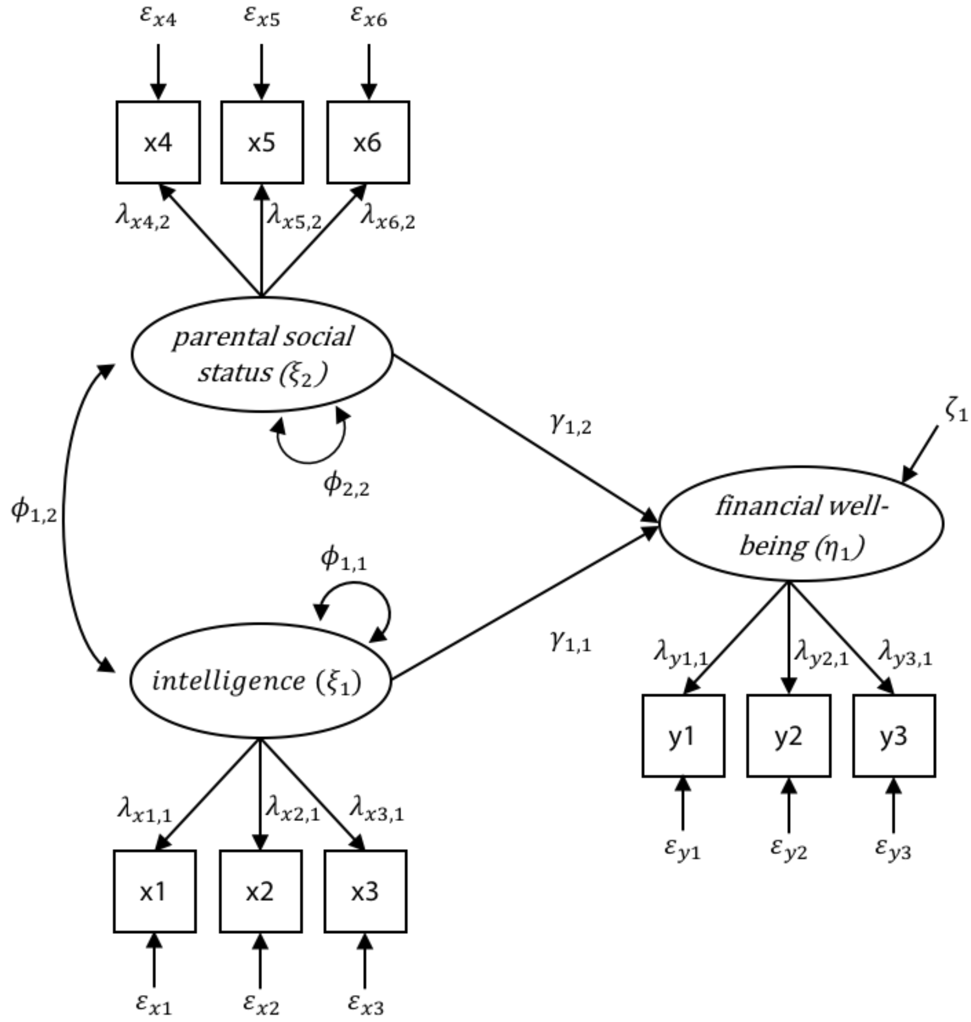
[3] The analyses of the present article were performed using R version 3.6.3, semPower version 1.1.0, and lavaan version 0.6-6.

[4] The general form of structural equation models can be found in the Appendix.

Furnham and Cheng (2017) investigating the influence of childhood intelligence and parental social status on adult financial well-being. For illustration purposes, we simplified their model so that it comprises three latent factors with three indicator variables each, leading to a total of $p = 9$ indicator variables and thus $0.5 \cdot (9 \cdot (9 + 1)) = 45$ unique elements in the variance-covariance matrix of the empirical data. As can be seen in Figure 2, the latent factor “financial well-being” is predicted by the latent factors “intelligence” and “parental social status”. In total, the model estimates 21 free parameters: 6 loadings (because the unstandardized loadings of three indicator variables are fixed to 1 to identify the model), 2 variances and 1 covariance of the exogenous factors “intelligence” and “parental social status”, 10 error terms (9 errors of indicator variables and 1 error of the endogenous factor “financial well-being”), as well as 2 slope parameters. This results in $45 - 21 = 24$ *df* of the LRT statistic for this model.

Figure 2

Exemplary Structural Equation Model Used to Perform Power Analyses



Note. The model comprises three latent factors (in ellipses) and nine indicator variables (in boxes). The two latent exogenous factors "intelligence" (ξ_1) and "parental social status" (ξ_2) predict the endogenous latent factor "financial well-being" (η_1). This relation is modeled by slope parameters (γ). All latent factors are indicated by three manifest indicators each (x_1 to x_6 for the exogenous factors and y_1 to y_3 for the endogenous factor) and the strength of this relation is indicated by the loadings (λ). Each predicted variable is associated with an error term, which is denoted by ε for errors of manifest indicator variables and by ζ for errors of latent factors. The variances ($\phi_{1,1}$ and $\phi_{2,2}$) and covariance ($\phi_{1,2}$) of the two exogenous latent factors are also considered in the model.

Power Analyses for Global Hypothesis Testing

Global hypothesis tests are performed when the overall structure of a model rather than a particular model parameter is of interest. Different types of hypotheses arise depending on the comparison model. If the comparison model is the perfectly fitting saturated model, power refers to the probability to detect whether the hypothesized model is misspecified (i.e., whether or not it describes the data perfectly). Rather than asking whether a model is perfect, it is sometimes of interest to compare the hypothesized model against another actually misspecified model (e.g., a comparison model showing an RMSEA of .040). In this case, it is not required to specify the comparison model explicitly, and power refers to the probability to differentiate the more parsimonious but poorer fitting model from the more general but better fitting model. Finally, global hypothesis tests also arise when comparing two explicitly specified models that differ by multiple parameters, so power refers to the probability to detect whether the more restrictive model fits the data significantly worse than the more general model.

Power Analyses for Global Hypothesis Testing Between the Saturated and a Hypothesized Model

A Priori Power Analysis. A priori power analyses for global hypothesis testing are used to determine the sample size that is required to detect a certain degree of misspecification of a model (e.g., a model misfit corresponding to RMSEA ≥ .050). Note again that whereas we rely on the RMSEA as an effect size measure of model misspecification, any noncentrality based fit index (such as F_0 or Mc; for a number of alternative indices, see West et al., 2012) can be used to quantify the discrepancy between the saturated and a hypothesized model. Performing an a priori power analysis requires to specify the magnitude of effect, the effect size measure, the α and β error probabilities, as well as the *df* of the model. Let us assume that we are interested in detecting an RMSEA of at least .050

with a power of 80.00% (so that $\beta = .20$) based on an α error probability of .05. The *df* of the test statistic depend on the hypothesized model; in the example model, the *df* equal 24. The required sample size to detect this particular effect for the hypothesized model can be determined via the command `semPower.aPriori` in the semPower package:

```
apriori <- semPower.aPriori(effect = .050, effect.measure =
     'RMSEA', alpha = .05, beta = .20, df = 24)
summary(apriori)
```

The magnitude of misfit is defined via the argument `effect`, `'RMSEA'` denotes the effect size measure, `alpha` and `beta` represent the respective error probabilities and `df` defines the degrees of freedom of the test statistic. Note that semPower allows for specifying either the β error through the `beta` argument or statistical power through the `power` argument. The results are stored in a variable named `apriori`. The following output is obtained by applying the `summary` command on the result variable:

```
semPower: A-priori power analysis

 F0                          0.060000
 RMSEA                       0.050000
 Mc                          0.970446

 df                          24
 Required Num Observations  375

 Critical Chi-Square         36.41502
 NCP                         22.44000
 Alpha                       0.050000
 Beta                        0.201086
 Power (1-beta)              0.798914
 Implied Alpha/Beta Ratio    0.248650
```

As can be seen, $N = 375$ observations are required to obtain a power of approximately 80.00% to detect a model misspecification corresponding to an RMSEA of at least .050. The output also shows the values of other effect size measures (F_0 and Mc) corresponding to the specified value of the RMSEA of .050. In addition, the output gives the critical chi-square value T_c, the noncentrality parameter (NCP), as well as the ratio between α and β error probabilities. Here, the implied α/β ratio is .25, meaning that it is four times as likely to commit a β error compared to an α error, which is a necessary consequence of specifying an α error probability of .05 and a β error probability of .20.

Post Hoc Power Analysis. The actually achieved power of a test to identify a certain effect (e.g., RMSEA ≥ .050) with a given sample size (e.g., *N* = 500) can be determined by post hoc power analyses. The sample size used, the effect of interest, the α error probability, as well as the *df* of the LRT statistic are required to compute the statistical power of the performed test. Suppose we are interested in determining the achieved power to detect an RMSEA ≥ .050 for the example model with 24 *df* based on an α error probability of .05 with a given sample of *N* = 500. This type of power analysis can be performed using the command `semPower.postHoc`.

```
posthoc <- semPower.postHoc(effect = .050, effect.measure =
     'RMSEA', alpha = .05, N = 500, df = 24)
summary(posthoc)
```

```
semPower: Post-hoc power analysis

 F0                          0.060000
 RMSEA                       0.050000
 Mc                          0.970446

 df                          24
 Num Observations            500
 NCP                         29.94000

 Critical Chi-Square         36.41502
 Alpha                       0.050000
 Beta                        0.075662
 Power (1-beta)              0.924338
 Implied Alpha/Beta Ratio 0.660835
```

The output provides the same information as in an a priori power analysis. The line `Power(1-beta)` gives the probability to detect an effect quantified by an RMSEA of at least .050 based on the above described scenario. That is, the probability to falsify our model when it is actually wrong (at least to an extent corresponding to RMSEA ≥ .050) is 92.43%

for a given sample size of $N = 500$. This implies that the β error probability is about 1.5 times as large as the α error probability (α/β = .66).

Compromise Power Analysis. Compromise power analyses are a useful tool if it is desired to identify a critical chi-square value T_c that balances the probabilities of committing an α error and a β error in deciding whether to retain or to reject the null hypothesis (Faul et al., 2007; Moshagen & Erdfelder, 2016). Performing a compromise power analysis requires the specification of an effect size measure and the magnitude of this effect (such as an RMSEA ≥ .100), the *df* of the LRT statistic, and, crucially, the desired ratio between α and β error probabilities (e.g., $\alpha/\beta = 1$ for equal probabilities). For example, let us perform a compromise power analysis to obtain equal error probabilities ($\alpha = \beta$) for the example model with 24 *df* in a sample of $N = 500$. Because compromise power analysis provides a criterion to decide whether the model is rather aligned with the hypothesis of perfect fit or with the hypothesis of an inacceptable degree of misspecification, we define the magnitude of effect as RMSEA = .100, which is often deemed an inacceptable extent of misfit (Browne & Cudeck, 1992). The analysis can be performed via the command `semPower.compromise` by specifying the desired ratio between α and β errors via the `abratio` argument.

```
compromise <- semPower.compromise(effect = .100,
     effect.measure = 'RMSEA', abratio = 1, N = 500, df = 24)
summary(compromise)
```

```
semPower: Compromise power analysis

 F0                          0.240000
 RMSEA                       0.100000
 Mc                          0.886920

 df                          24
 Num Observations            500
 Desired Alpha/Beta Ratio    1.000000

 Critical Chi-Square         64.18626
 Implied Alpha               0.000016
 Implied Beta                0.000016
 Implied Power (1-beta)   0.999984
 Actual Alpha/Beta Ratio  1.000000
```

The output indicates that a critical chi-square value of $T_c = 64.19$ is associated with balanced error probabilities in the above described scenario. If the observed chi-square value equals or exceeds this critical value, this would support the conclusion that the hypothesized model is rather aligned with the alternative hypothesis of an inacceptable degree of misfit, whereas smaller values would support the conclusion that the hypothesized model is rather aligned with the null hypothesis of perfect fit. By using this critical value, the probability of committing an α error equals the probability of committing a β error and amounts to < 0.01%.

A Priori Power Analyses for Global Hypothesis Testing Between Two Misspecified but Not Explicitly Specified Models

Rather than asking whether the hypothesized model differs from the saturated and thus perfectly fitting model, one might also ask whether the hypothesized model differs from another overidentified but not directly specified model. For example, suppose that one is interested in obtaining sufficient power to discriminate a more restrictive model associated with an RMSEA of .060 from a more general model associated with an RMSEA of .040. Note that one can often rely on prior research to identify plausible values for the RMSEA (or any other effect size measure). If no prior evidence is available, one needs to make a reasonable guess based on theoretical assumptions.

The effect of interest is now given by the discrepancy between these models in terms of the difference of the associated RMSEA values. As described in greater detail in MacCallum et al. (2006), performing power analyses for differences in the RMSEA also depends on the particular values of the individual RMSEAs, so a difference of ΔRMSEA = .020 translates into different effects in terms of F_0:

$$\Delta F_0 = (df_1 \cdot RMSEA_1^2) - (df_2 \cdot RMSEA_2^2). \tag{6}$$

Let us assume that we are interested in the required sample size to distinguish a model associated with RMSEA = .060 from a model associated with RMSEA = .040. In the running example, this might reflect a single-factor representation with 27 *df* and a three-factor representation with 24 *df*. We first compute the implied effect in terms of F_0 as per Equation 6 and then plug the obtained value along with the difference[5] in the *df* in the `semPower.aPriori` command:

```
RMSEA1 <- 0.06
RMSEA2 <- 0.04
df1 <- 27
df2 <- 24
deltaF <- df1*RMSEA1^2 - df2*RMSEA2^2

apriori <- semPower.aPriori(effect = deltaF, effect.measure =
     "F0", alpha = .05, beta = .20, df = df1 - df2)
summary(apriori)
```

```
semPower: A-priori power analysis

F0                          0.058800
RMSEA                       0.140000
Mc                          0.971028

df                          3
Required Num Observations   186

Critical Chi-Square         7.814728
NCP                         10.87800
Alpha                       0.050000
Beta                        0.200987
Power (1-beta)              0.799013
Implied Alpha/Beta Ratio    0.248772
```

[5] Update: the current version of semPower (1.3.0) allows for an automatic computation of the difference in the effect sizes as well as in the *df*. The present example can now directly be computed via

```
apriori <- semPower.aPriori(effect = c(0.06, 0.04), effect.measure =
"RMSEA", alpha = .05, beta = .20, df = c(27, 24))
summary(apriori)
```

The output indicates that $N = 186$ yields the desired power of approximately 80.00% on an alpha level of .05 to distinguish a model with $df_1 = 27$ and $\text{RMSEA}_1 \geq .060$ from a model with $df_2 = 24$ and $\text{RMSEA}_2 \leq .040$.

A Priori Power Analyses for Global Hypothesis Testing Between Two Explicitly Specified Models

As depicted in Figure 2, the example model comprises three latent factors. Although somewhat odd from a theoretical point of view, for the sake of illustration, let us assume we want to compare a correlated three-factor model (one factor each for "parental social status", "intelligence", and "financial well-being") against an alternative model specifying just a single factor affecting all indicators. Suppose that we want sufficient power to reject a single-factor model when the mutual correlations between the three factors factually are $r \leq .90$. Thus, the discrepancy between the single-factor model and a three-factor model with factor intercorrelations of $r = .90$ defines the magnitude of effect. However, it is not easily possible to say in advance how this effect translates into particular values of a certain effect size measure such as F_0 or RMSEA. Instead, we need to obtain the variance-covariance matrices implied by the three-factor and the single-factor model.

To this end, we first define the more general three-factor model (involving a correlation of .90 between all factors) as a population model to obtain the population variance-covariance matrix. Then, we fit the nested, more restrictive single-factor model to the population variance-covariance matrix and obtain the associated model implied variance-covariance matrix of the nested model. We rely on the R package lavaan (Rosseel, 2012), but any other suitable SEM software can be employed as well.

First, we need to define one model representing the true relations in the population. Given that the magnitude of effect (and thus power) depends on both the correlation between the factors and the values of other parameters in the model (e.g., Moshagen & Auerswald, 2018), one needs to specify all model parameters based on existing knowledge or some reasonable guess. The definition of the population model can be written as follows, complying with the syntax requirements of lavaan (see the lavaan documentation for details):

```
population.model <- '
# define factors and set all loadings to .7
     intelligence =~ .7*x1 + .7*x2 + .7*x3
     social.status =~ .7*x4 + .7*x5 + .7*x6
     well.being =~ .7*y1 + .7*y2 + .7*y3
# set all item residual variances to .51 (= 1 - .7*.7)
     x1 ~~ .51*x1
     x2 ~~ .51*x2
     x3 ~~ .51*x3
     x4 ~~ .51*x4
     x5 ~~ .51*x5
     x6 ~~ .51*x6
     y1 ~~ .51*y1
     y2 ~~ .51*y2
     y3 ~~ .51*y3
# set factor variances to 1
     intelligence ~~ 1*intelligence
     social.status ~~ 1*social.status
     well.being ~~ 1*well.being
# set factor covariances to .9
     intelligence ~~ .9*social.status
     intelligence ~~ .9*well.being
     well.being ~~ .9*social.status'
```

Here, *population.model* indicates the name of the model. It is defined by (<-) a model string specifying three latent factors ("intelligence", "social status", and "well-being") that are measured by (=~) three manifest indicator variables each. For simplicity, we assume that all loadings are .70. We also fix all item residual variances (~~) to be equal to $1 - .70^2 = .51$, so that the loadings are in a standardized metric. We further fix the variances of the factors

(~~) to 1, so that the covariances between the latent factors (~~) are also standardized and thus reflect correlations. Importantly, we define the mutual correlations between all three factors to be equal to *r* = .90.

The variance-covariance matrix defined by this model is the population variance-covariance matrix. This matrix can be obtained using the lavaan package via

```
cov.population.model <- fitted(sem(population.model))$cov
```

As an optional step, one might want to verify that all parameters have been appropriately defined in the population model by simply fitting the intended model to the obtained population variance-covariance matrix and checking whether the estimates correspond to the specifications in the population model.

```
myModel <- '
     intelligence =~ x1 + x2 + x3
     social.status =~ x4 + x5 + x6
     well.being =~ y1 + y2 + y3'

fit.myModel <- sem(myModel, sample.cov = cov.population.model,
     sample.nobs = 1000, sample.cov.rescale = FALSE)
summary(fit.myModel, standardized = TRUE)
df.myModel <- fit.myModel@test[[1]]$df
```

The lavaan command `sem` fits the model *myModel* to the population variance-covariance matrix (*cov.population.model*) as supplied via the argument `sample.cov`. The number of observations specified in `sample.nobs` is arbitrary in the present context, because the population variance-covariance matrix is independent of sample size. Moreover, we suppress a rescaling of the sample variance-covariance matrix via the argument `sample.cov.rescale = FALSE`, because we do not want lavaan to multiply the input variance-covariance matrix *cov.population.model* by (*N*-1)/*N*. By fitting *myModel*, one can verify that the model replicates the parameter values used to define the population model.

Moreover, as *myModel* represents the intended relations in the population, its *df* can be used to compute the difference in *df* between the population model and the hypothesized model. We save the information about the *df* in *df.myModel* by using the @ operator, which extracts a particular content of an object (such as the *df* of the model).

In a next step, we define the nested model *hypothesized.model* that specifies only a single latent factor and fit the analysis model to the population variance-covariance matrix obtained above.

```
hypothesized.model <- '
     single.factor =~ x1 + x2 + x3 + x4 + x5 + x6 + y1 + y2 +
     y3'

fit.hypothesized.model <- sem(hypothesized.model, sample.cov =
    cov.population.model, sample.nobs = 1000,
    sample.cov.rescale = FALSE)
```

After fitting *hypothesized.model*, the *df* as well as the model-implied variance-covariance matrix are stored and then plugged into an a priori power analysis requesting the required sample size to achieve a power of 80.00%.

```
df.hypothesized.model <- fit.hypothesized.model@test[[1]]$df
cov.hypothesized.model <- fitted(fit.hypothesized.model)$cov

apriori <- semPower.aPriori(SigmaHat = cov.hypothesized.model,
     Sigma = cov.population.model, alpha = .05, power = .80,
     df = (df.hypothesized.model - df.myModel))
summary(apriori)
```

The underlying effect is quantified by the discrepancy between the model-implied variance-covariance matrix of *hypothesized.model* and the population variance-covariance matrix obtained in *population.model*. If one is interested in examining both matrices in greater detail, one can just run the code `cov.hypothesized.model` and `cov.population.model` to generate the respective output. The command `semPower.aPriori` uses the argument `SigmaHat` for the model-implied variance-covariance matrix and the argument `Sigma` for the population variance-covariance matrix. This yields the output:

```
semPower: A-priori power analysis

F0                          0.050101
RMSEA                       0.129230
SRMR                        0.018978
Mc                          0.975261
GFI                         0.988989
AGFI                        0.834836
CFI                         0.985038

df                          3
Required Num Observations   219

Critical Chi-Square         7.814728
NCP                         10.92199
Alpha                       0.050000
Beta                        0.199222
Power (1-beta)              0.800778
Implied Alpha/Beta Ratio    0.250977
```

The output now also reports a number of additional fit indices that can be computed based on the variance-covariance matrices. For example, the degree of discrepancy between the single-factor model and the three-factor model with factor intercorrelations of .90 corresponds to RMSEA = .129 or SRMR = .019. Importantly, the output also shows that given an α error probability of .05, a sample size of *N* = 219 is required to reject the single-factor model with a probability of 80.08% when there are actually three factors that correlate by $r \leq .90$.

Power Analyses for Local Hypothesis Testing

Beyond global hypothesis testing (i.e., assessing the overall fit of a model), it is often desired to perform power analyses for local hypotheses regarding a single or a few particular model parameter(s), such as loadings or slope parameters. With respect to the example model, this type of power analysis can for instance be used to determine the sample size that is required to detect a particular difference in the regression slope modeling the relation between "intelligence" and "financial well-being".

Let us assume that the slope parameter between said factors is .30 in the population, and – for simplicity – that the exogenous factors are uncorrelated. We want to achieve a power of .90 based on an α error probability of .05 to reject the factually false null hypothesis that the slope parameter equals zero. Again, it is not straightforward to translate this effect into a particular value of an effect size measure. For this reason, we specify the variance-covariance matrices associated with the model assuming a regression slope larger than zero and the model assuming a slope equal to zero in the same way as described in greater detail in the previous section. To obtain a variance-covariance matrix that matches the data perfectly, we define *population.model* as the population model and define the "true" loadings and regression slopes. Then, we specify the model *hypothesized.model* representing the incorrect null hypothesis of no relation between intelligence and well-being and fit it to the population variance-covariance matrix.

```
population.model <- '
# define factors and set all loadings to .7
     intelligence =~ .7*x1 + .7*x2 + .7*x3
     social.status =~ .7*x4 + .7*x5 + .7*x6
     well.being =~ .7*y1 + .7*y2 + .7*y3
# set all item residual variances to .51 (= 1 - .7*.7)
     x1 ~~ .51*x1
     x2 ~~ .51*x2
     x3 ~~ .51*x3
     x4 ~~ .51*x4
     x5 ~~ .51*x5
     x6 ~~ .51*x6
     y1 ~~ .51*y1
     y2 ~~ .51*y2
     y3 ~~ .51*y3
# set factor variances to 1
     intelligence ~~ 1*intelligence
     social.status ~~ 1*social.status
```

```
# define orthogonal exogenous factors
     intelligence ~~ 0*social.status
# define regression relationship
     well.being ~ .3*intelligence + .6*social.status
# define residual variance of well.being (= 1 - (.3^2 + .6^2))
     well.being ~~ .55*well.being'

hypothesized.model <- '
     intelligence =~ x1 + x2 + x3
     social.status =~ x4 + x5 + x6
     well.being =~ y1 + y2 + y3
# fix slope of intelligence to zero
     well.being ~ 0*intelligence + social.status'

cov.population.model <- fitted(sem(population.model))$cov

fit <- sem(hypothesized.model, sample.cov =
     cov.population.model, sample.nobs = 1000,
     sample.cov.rescale = FALSE)
```

After fitting *hypothesized.model* to the variance-covariance matrix implied by *population.model*, we extract the model-implied variance-covariance matrix of *hypothesized.model*. To answer the question which sample size is required to detect the specified difference in the slope parameter with a probability of .90 and an α error probability of .05, we use the `semPower.aPriori` command, defining the model-implied variance-covariance matrix as `SigmaHat` and the population variance-covariance matrix as `Sigma`. In addition, there is now just a single *df* associated with this hypothesis, because the nested model differs from the more general model in only a single parameter (namely the omitted regression slope).

```
cov.hypothesized.model <- fitted(fit)$cov

apriori <- semPower.aPriori(SigmaHat = cov.hypothesized.model,
     Sigma = cov.population.model, alpha = .05, power = .90,
     df = 1)
summary(apriori)
```

The command leads to the following output:

```
semPower: A-priori power analysis

F0                          0.062066
RMSEA                       0.249130
SRMR                        0.059513
Mc                          0.969444
GFI                         0.986395
AGFI                        0.387784
CFI                         0.972728

df                          1
Required Num Observations   170

Critical Chi-Square         3.841459
NCP                         10.48915
Alpha                       0.050000
Beta                        0.100496
Power (1-beta)              0.899504
Implied Alpha/Beta Ratio    0.497534
```

The output reveals that fitting *hypothesized.model* assuming a slope parameter of zero for the regression of well-being on intelligence when the true relation in the population is .30, leads to a minimum of the fit function of 0.062. A sample size of $N = 170$ is required to achieve a power of .90 to detect this effect based on an α error probability of .05.

Illustrations of Power Analyses in Multiple Group Settings

SEM is often used in a multiple group context, where a certain model is simultaneously fit to several groups, such as gender groups or different cultures. Common hypotheses in such contexts refer to cross-group comparisons of the model parameters. A typical example is the assessment of measurement invariance to evaluate whether and to which extent latent constructs can be compared across groups. Measurement invariance can be tested based on a multi-group confirmatory factor analysis (MGCFA) comparing a sequence of increasingly restrictive, nested factor models (for details, see Meredith, 1993). For example, the test of metric invariance compares two models, one model with freely estimated loadings per group (i.e., configural invariance model) and one with loadings constrained to be equal across groups (i.e., metric invariance model). If the LRT does not indicate a significant difference between both models, one can assume that the model with equal loadings is not significantly worse than the unrestricted model and, in turn, that the loadings do not significantly differ across groups, so that metric invariance is supported.

Concerning power analyses, a relevant question is the required sample size to identify a lack of measurement invariance. Given that these hypotheses involve constraints on multiple parameters at once (e.g., all loadings), we treat this scenario as an instance of global hypothesis testing in a multiple group context. Of course, local hypothesis tests regarding comparisons of a single parameter occur in multiple group contexts as well, for example when investigating whether two constructs relate equally strong in several groups. We illustrate global as well as local hypothesis tests in multi-group settings based on a model similar to our running example (see Figure 2) but with orthogonal exogenous factors to simplify the presentation.

A Priori Power Analyses for Global Hypothesis Testing in Multiple Group Models

When comparing the fit of a configural invariance model against the more restrictive metric invariance model, a deterioration in fit is expected due to the applied constraints on the loadings. Cheung and Rensvold (2002) suggested that a decline of the Mc associated with the metric invariance model as compared to the configural invariance model of ≥ .02 indicates a meaningful departure from invariance. As mentioned above, different rationales exist to determine a meaningful effect. Here we rely on the recommendations of Cheung and Rensvold (2002) for illustration purposes, but alternative suggestions of meaningful effects in the context of measurement invariance testing can also be applied (e.g., Chen, 2007; Jorgensen et al., 2017; Meade et al., 2008).

Let us now assume that we want to obtain the required sample size to detect a difference of ΔMc ≥ .02 between the configural and the metric invariance model in a two-group scenario (say, in the comparison of gender groups). Because the effect size in terms of F_0 does not only depend on ΔMc but also on the particular Mc values, the Mc for all groups needs to be supplied in the `effect`[6] argument of the `semPower.aPriori` command. Beyond the desired α and β error probabilities, the difference in *df* between the configural and the metric invariance model needs to be provided. In the present case, there are 6 *df*, as 6 out of 9 loadings each are freely estimated in the configural model but constrained in the metric invariance model. The *df* of the configural invariance model and the metric invariance model can be determined via `semPower.getDf(model, nGroups = 2)` and `semPower.getDf(model, nGroups = 2, group.equal = c('loadings'))`, respectively,[7] where `model` represents the name of the specified model, `nGroups` indicates the number of compared groups, and `group.equal = c('loadings')` imposes the equality constraints, i.e., we assume equal loadings across groups. Finally, power analyses involving multiple groups require the specification of group sizes in terms of a list of weights via the `N` argument. By using `N = list(1,1)`, we request equally sized groups.

[6] Please note that there is a syntax error in the published version of the manuscript. The correct specification of the `effect` argument is `c(effect1, effect2)`. In the published version it is erroneously specified via `list(effect1, effect2)`. Additionally, if you use an older version of semPower (i.e., not 1.3.0), please directly define the difference in *df* in the `df` argument (e.g., `df = 6`).

[7] Please note that this function is only available in the current semPower version (1.3.0).

```
model<- '
intelligence =~ x1 + x2 + x3
social.status =~ x4 + x5 + x6
well.being =~ y1 + y2 + y3'

mc1 <- 0.99
mc2 <- 0.97
df1<- semPower.getDf(model, nGroups = 2)
df2<- semPower.getDf(model, nGroups = 2, group.equal =
c('loadings'))

apriori <- semPower.aPriori(effect = c(mc1, mc2),
     effect.measure = "Mc", alpha = .05, beta = .20, df =
     c(df1, df2), N = list(1,1))
summary(apriori)
```

```
semPower: A-priori power analysis

F0                          0.040818
RMSEA                       0.116644
Mc                          0.979798

df                          6
Required Num Observations 336
                            (168, 168)

Critical Chi-Square         12.59158
NCP                         13.63312
Alpha                       0.050000
Beta                        0.199697
Power (1-beta)              0.800303
Implied Alpha/Beta Ratio    0.250379
```

As can be seen, a total sample size of *N* = 336 (168 per group) is required to identify a difference between two models with Mc ≥ .99 versus Mc ≤ .97 with a power of approximately 80.00% on α = .05. Note that the effect sizes provided in the output refer to a model that is simultaneously fit to all groups.

A Priori Power Analyses for Local Hypothesis Testing in Multiple Group Models

After measurement invariance concerning the relation between latent variables and the manifest indicators has been established, it is common to compare the groups with respect to structural differences regarding the modeled constructs. In the running example, one might be interested whether there is a group difference in the strength of the regression slope linking the factors “intelligence” and “financial well-being”. The magnitude of the group difference defines the effect of interest and is to be determined by obtaining the discrepancy of a more general model that freely estimates the regression slopes in both groups versus a more restrictive model that restricts the regression slope to be equal across groups.

As described above, determining the magnitude of effect when comparing explicitly specified models usually requires obtaining the associated variance-covariance matrices, so we first need to define a population model to obtain the population variance-covariance matrix. In the context of multiple group analyses, the population models for all groups considered need to be defined, which we call *pop.groupA* and *pop.groupB*. In the population, the only difference between the groups pertains to the magnitude of the slope of the factor “intelligence”, defining the effect of interest. Let us assume that the slope is .30 in the first group, but .10 in the second group.

```
pop.groupA <- '
# define factors and set all loadings to .7
     intelligence =~ .7*x1 + .7*x2 + .7*x3
     social.status =~ .7*x4 + .7*x5 + .7*x6
     well.being =~ .7*y1 + .7*y2 + .7*y3
# set all item residual variances to .51 (= 1 - .7*.7)
     x1 ~~ .51*x1
     x2 ~~ .51*x2
     x3 ~~ .51*x3
     x4 ~~ .51*x4
     x5 ~~ .51*x5
     x6 ~~ .51*x6
     y1 ~~ .51*y1
     y2 ~~ .51*y2
     y3 ~~ .51*y3
# set factor variances to 1
     intelligence ~~ 1*intelligence
     social.status ~~ 1*social.status
# define orthogonal exogenous factors
     intelligence ~~ 0*social.status
# define regression relationship
     well.being ~ .3*intelligence + .6*social.status
# define residual variance of well.being (= 1 - (.3^2 + .6^2))
     well.being ~~ .55*well.being'

pop.groupB <- '
# define factors and set all loadings to .7
     intelligence =~ .7*x1 + .7*x2 + .7*x3
     social.status =~ .7*x4 + .7*x5 + .7*x6
     well.being =~ .7*y1 + .7*y2 + .7*y3
# set all item residual variances to .51 (= 1 - .7*.7)
     x1 ~~ .51*x1
     x2 ~~ .51*x2
     x3 ~~ .51*x3
     x4 ~~ .51*x4
     x5 ~~ .51*x5
     x6 ~~ .51*x6
     y1 ~~ .51*y1
     y2 ~~ .51*y2
     y3 ~~ .51*y3
# set factor variances to 1
     intelligence ~~ 1*intelligence
     social.status ~~ 1*social.status
# define orthogonal exogenous factors
     intelligence ~~ 0*social.status
```

```
# define regression relationship; note smaller slope for iq
     well.being ~ .1*intelligence + .6*social.status
# define residual variance of well.being (= 1 - (.1^2 + .6^2))
     well.being ~~ .63*well.being'
```

We now use the population variance-covariance matrices of both groups to fit the analysis model assuming strict invariance with equal latent variances and, crucially, equal slope parameters for “intelligence” across groups. Note again that the `sample.nobs` are arbitrary, because we fit the model to variance-covariance matrices.

```
cov.pop.groupA <- fitted(sem(pop.groupA))$cov
cov.pop.groupB <- fitted(sem(pop.groupB))$cov

analysis.model<- '
     intelligence =~ x1 + x2 + x3
     social.status =~ x4 + x5 + x6
     well.being =~ y1 + y2 + y3
# fix slope of intelligence to equality across groups
     well.being ~ c(s1,s1)*intelligence + social.status'

fit <- sem(analysis.model, sample.cov = list(cov.pop.groupA,
     cov.pop.groupB), sample.cov.rescale = FALSE, sample.nobs
     = list(1000,1000), group.equal = c('loadings',
     'residuals', 'lv.variances'))
```

The number of population variance-covariance matrices as well as model-implied variance-covariance matrices depends on the number of groups. In this example, we compare two groups, which is why two matrices (*cov.pop.groupA* and *cov.pop.groupB*) are used in the argument `sample.cov`. In a next step, we obtain the model-implied variance-covariance matrices per group.

```
cov.analysis.model.groupA <- fitted(fit)$'Group 1'$cov
cov.analysis.model.groupB <- fitted(fit)$'Group 2'$cov
```

Then, we conduct the a priori power analysis based on the desired α error probability of .05 and the β error probability of .20. The argument `SigmaHat` refers to the model-implied variance-covariance matrices, whereas the argument `Sigma` refers to the population variance-covariance matrices, both of which are passed as a list. Furthermore, as noted above, the group sizes have to be specified in terms of weights in a multi-group context. For illustration, let us assume that the first group is twice as large as the second group, leading to weights of 2:1.

```
apriori<- semPower.aPriori(SigmaHat =
      list(cov.analysis.model.groupA,
      cov.analysis.model.groupB), Sigma = list(cov.pop.groupA,
      cov.pop.groupB), alpha = .05, beta = .20, df = 1, N =
      list(2,1))
summary(apriori)
```

```
semPower: A-priori power analysis

F0                          0.007321
RMSEA                       0.121007
SRMR                        0.019849
Mc                          0.996346
GFI                         0.998376
AGFI                        0.926906
CFI                         0.996756

df                          1
Required Num Observations 1056
                            (704, 352)

Critical Chi-Square         3.841459
NCP                         7.716765
Alpha                       0.050000
Beta                        0.206694
Power (1-beta)              0.793306
Implied Alpha/Beta Ratio    0.241904
```

The output shows that 704 and 352 observations, respectively (i.e., a total sample size of *N* = 1,056) are required to achieve a power of approximately .80 on an α error probability of .05 to detect that the slope parameters of “intelligence” of ≤ .10 versus ≥ .30 differ across groups. It is not surprising that a rather large sample size is required to detect this effect, considering the rather small difference in the slope parameter across groups and that the effect of interest refers to only a single model parameter. This emphasizes the importance of proper sample size planning especially if small or medium-sized effects are expected.

Discussion and Conclusion

In the present article, we sketched the theoretical foundations and provided an in-depth tutorial on how to perform power analyses for various types of hypotheses commonly arising in SEM. We illustrated a priori, post hoc, and compromise power analyses for the test of global and local hypotheses. Additionally, power analyses for common research questions occurring in multiple group contexts were exemplified.

In general, the herein described analytic approach for determining power is based on asymptotic theory, that is, assuming that the model test statistic asymptotically follows an underlying central chi-square distribution under the null hypothesis and a noncentral chi-square distribution under the alternative hypothesis (e.g., MacCallum et al., 1996, 2006; Satorra & Saris, 1985). Certain conditions need to be satisfied for these distributional assumptions to hold, such as multivariate normal data in case of ML estimation and a sufficiently large sample size in relation to the number of observed variables (for details, see Browne & Cudeck, 1992; MacCallum et al., 1996; Steiger et al., 1985). If the data are actually not multivariate normally distributed or if the number of observed variables is large, the empirical chi-square values obtained using ML estimation are biased upwards (e.g., Curran et al., 1996; Fouladi, 2000; Moshagen, 2012), so that the actual distribution of the model test statistic differs from the asymptotically expected distribution (so that the empirical rejection rates exceed the analytically determined level of power). For this reason, we strongly recommend using appropriate corrections for the empirical test statistic – so-called robust test statistics (e.g., Asparouhov & Muthén, 2010; Lin & Bentler, 2012; Satorra & Bentler, 1994) – to yield unbiased test statistics in such situations.

Further, it is important to note that we only considered sample size requirements from the perspective of obtaining sufficient power to test a certain hypothesis of interest. Clearly, if an a priori power analysis indicates that, say, $N = 100$ yields a reasonable statistical power, this does not imply that 100 observations are sufficient to support parameter estimation of the model. Beyond power analyses based on asymptotic theory, it is also possible to perform power analyses in SEM based on Monte Carlo simulations (e.g., Muthén & Muthén, 2002; Wang & Rhemtulla, in press). An advantage of the simulation approach is that it does not only provide information about power but also about convergence and thus the required sample size to support proper estimation of the model. Furthermore, simulations allow for a flexible and reliable handling of violated assumptions, as for instance non-normally distributed data, as long as these violations are appropriately considered in the simulation process. However, the simulation approach has shortcomings regarding global hypothesis testing, as it is not possible to consider the effect quantified by an overall measure of fit such as the RMSEA. Instead, the misspecification has to be induced on the parameter level, as for example in terms of misspecified loading estimates. Moreover, conducting simulations requires a substantial level of expertise and, finally, the computing time can be very extensive if complex models with many variables are estimated.

In sum, we highlighted the importance of sample size planning and argued that α as well as β error probabilities should always be considered to enable reasonable hypothesis tests. We encourage researchers to yield reliable – and thus more replicable – results based on thoughtful sample size planning especially if small or medium-sized effects are expected. Indeed, any study should take statistical power into account when planning sample sizes, or – at the very least – should report the actual power achieved. To this end, we hope that this tutorial advances the use of power analyses in applications of SEM.

References

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Asparouhov, T., & Muthén, B. (2010). *Simple second order chi-square correction*. https://www.statmodel.com/download/WLSMV_new_chi21.pdf

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*(1), 1–24.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*(1), 62-83.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*(19). https://doi.org/10.1186/s13717-016-0063-3

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*, 356–410. https://doi.org/10.1207/S15328007SEM0703_2

Furnham, A., & Cheng, H. (2017). Socio-demographic indicators, intelligence, and locus of control as predictors of adult financial well-being. *Journal of Intelligence, 5*, 11. https://doi.org/10.3390/jintelligence5020011

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jorgensen, T. D., Kite, B. A., Chen, P-Y., & Short, S. D. (2017). Finally! A valid test of configural invariance using permutation in multigroup CFA. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 93–103). Springer. https://doi.org/10.1007/978-3-319-56294-0_9

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Sage.

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology, 9,* 2207–2230. https://doi.org/10.4236/psych.2018.98126

Lin, J., & Bentler, P. M. (2012). A third moment adjusted test statistic for small sample factor analysis. *Multivariate Behavioral Research*, *47*(3), 448–462. https://doi.org/10.1080/00273171.2012.673948

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226. https://doi.org/10.1146/annurev.psych.51.1.201

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*(1), 19–35. https://doi.org/10.1037/1082-989X.11.1.19

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis?: What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*, 97–103. https://doi.org/10.1007/BF01908590

McNeish, D., & Wolf, M. G. (2020). *Dynamic fit index cutoffs for confirmatory factor analysis models.* https://doi.org/10.31234/osf.io/v8yru

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(1), 86–98.

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods, 23*(2), 318–336. https://doi.org/10.1037/met0000122

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Olsson, U. H., Foss, T., & Breivik, E. (2004). Two equivalent discrepancy functions for maximum likelihood estimation: Do their test statistics follow a non-central chi-square distribution under model misspecification? *Sociological Methods and Research*, *32*(4), 453–500. https://doi.org/10.1177/0049124103258131

R Core Team (2020). R: A language and environment for statistical computing. Vienna, Austria. https://www.R-project.org

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*(1), 83–90. https://doi.org/10.1007/BF02294150

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Steiger, J. H. (2016). Notes on the Steiger–Lind (1980) handout. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 777–781. https://doi.org/10.1080/10705511.2016.1217487

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika*, *50*(3), 253–263. https://doi.org/10.1007/BF02294104

Wang, Y. A., & Rhemtulla, M. (in press). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. (pp. 209–231). Guilford Press.

Williams, L. J., & O'Boyle, E., Jr. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods, 14*(2), 350–369. https://doi.org/10.1177/1094428110391472

Appendix

Definition of the General Structural Equation Model

The relations of the indicator variables and the latent factors are defined by the measurement model, which is given by $\boldsymbol{x} = \boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\xi} + \boldsymbol{\varepsilon}_{\boldsymbol{x}}$ for indicators that are predicted by exogenous latent factors $\boldsymbol{\xi}$ (i.e., factors that are not predicted by other variables in the model) and by $\boldsymbol{y} = \boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta} + \boldsymbol{\varepsilon}_{\boldsymbol{y}}$ for indicators that are predicted by endogenous latent factors $\boldsymbol{\eta}$ (i.e., factors that are predicted by other variables in the model; for details, see Bollen, 1989). The vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, collect the indicator variables, matrix $\boldsymbol{\Lambda}$ contains the factor loadings and the residuals are collected in vector $\boldsymbol{\varepsilon}_{\boldsymbol{x}}$ and $\boldsymbol{\varepsilon}_{\boldsymbol{y}}$, respectively. The structural model defines the relations between exogenous and endogenous factors by $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$, where matrices $\mathbf{B}$ and $\boldsymbol{\Gamma}$ contain the regression weights of endogenous and exogenous factors, respectively, and the residuals of endogenous variables are stacked in vector $\boldsymbol{\zeta}$.

Referring to our running example (see Figure 2), the measurement model in matrix notation is given by

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} \lambda_{x1} & 0 \\ \lambda_{x2} & 0 \\ \lambda_{x3} & 0 \\ 0 & \lambda_{x4} \\ 0 & \lambda_{x5} \\ 0 & \lambda_{x6} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{x1} \\ \varepsilon_{x2} \\ \varepsilon_{x3} \\ \varepsilon_{x4} \\ \varepsilon_{x5} \\ \varepsilon_{x6} \end{pmatrix} \tag{7}$$

and by

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \lambda_{y1} \\ \lambda_{y2} \\ \lambda_{y3} \end{pmatrix} (\eta_1) + \begin{pmatrix} \varepsilon_{y1} \\ \varepsilon_{y2} \\ \varepsilon_{y3} \end{pmatrix} \tag{8}$$

for the manifest indicators of the exogenous latent factors “intelligence” (ξ_1) and “parental social status” (ξ_2) and for the manifest indicators of the endogenous latent factor “financial well-being” (η_1), respectively. The structural model of the running example in matrix notation is defined as

$$(\eta_1) = (0)(\eta_1) + (\gamma_{\xi_1,\eta_1} \quad \gamma_{\xi_2,\eta_1})\begin{pmatrix}\xi_1 \\ \xi_2\end{pmatrix} + (\zeta_{\eta_1}). \tag{9}$$

The slopes γ_{ξ_1,η_1} and γ_{ξ_2,η_1} of the regression of financial well-being on intelligence and parental social status are collected in matrix $\mathbf{\Gamma}$. As our running example models no regression between endogenous factors (as there is only a single endogenous factor), $\mathbf{B}$ equals zero. Under the assumption of independent factors and errors, the model-implied variance-covariance matrix based on the measurement and the structural model is given by

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix}\mathbf{\Sigma}(\boldsymbol{\theta})_{yy} & \mathbf{\Sigma}(\boldsymbol{\theta})_{yx} \\ \mathbf{\Sigma}(\boldsymbol{\theta})_{xy} & \mathbf{\Sigma}(\boldsymbol{\theta})_{xx}\end{pmatrix} =$$

$$\begin{pmatrix}\mathbf{\Lambda}_y(\boldsymbol{I} - \mathbf{B})^{-1}(\mathbf{\Gamma\Phi\Gamma}' + \mathbf{\Psi})[(\boldsymbol{I} - \mathbf{B})^{-1}]'\mathbf{\Lambda}'_y + \mathbf{\Theta}_{\boldsymbol{\varepsilon}_y} & \mathbf{\Lambda}_y(\boldsymbol{I} - \mathbf{B})^{-1}\mathbf{\Gamma\Phi\Lambda}'_x \\ \mathbf{\Lambda}_x\mathbf{\Phi}\Gamma'[(\boldsymbol{I} - \mathbf{B})^{-1}]'\mathbf{\Lambda}'_y & \mathbf{\Lambda}_x\mathbf{\Phi\Lambda}'_x + \mathbf{\Theta}_{\boldsymbol{\varepsilon}_x}\end{pmatrix} \tag{10}$$

with $\boldsymbol{I}$ as identity matrix, $\mathbf{\Phi}$ as the variance-covariance matrix of the exogenous factors, $\mathbf{\Psi}$ as the variance-covariance matrix of $\boldsymbol{\zeta}$, and $\mathbf{\Theta}$ as the variance-covariance matrix of the residuals in $\boldsymbol{\varepsilon}_{\boldsymbol{x}}$ and $\boldsymbol{\varepsilon}_{\boldsymbol{y}}$. Note that for factor analytic models $\mathbf{\Sigma}(\theta)$ equals $\mathbf{\Sigma}(\theta)_{xx}$ given by the lower diagonal element in Equation 10 (for details, see Bollen, 1989; Kaplan, 2009).