5

Sample Size Planning for Confirmatory Factor Models: Power and Accuracy for Effects of Interest

Ken Kelley and Keke Lai

Introduction

Confirmatory factor analysis (CFA) holds a special place in psychometrics because of the model's historical significance and widespread use when latent constructs are theorized and empirically evaluated. A special case of a CFA model is simply a composite measure.[1] Confirmatory factor models are themselves special cases of structural equation models. Within a structural equation model (SEM), the measurement model of a given construct is a confirmatory factor model. In an SEM, constructs are used to test a hypothesized structural relationship among latent variables. Thus, an SEM, the measurement model, itself a confirmatory factor model, is used in a larger framework of modeled constructs called the structural model. Although widely used in applied research, much of the writing on CFA is within an SEM framework treating CFA as a special case. In the introduction to Brown's book on CFA (2006), David Kenny says, "it is ironic that SEM has received so much more attention than CFA, because the social and behavioral sciences have learned much more from CFA than from SEM" (p. ix). With that strong endorsement of the importance of CFA, our focus in this chapter is sample size planning for CFA models. The ideas, however, are directly applicable to other related models, such as multiple regression, path analysis, and SEM among others. In fact, the ideas we discuss here are general and they are widely applicable to many contexts.

Consider the idea of sample size planning. First, in our experience, this is a topic that is often ignored by many researchers. We say this because so often the idea of planning an appropriate sample size is an afterthought to other aspects of a study. We believe one reason why sample size is sometimes an afterthought is because there are often a priori built-in limitations of a study, such as when there is only a short amount of time available to collect data, the number of potential participants is necessarily restricted (e.g., from a shared participant pool, from within an organization, from a special or limited

[1] For example, the sum of a set of items, with unit weights or unequal weights, can be conceptualized as a constrained CFA model.

The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, First Edition. Edited by Paul Irwing, Tom Booth, and David J. Hughes.

population), or financial resources to collect data dictate only a small number of participants' data can be collected. Any of these limitations may be a very practical reason why, regardless of what a formal planning of sample size would suggest, researchers use the sample size that is available to them, at least conveniently. In these sorts of situations, the idea of doing an analysis to choose an appropriate sample size does not seem to be highly valued. However, as we explain momentarily, we believe that even in such constrained situations it is still important to plan sample size and the reason many studies fail to accomplish their original goals is often, but certainly not always, directly related to sample size.

An alternative scenario to the one just considered is one in which sample size planning is explicitly and carefully considered, such as in many proposals for funding (i.e., grant applications). Many funding agencies and reviewers will carefully consider the sample size justifications in proposals. The scrutiny of sample size considerations is widely known by authors of successful proposals. There is no technical reason of course why studies that are not being considered for funding should ignore sample size planning.

Although we realize that to some, perhaps many, researchers, planning sample size is not as exciting as planning other aspects of a study (e.g., selecting measures to use, theoretical model to evaluate, population from which to sample, hypothesis generation) or performing analyses on the collected data, sample size planning is itself a fundamentally important task to a well-designed study. We believe, in fact, that an appropriate sample size is a necessary, but certainly not a sufficient, requirement for a well-designed study. One reason why sample size should be explicitly considered before starting a study is to evaluate the potential success of a study with a specified sample size that the researcher wants, or is able, to use. That is to say, how likely is it that a study will be successful at the sample size that a researcher wants or is able to use? Alternatively, sample size should be explicitly considered in order to plan for a sample size that would have a sufficiently high likelihood of leading to a successful study.

A very practical reason why sample size should be considered is to avoid starting a study in which the a priori success rate is unacceptably low given the sample size that is available or that the researcher has the time or funds to collect. That is, if a researcher knows that he or she will only be able to have, say, 100 participants in the study, yet the sample size planning procedure suggests that 350 participants are needed, the researcher may elect not to start the study due to the less-than-desired probability of accomplishing the study's goals. For example, it could be the case that the goals of the study will only have a 10% chance of being successful. Is a 10% chance of success really worth putting in the substantial amount of time, effort, and resources required? Maybe it is still worth conducting a study that has a low chance of success, as there is often more than a single goal of a study. Either way, however, considering sample size before the start of the study allows the researcher to consider this information before conducting the study, rather than only learning of the likely reason why the failure after the study's completion.

We have said much about the "goals of a study" without identifying exactly what these "goals" are. First, consider an effect size, which we conceptualize along the lines of Kelley and Preacher (2012) as "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (p. 140). In this context, mean differences, regression coefficients, path coefficients, factor loadings, correlations, proportion of variance accounted for type measures (e.g., the

squared multiple correlation, eta-squared), and measures of fit, among others, are effect sizes. When we say "the effect," we mean the effect size of interest.

When considering a study's goals, Maxwell, Kelley, and Rausch (2008) consider the idea of showing (1) the existence of an effect (and possibly its direction), (2) accurately estimating the magnitude of an effect, or (3) showing the existence of an effect **and** estimating an effect accurately. By "showing the existence of an effect," we mean that the specified null value of the effect is rejected with a null hypothesis significance test. By "accurately estimating the magnitude of the effect," we mean that the confidence interval for the parameter of interest is sufficiently narrow.

More formally, in the context of **showing the existence of an effect**, the study will often be considered a "success" if the p-value $< \alpha$, where the p-value is the probability of obtaining results as or more extreme than observed under the null hypothesis and α is the Type I error rate. That is to say, when the null hypothesis of the effect of interest can be rejected, support is shown for the alternative hypothesis, which is usually the research hypothesis of interest. When the goal of a study is to reject a false null hypothesis, a power analysis should be performed so that the sample size that leads to sufficient power, under the specified assumptions and study design, can be planned. This type of sample size planning is the traditional approach and is what Jacob Cohen spent considerable time positing should be done to improve research quality throughout much of his career (e.g., as is exemplified in his power analysis book, 1988).

In the context of **accurately estimating the magnitude of the effect**, such a conceptualization can be considered a success when the $(1-\alpha)100\%$ confidence interval for the corresponding population value is sufficiently narrow. That is to say, when sampling error is reduced (precision is improved) and the estimator is not more biased in the process of improving precision, the estimate is made more accurate. Accuracy of an estimate is a function of two quantities: precision and bias. If bias is zero, and thus the estimate is unbiased, improving precision improves accuracy: in the case of an unbiased estimate precision and accuracy are the same (e.g., Kelley & Maxwell, 2003). Thus, holding everything else constant, a narrower confidence interval equates to a more accurate estimate. Sample size planning in this framework has been termed accuracy in parameter estimation (AIPE; e.g., Kelley & Maxwell, 2003). Although, as noted, Cohen spent considerable time discussing the importance of appropriately powered studies, in the latter part of his career he shifted focus to the importance of effect sizes and their corresponding confidence intervals. A useful way of understanding the goal of AIPE is illustrated by solving the problem noted here with regards to why Cohen suspected, at that time, so few researchers reported confidence intervals: "I suspect that the main reason they [confidence intervals] are not reported is that they are so embarrassingly large!" (1994, p. 1002). The AIPE approach to sample size planning seeks to plan sample size so as to avoid "embarrassingly large" confidence intervals. All of these ideas are general and in no way limited to the CFA context.

Effect Size

When considering estimates in CFA, it is useful to frame them in the context of an effect size. In the context of CFA, factor loadings, correlations, and measures of fit are each effect sizes. There is an important distinction between two general types of effect sizes,

namely omnibus and targeted. An omnibus effect size relates to the overall model, whereas a targeted effect size relates to a specific well-defined part of the model. In the context of CFA, a targeted effect size could be a factor loading or correlation coefficient among two constructs, whereas an omnibus effect size could be a fit index, such as the root mean square error of approximation (RMSEA), comparative fit index (CFI), or Tucker–Lewis index (TLI), for example.

In addition to the distinction between targeted and omnibus effect sizes, one can consider whether effect sizes are standardized or unstandardized. In general, unstandardized path coefficients in CFA (and SEM) models are preferred from a methodological perspective, as the sampling distribution of standardized coefficients has some technical issues associated with their standard errors (see Cudeck, 1989, for a discussion of issues associated with standardized coefficients in CFA and SEM models).[3] Fit indices are a different type of effect size that are standardized because they are not wedded to the particular measurement scales themselves. Unstandardized path coefficients, on the other hand, are wedded to the measurement scales of the variables used. Correspondingly, their interpretation is necessarily in the context of the variances and covariances of the manifest variables. This is no different than the interpretation of a regression coefficient. Consider the case of simple regression, where the regression coefficient of Y on X is estimated as $\frac{Covariance(X,Y)}{Variance(X)} = s_{xy}/s_x$, which clearly shows that the scaling depends on both the regressor (X) and the outcome variable (Y).

As we have illustrated, effect size can be considered in a two-by-two conceptualization comprised of scaling (standardized or unstandardized) and scope (targeted or omnibus). Coupling the type of effect size(s) with the goal of showing the existence of an effect or its magnitude is needed in order to plan an appropriate sample size. Kelley and Maxwell (2008) provide a two-by-two conceptualization for sample size planning. That conceptualization consists of one dimension in which the interest concerns whether the approach to planning sample size is statistical power or AIPE and another dimension in which interest concerns whether the effect of interest is omnibus or targeted. A modified version of that table was given in Kelley and Maxwell (2012) and is provided here in Table 5.1.

Table 5.1 shows that sample size can be considered in a two-by-two framework. Although the table does not make the distinction between standardized and unstandardized values, implicit is that the effect size of interest is either standardized or not.[4] Use of Table 5.1 is only a guide, as the table supposes that only a single effect size is of interest. Simultaneously considering multiple effect sizes (e.g., path coefficients and fit indices) or multiple goals (e.g., power and accuracy) is beyond the discussion of this chapter (but see Maxwell et al., 2008). However, a simple way to consider an appropriate sample size is to plan for effects of interest for the particular goal(s) and use as the necessary sample size the largest of the multiple sample sizes.

[3] The issue has to do with there being extra randomness in the model not explicitly accounted for due to the standardization process. Consider standardization in multiple regression in which an unstandardized regression coefficient is multiplied by a random variable that is the quotient of two estimated standard deviations, one for the dependent variable and one for the regressor of interest. In repeated samplings from the same population, the ratio of standard deviations (that is multiplied by the unstandardized regression coefficient) will vary. This extra variability is unaccounted for in the estimated standard errors of standard regression coefficients.

[4] It is also possible to consider partially standardized effect sizes. Consider, for example, a multiple regression model in which a regressor is standardized but the outcome variable is kept in its raw score form.

Table 5.1 Goals of statistical inference for a two-by-two conceptualization of possible scenarios when the approach (statistical power or AIPE) is crossed with the type of effect size (omnibus or targeted).

		Type of Effect Size	
		Omnibus	*Targeted*
Approach	Statistical Power	*Establish existence of an omnibus effect by rejecting the null hypothesis that the population value of the omnibus effect is consistent with the specified null hypothesis*	*Establish existence of a targeted effect by rejecting the null hypothesis that the population value of the targeted effect is consistent with the specified null hypothesis*
	Accuracy in Parameter Estimation	*Establish the magnitude of the omnibus effect by obtaining a narrow confidence interval for the population omnibus effect*	*Establish the magnitude of the targeted effect by obtaining a narrow confidence interval for the population targeted effect*

Table taken from Kelley & Maxwell (2012).

In the remainder of this chapter, we discuss how to plan sample size for the case of a single factor model and also for a bifactor model. We use R (R Core Team, 2014) and our R software package, MBESS (Kelley & Lai, 2014), both of which are open source and freely available. Many good introductions to R and their use in different fields are available, both online and in book form (and some online books, such as Venables, Smith, & the R Core Team, 2014). Some of the sample size planning methods discussed in this chapter are also available in other statistical packages. We use R because it is freely available, it is easy to use, and we have provided functions that can be used for each of the scenarios we discuss. We hope that our chapter is able to stimulate researchers to consider the multiple ways in which sample size planning can be conceptualized when interest concerns parameters from CFA or related models.

Empirical Demonstration 1

We provide two worked examples to illustrate the sample size planning theories discussed previously. The examples are in the context of studies in social and personality psychology, but they can be easily generalized to other psychometric problems. In particular, the first example is based on a one-factor CFA model, and the second example is a scenario in which a bifactor and a second-order CFA model are compared.

Suppose a researcher is interested in the cognition of a certain group of adults and selects five manifest variables to measure cognition. The proposed model is a one-factor CFA model and is depicted in Figure 5.1. Appropriate manifest variables in the present example can be (1) scores from questions such as "Have difficulty reasoning and solving problems?" or "Forget where you put things or appointments?"; (2) the summary scores of a group of items; or (3) scores from a standardized test. After selecting the five indicators for the model, the researcher plans the necessary sample size for the study. Recall that we previously explained four different methods to plan

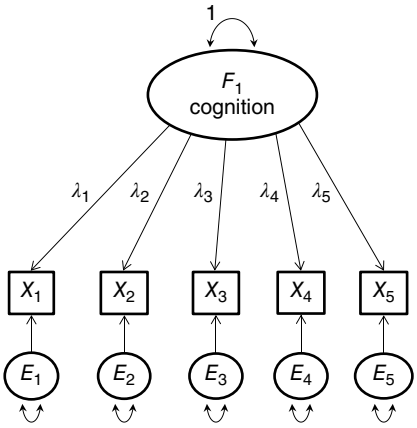

Figure 5.1 Confirmatory factor model for Cognition with five manifest variables.

the sample size in a factor analysis study, and in the present example sample size planning will be illustrated using all four of those methods. However, note that in practice the researcher may not be interested in all four questions, but only one or two of them and, accordingly, the sample size only needs to be large enough to address the one or two questions of interest.

Approach 1: Power analysis for the model's population RMSEA At least two types of null hypotheses can be stated with respect to the model's population RMSEA, denoted as ε: (1) the traditional null hypothesis in which $H_0 : \varepsilon = 0$ (often called the nil hypothesis because the hypothesized value is 0); (2) a more realistic null hypothesis that tests the minimum effect of interest, such as $H_0 : \varepsilon \geq .08$. Rejecting the first type of null hypothesis will lead to the conclusion that the model's fit is not perfect, but this conclusion is not informative because, even without carrying out the study, the researcher already knows that the proposed model is unlikely to have perfect fit. The second type of null hypothesis states that the proposed model's fit is worse than some value of interest and rejecting the null hypothesis will lead to the conclusion that the model's fit is better than the cutoff value. In the present example, we use .08 as the cutoff value of interest, and demonstrate how to test the null hypothesis $H_0 : \varepsilon \geq .08$. Depending on the researcher's interest and prior knowledge of the model's statistical adequacy, the second type of null hypothesis can use many other cutoff values to describe ε (e.g., $H_0 : \varepsilon \geq .05$, $H_0 : \varepsilon \geq .06$, $H_0 : \varepsilon \geq .10$, etc.). Regardless of which null hypothesis the researcher chooses to test, calculating the sample size in order to have adequate power to test either type of hypothesis with respect to the RMSEA requires the following input information: (1) the model's degrees of freedom (df), (2) the Type I error rate (α), (3) the desired power ($1-\beta$, with β being the Type II error rate), (4) the population RMSEA under the null hypothesis (ε_0), and (5) the true population RMSEA (ε). Based on the path diagram, $df = 5$. For this example, we use $\alpha = .05$ and $1-\beta = .80$. The value for ε_0 is the value H_0 states, and thus ε_0 is 0 or .08, depending on which null hypothesis is of interest. Then the last and most difficult piece of input information is the true population

Table 5.2 Covariance matrix (upper triangle) and correlation matrix (lower triangle) of a hypothetical previous study in Empirical Demonstration 1.

	X_1	X_2	X_3	X_4	X_5
X_1	$(1.1)^2$	.548	.320	.828	.343
X_2	.415	$(1.2)^2$	.430	1.024	.445
X_3	.364	.448	$(.80)^2$	.610	.326
X_4	.502	.569	.509	$(1.5)^2$	.619
X_5	.346	.412	.453	.459	$(.90)^2$

The principal diagonal is the standard deviation in parentheses squared (i.e., the variance)

RMSEA, which can be estimated by the researcher based on the literature, previously collected data, or even a pilot study.

Suppose a previous study using $N = 420$ reported a covariance matrix of X_1 through X_5 as Table 5.2 indicates. Fitting the proposed model to that sample covariance matrix will yield $\hat{\varepsilon} = .035$. Suppose, based on this previous study result and substantive theories, the researcher believes that the population RMSEA is likely to be around .045 and uses $\varepsilon = .045$ as the input information. To calculate the necessary sample size so as to reject $H_0 : \varepsilon = 0$, the researcher specifies the input information for the R function from the MBESS package `ss.power.sem()` as follows:

```
> ss.power.sem(RMSEA.null=0, RMSEA.true=.045, df=5,
alpha=.05, power=.80),
```

where `RMSEA.null`, `RMSEA.true`, `df`, `alpha`, and `power` refer to the RMSEA under the null hypothesis, the true population RMSEA, the model's degrees of freedom, the Type I error rate, and the desired power, respectively. The necessary sample size estimated based on this input information is 1,269.

Alternatively, if the researcher is interested in rejecting a more realistic null hypothesis, say $H_0 : \varepsilon \geq .08$, then the sample size can be calculated as follows:

```
> ss.power.sem(RMSEA.null=.08, RMSEA.true=.045, df=5,
alpha=.05, power=.80).
```

Note that in the present case `RMSEA.null` is .08 as it is the RMSEA value under H_0, and `RMSEA.true` remains 0.045 as the true population RMSEA does not change regardless of what the null hypothesis states. The necessary sample size estimate based on this input information is 1,113. Note also that in this case testing $H_0 : \varepsilon \geq .08$ requires a smaller sample size than does testing $H_0 : \varepsilon = 0$, but one should not over-generalize this result. Depending on the input information and how the null hypothesis is stated (e.g., $H_0 : \varepsilon \geq .05$ instead of $H_0 : \varepsilon \geq .08$), the minimum-effect hypothesis may require a larger or smaller sample size than does the nil hypothesis.

Approach 2: AIPE for the model's population RMSEA Consider again the covariance matrix in Table 5.2. Fitting the proposed model to this covariance matrix of $N = 420$ yields $\hat{\varepsilon} = .035$, as well as a 90% confidence interval for ε, namely [0, .082]. That is, if this CI indeed includes ε, then the population RMSEA may be very close to 0, meaning

the model has nearly perfect fit, or may be as large as .082, meaning the model fit is just about "fair" according to common standards. Therefore, there is much uncertainty about the population RMSEA, although the sample data indicates that the fit is relatively "good." If the researcher is interested in estimating ε with a higher degree of accuracy, they can plan the sample size with the goal to achieve a sufficiently narrow CI. Planning the sample size for the RMSEA from the AIPE perspective requires the following input information: (1) model's degrees of freedom df, (2) confidence level $1-\alpha$, (3) the true population RMSEA ε, and (4) desired CI width ω. The first three pieces of input information are the same as before, so let us focus on the desired CI width ω. The width of the CI based on $N = 420$ is .082, and suppose for the future study the researcher desires a CI whose width is no larger than .060; that is, $\omega = .060$. Then the sample size can be calculated using the function `ss.aipe.rmsea()` as follows:

```
> ss.aipe.rmsea(RMSEA=.045, df=5, width=.060, conf.
level=0.90)
```

where `RMSEA`, `df`, `width`, and `conf.level` refer to the population RMSEA, model's degrees of freedom, desired CI width, and confidence level, respectively. The necessary sample size estimated based on this input information is 802.

Note there is no such thing as "the correct desired CI width" or "correct power," as the researcher sets either (or both) based on the goals he or she has for the study. In the present example, we selected $\omega = .060$ because we want to achieve a narrower width compared to the one observed in a previous study (i.e., .082). This estimated sample size, $N = 802$, means that, if the input information is correct and all the statistical assumptions are satisfied, the researcher can expect to obtain a CI whose width is no larger than .060. An implication of the expectation is that about half of the time the interval will be narrower than desired and about half of the time the interval will be wider than desired. The researcher could well use other values for ω, in a sensitivity analysis, so as to better understand the tradeoff between the increase in sample size and the reduction in CI width. That is, given ε, df, and α, a decrease in ω (i.e., narrower CI) will result in an increase in N, but the relationship between ω and N is usually nonlinear, and a small decrease in ω may require a small or large increase in N, depending on all factors. To illustrate, consider using $\omega = .050$ instead of $\omega = .060$ as input:

```
> ss.aipe.rmsea(RMSEA=.045, df=5, width=.050, conf.
level=0.90)
```

the estimated sample size would be 1,059. Therefore, in order to reduce the expected CI width from 0.060 to 0.050, it requires 257 additional participants. If one further reduces the desired expected CI width to $\omega = .040$, the necessary sample size would be 1,544. A 0.010 decrease from $\omega = .060$ to $\omega = .050$ requires an increase of 257 in N, whereas a 0.010 decrease from $\omega = .050$ to $\omega = .040$ requires an increase of 485 in N. By comparing the sample size estimates at various ω values, the researcher can better understand the interplay between resources (in the present context, N) and the estimation certainty (in the present context, expected CI width), and consider whether the gain in estimation certainty is worth the extra participants. Because there is no "correct" CI width, reasonable values for ω can be those that help achieve a balance between the sample size invested and the knowledge obtained about the population parameters. Lin

and Weng (2014) provide a graphical approach to sample size planning for AIPE in the context of the RMSEA, where one can assess the effect sizes of desired width and the population RMSEA on the necessary sample size.

Approach 3: Power analysis for the population model parameter of interest In addition to the model's overall fit, some specific model parameters are frequently of interest. Suppose X_1. represents the scores on the item "Have difficulty reasoning and solving problems" and X_2 represents the scores on the item "React slowly to questions," and it is of interest to infer whether these two items have the same reliability in measuring the latent construct cognition. Put another way, it is of interest to know whether X_1 and X_2 have the same population factor loadings. The null hypothesis is $H_0 : \lambda_1 = \lambda_2$, and rejecting this null hypothesis will lead to the conclusion that $H_a : \lambda_1 \neq \lambda_2$. To test this null hypothesis, one can carry out a chi-square likelihood ratio test that compares models with and without the constraint $\lambda_1 = \lambda_2$. To calculate the necessary sample size in order to have enough power to perform such a test, the input information is (1) the model, (2) Type I error rate, (3) desired power, and (4) the population covariance matrix of the manifest variables, $\boldsymbol{\Sigma}$. Note this approach requires the complete specification of the statistical model, not just the model's degrees of freedom, and thus input item (1) in the present context is different from the input item (1) in the context of power analysis for the RMSEA.

Let us now consider input item (4), the population covariance matrix. Broadly speaking, there are two ways to specify $\boldsymbol{\Sigma}$. The first way is to specify all the variances and covariances in $\boldsymbol{\Sigma}$, and the second way is to specify the model parameter values first and then use the model-implied covariance matrix as $\boldsymbol{\Sigma}$. Methods to facilitate specifying $\boldsymbol{\Sigma}$ are, for example, discussed in Lai and Kelley (2011). In the present example, we demonstrate how to specify $\boldsymbol{\Sigma}$ based on a previous study. Consider the sample covariance matrix in Table 5.2. Fitting the model to this covariance matrix will return the model parameter estimates shown in Table 5.3. Because the estimation is based on a covariance matrix, the resulting model parameter estimates are in the unstandardized metric and not straightforward to interpret. In Table 5.3 we have also provided model parameter estimates in the standardized metric, so let us focus on the standardized estimates for the moment. The researcher must then specify the input information $\boldsymbol{\Sigma}$ using the following

Table 5.3 Model parameter estimates based on the sample covariance matrix in Table 5.1.

	Unstandardized estimate	*Standard error*	*Standardized estimates*	*Standardized input information*
λ_1	.659	.054	.600	.550
λ_2	.839	.056	.700	.730
λ_3	.519	.038	.650	.680
λ_4	1.199	.068	.800	.750
λ_5	.539	.044	.600	.600
e_1	.773	.061	.640	.6636
e_2	.773	.065	.510	.4816
e_3	.369	.031	.577	.5376
e_4	.808	.093	.360	.4375
e_5	.517	.041	.640	.640

three steps. First, based on these standardized model parameter estimates, the researcher must make educated guesses as to the population model parameters in the standardized context. Second, using the population standardized model parameters, the model-implied correlation matrix $\mathbf{P}(\boldsymbol{\theta})$ is calculated. Third, the researcher specifies the standard deviations of the manifest variables, and using these transforms $\mathbf{P}(\boldsymbol{\theta})$ into the covariance matrix.

Suppose that, based on substantive theories, the researcher believes that X_1 should have lower reliability in measuring the latent factor and thus uses $\lambda_1 = .550$ as input. Similarly, substantive theory also suggests that the reliability of X_2 and X_3 should be higher than has been reported in the previous study, and accordingly the researcher uses $\lambda_2 = .730$ and $\lambda_3 = .680$ as input. To be conservative, the researcher uses a lower factor loading for X_4: $\lambda_4 = .750$. The input for λ_5 remains the same as reported in the previous study: $\lambda_5 = .600$. Because the model parameters are in the standardized metric, the error variance in the present model can be easily calculated as unity minus the square of the factor loading (i.e., $1 - \lambda^2$). Appendix Code 1 demonstrates how to obtain the model-implied correlation matrix using the MBESS package, and the resulting $\mathbf{P}(\boldsymbol{\theta})$ is included in Table 5.4. Based on the previous study, suppose the standard deviations of X_1 through X_5 are believed to be 1.20, 1.20, .70, 1.30, and 1.10, respectively. Now the input covariance matrix can be obtained and this is provided in Table 5.4. We use $\boldsymbol{\Sigma}_1$ to denote this input covariance matrix.

Now that all of the input information is complete, the researcher can calculate the necessary sample size using the method Satorra and Saris (1985) developed. In essence, Approach 3 is based on the sampling distributions of the model chi-square statistic under H_0 and H_a, and the Satorra–Saris (1985) method helps to obtain a key distribution parameter (namely the noncentrality parameter) for the sampling distribution under H_a using the population covariance matrix of manifest variables. We will explain the underlying statistical theories in more detail in a later section, and for the moment let us continue to demonstrate how to use Approach 3 to plan the sample size. Recall the present task is after specifying the model, Type I error rate, desired power, and $\boldsymbol{\Sigma}_1$ as input information, how to calculate the sample size so as to reject $H_0 : \lambda_1 = \lambda_2$ with a sufficiently high probability (i.e., desired power). Let the proposed model for the present study be referred to as the full model and the model that constrains $\lambda_1 = \lambda_2$ be referred to as the restricted model. The full model will fit the input covariance matrix $\boldsymbol{\Sigma}_1$ perfectly, but the restricted model will have some misfit to $\boldsymbol{\Sigma}_1$. In particular, fitting the restricted model to $\boldsymbol{\Sigma}_1$ will yield a maximum likelihood (ML) discrepancy value $F_{Res} = .02346$ and $\varepsilon_{Res} = .0625$, where F_{Res} is a shorthand for the F_{ML} value in the restricted model. The F_{ML} value in the full model is 0 ($F_{Full} = 0$). The difference in degrees of freedom between the full model and the restricted model is 1. Given this information the sample size can then be planned using the function `ss.power.sem()` as follows:

Table 5.4 Covariance matrix (upper triangle) and correlation matrix (lower triangle) of input information for sample size planning.

	X_1	X_2	X_3	X_4	X_5
X_1	$(1.20)^2$	.418	.394	.435	.348
X_2	.601	$(1.20)^2$	.490	.540	.432
X_3	.331	.411	$(.70)^2$	.510	.408
X_4	.679	.842	.464	$(1.30)^2$	.450
X_5	.459	.570	.314	.644	$(1.10)^2$

```
> ss.power.sem(F.full=0, F.res=.02346, df.full=5, df.res=6,
alpha=.05, power=.80)
```

where `F.full`, `F.res`, `df.full`, and `df.res` refer to the full model's F_{ML}, the restricted model's F_{ML}, the full model's degrees of freedom, and the restricted model's degrees of freedom, respectively. The necessary sample size is 336. Therefore, in order to have .80 power to demonstrate that X_1 and X_2 have different factor loadings, sample size needs to be $N = 336$.

Note the function `ss.power.sem()` asks for F_{ML} values as input only instead of the population covariance matrix of the manifest variables, and the Satorra–Saris method is sometimes difficult to employ because it usually requires a large amount of information to specify the population covariance matrix. Alternatively, MacCallum, Browne, and Cai (2006) proposed a method that utilizes the relationship between RMSEA and F_{ML}, so as to specify the F_{ML} values in terms of the RMSEA. The MacCallum et al. method has the same goal as does the Satorra–Saris method, namely to obtain a key distribution parameter for the sampling distribution under H_a, but it requires only the population RMSEA values under H_0 and H_a, and thus circumvents the task of specifying the population covariance matrix. We will explain this method's underlying theories in a later section, and for the moment continue to demonstrate the MacCallum et al. method. Recall we obtained the population RMSEA values earlier when fitting $\boldsymbol{\Sigma}_1$ to the full and restricted models; that is, $\varepsilon_{Res} = .0625$ and $\varepsilon_{Full} = 0$. Accordingly, we can employ MacCallum et al. method and use the following call to the `ss.power.sem()`:

```
> ss.power.sem(RMSEA.full=0, RMSEA.res=.0625, df.full=5,
df.res=6, alpha=.05, power=.80),
```

which also returns $N = 336$, equivalent to the N obtained with the Satorra–Saris method using $\boldsymbol{\Sigma}_1$ as input. In practice the MacCallum et al. method is easier to implement as compared to the Satorra–Saris method, because it is easier to obtain knowledge about the model's general adequacy (operationalized as the population RMSEA) as compared to the variances and covariances of all the manifest variables in the model.

The above example demonstrated that, if the input RMSEA equals the population RMSEA, the sample size estimate will equal the idealized sample size based on the population RMSEA. But if the input information is not exactly accurate, the sample size returned will be different from the idealized sample size. To illustrate, suppose the researcher does not have enough knowledge about the phenomenon under study to accurately specify $\boldsymbol{\Sigma}_1$, but can only judge that the restricted model should have a "fair" to "good" statistical fit. Accordingly, the researcher chooses .07 as a proxy for ε_{Res} and plans the sample size as follows:

```
> ss.power.sem(RMSEA.full=0, RMSEA.res=.07, df.full=5,
df.res=6, alpha=.05, power=.80).
```

The resulting sample size estimate is $N = 268$, falling short of the theoretically optimal necessary sample size based on the correct population RMSEA (had it been known to the researcher). However, if the researcher uses .06 as the input RMSEA value, the resulting N will be 364, which is close to the idealized N. How much the sample size estimate differs from the idealized sample size hinges upon how close the input parameter is to the population parameters (e.g., population covariance matrix of manifest

variables, population RMSEA of the restricted model), and therefore effort is required to ensure the quality of the input information.

Statistical theories for SEM power analysis In this subsection, we briefly explain the underlying statistical theories for the Satorra–Saris method and the MacCallum et al. method. Detailed explanations are available in Satorra and Saris (1985) and MacCallum et al. (2006). The present task is to assess the difference in statistical fit between a full model and a restricted model. Statistical fit in the current context is derived from the model estimation results using the following ML discrepancy function:

$$F_{ML} = \ln\left|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right| - \ln|\mathbf{S}| + tr\left(\mathbf{S}\cdot\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\right) - p, \tag{5.1}$$

where $\ln(\cdot)$ is the natural logarithm, $\mathbf{S}$ is the sample covariance matrix, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ is the model-implied covariance matrix given $\mathbf{S}$, and p is the number of manifest variables. We use $\hat{F}$ and F^* to denote the sample and population discrepancy function value, respectively. If the data follow a multivariate normal distribution and the models are not badly misspecified, we have the following three approximate chi-square distributions:

$$n\hat{F}_{Full} \approx \chi^2(df_{Full}, \delta_{Full}); \tag{5.2}$$

$$n\hat{F}_{Res} \approx \chi^2(df_{Res}, \delta_{Res}); \tag{5.3}$$

$$n\left(\hat{F}_{Res} - \hat{F}_{Full}\right) \approx \chi^2(df_{Res} - df_{Full}, \delta_{Res} - \delta_{Full}); \tag{5.4}$$

where df_{Full} and df_{Res} refer to the degrees of freedom of the full and restricted models, respectively, δ_{Full} and δ_{Res} refer to the noncentrality parameter of their respective distributions, $n = N - 1$, and "$\approx$" signifies that the (random) quantity on the left approximately follows the distribution on the right. The noncentrality parameter is based on the population discrepancy function value:

$$\delta_{Full} = nF^*_{Full}; \tag{5.5}$$

$$\delta_{Res} = nF^*_{Res}. \tag{5.6}$$

Accordingly, the noncentrality parameter for the distribution in Equation 5.4 is simply $\delta_{Res} - \delta_{Full} = n\left(F^*_{Res} - F^*_{Full}\right)$. Recall of interest is whether the full and restricted models have equal statistical fit, and we use the test statistic $n\left(\hat{F}_{Res} - \hat{F}_{Full}\right)$ and its sampling distribution (i.e., Equation 5.4) to perform a hypothesis test and examine the question of interest. To simplify the exposition, let T_{Diff} denote this test statistic; accordingly, the sampling distribution can be expressed as $T_{Diff} \approx \chi^2(df_{Res} - df_{Full}, \delta_{Res} - \delta_{Full})$.

Now let us take these statistical theories back to Approach 3's example, where the interest is to infer whether the two factor loadings, λ_1 and λ_2, are equal in the population. If the null hypothesis $H_0 : \lambda_1 = \lambda_2$ is true, the full and restricted models fit equally well and are both correctly specified. Thus, under the null hypothesis, $F^*_{Res} = F^*_{Full} = 0$ and T_{Diff} approximately follows a chi-square distribution with degrees of freedom $(df_{Res} - df_{Full})$ and a noncentrality parameter of 0. That is,

$$T_{Diff} \approx \chi^2(df_{Res} - df_{Full}, 0) \tag{5.7}$$

under H_0. Under the alternative hypothesis $H_a : \lambda_1 \neq \lambda_2$, the full model is correctly specified but the restricted model is misspecified, and therefore $F^*_{Full} = 0$ and $F^*_{Res} > 0$, implying $\delta_{Full} = 0$ and $\delta_{Res} > 0$. The noncentrality parameter $\delta_{Res} - \delta_{Full}$ reduces to δ_{Res}, which is equivalent to nF^*_{Res}. That is,

$$T_{Diff} \approx \chi^2 \left(df_{Res} - df_{Full}, nF^*_{Res} \right) \tag{5.8}$$

under H_a. Given the distributions under H_0 and H_a, one can plan the sample size by finding the smallest value of n in Equation 5.8, such that there is $(1-\beta)100\%$ probability of rejecting H_0. But note F^*_{Res} in Equation 5.8 is a population parameter unknown to the researcher, and thus the specific form of the distribution is not determined and Equation 5.8 cannot be readily used for sample size planning. How to provide a value for F^*_{Res} (or equivalently speaking, the noncentrality parameter δ_{Res}) is where the Satorra–Saris method and the MacCallum–Browne–Cai method come into play.

In particular, Satorra and Saris proposed using the model-implied covariance matrix of the full model to obtain F^*_{Res}. That is, because the full model is specified correctly, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{Full}(\boldsymbol{\theta})$. If one supplies all the model parameter values to the full model, the model-implied covariance matrix $\boldsymbol{\Sigma}_{Full}(\boldsymbol{\theta})$ will be equivalent to $\boldsymbol{\Sigma}$. Next one fits the restricted model to $\boldsymbol{\Sigma}$, and the resulting discrepancy function value will be F^*_{Res}. Alternatively, MacCallum et al. utilized the relationship between the discrepancy function value and RMSEA. Commonly, RMSEA is defined as

$$\varepsilon = \sqrt{\frac{F_{ML}}{df}} \tag{5.9}$$

in the population, and thus $F_{ML} = df \cdot \varepsilon^2$. If the population RMSEA of the restricted model is known, one can obtain F^*_{Res} using the identity $F^*_{Res} = df_{Res} \cdot \varepsilon^2_{Res}$. Therefore, the Satorra–Saris method and the MacCallum et al. method try to solve the same problem: namely how to obtain F^*_{Res}-- from two different perspectives, and they are independent approaches. After F^*_{Res} is obtained with either method, it is supplied to the sample size planning process along with other input information, such as the full and restricted models' degrees of freedom, Type I error rate, and desired power.

Approach 4: AIPE for the population model parameters of interest The fourth perspective on sample size planning concerns the case when it is desired to estimate a model parameter of interest with sufficient accuracy, as expressed by the width of the confidence interval. Consider again the model parameter estimates reported from the previous study as shown in Table 5.3. Suppose the researcher is interested in knowing the true value of λ_1 in the population. Based on the previous study $N = 420$, $se(\lambda_1) = .054$ and the 95% CI for λ_1 is [.553, .765]. The width of this CI is .212. If the researcher desires a narrower CI in the study being designed, the necessary sample size can be calculated with the following input information: (1) the model, (2) confidence level, (3) desired value of the CI width, and (4) the population covariance matrix of the manifest variables, $\boldsymbol{\Sigma}$. Except for item (3), the required input information is the same as for Approach 3. The only difference in the input information between Approaches 3 and

4 is that the desired power is replaced with a desired CI width. In the present scenario, we continue to use all the input information we specified for Approach 3 except for item (3).

Recall that the 95% CI width for λ_1 reported in the previous study is .212 and that the researcher desires a narrower CI in the future study. If we used $\omega = .15$, the necessary sample size can be calculated using the function `ss.aipe.sem.path()` as follows:

```
> ss.aipe.sem.path(model=model.full, Sigma=Sigma.1, desired.
width=.15, which.path="l1", conf.level=0.95),
```

where `model`, `Sigma`, and `which.path` refer, respectively, to the model, the input population covariance matrix $\boldsymbol{\Sigma}_1$ (i.e., the covariance matrix in Table 5.4), and the name of the model parameter of interest. Due to space limitations, this call to `ss.aipe.sem.path()` omits the intermediate steps that specify objects `model.full` and `Sigma.1`. These intermediate steps are included in Appendix Code 1. Given this input information, the function will return a necessary sample size estimate $N = 1007$.

Planning N using $\omega = .15$ only guarantees that the expectation of the (random) CI widths in the long run is no larger than .15, but in the particular study to be carried out, the CI width will be less than .15 with a probability of about 50%. That is, in practice one forms a confidence interval based on a sample covariance matrix, and because values in the sample covariance matrix vary over repeated sampling, the confidence interval obtained is necessarily random, varying over repeated sampling. Accordingly, the width of a CI varies from sample to sample as well and is a random variable. Let w, a random variable, denote the width of a CI. Using $N = 1007$ will only ensure $E[w] \leq .15$, implying $w \leq .15$ about half of the time and $w \geq .15$ about the other half of the time. Therefore, in the particular study the researcher is going to conduct, there is only about 50% probability that the CI width to be observed will be less than .15. If the researcher desires a higher probability for the event $w \leq .15$ to appear in a particular study, the sample size can be planned with an extended procedure that ensures $\Pr(w \leq .15) = \gamma$, where γ is referred to as the assurance parameter. For example, in order to achieve $\Pr(w \leq .15) = .90$, we can calculate the sample size by specifying `ss.aipe.sem.path()` as follows:

```
> ss.aipe.sem.path(model=model.full, Sigma=Sigma.1, desired.
width=.15, which.path="l1", conf.level=0.95, assurance=.90),
```

where `assurance` is the assurance parameter. This call to the function will return $N = 1{,}064$. Thus, if the researcher increases the sample size from 1,007 to 1,064, there will be at least a 90% probability that $w \leq .15$ in the particular study to be carried out.

Empirical Demonstration 2

The second empirical demonstration is for a bifactor model and a second-order model, both of which are important special cases of CFA models. Suppose, in addition to the latent construct cognition, there are two other latent constructs in the study, coping and social support, and they are measured by three and four indicators, respectively. Further suppose that the researcher hypothesizes that there is a more

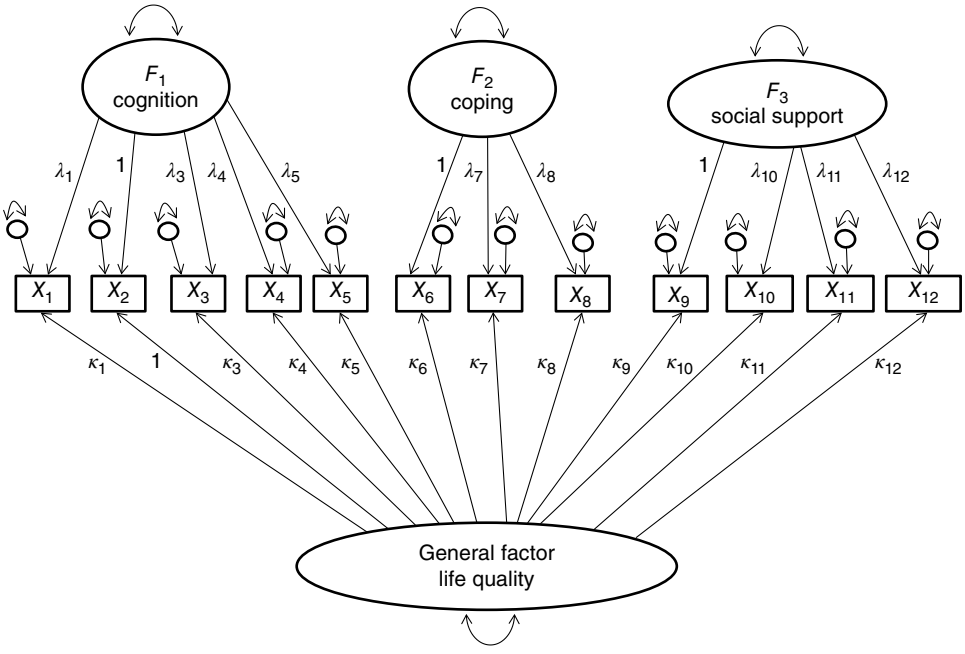

Figure 5.2 Bifactor model representation of items, where each item is measured from a specific factor (Cognition, Coping, or Social Support) and a general factor (Life Quality).

general factor than cognition, coping, and social support and terms this factor quality of life. Let cognition, coping, and social support be referred to as the specific factors, and let quality of life be referred to as the general factor. The dynamics among the general factor and specific factors can be conceptualized in terms of a bifactor model or a second-order CFA model. More specifically, the bifactor model hypothesizes that the general factor influences all of the 12 manifest variables directly, and the four latent factors are mutually independent. The second-order factor model hypothesizes that the general factor directly influences all the three specific factors, and thus the three specific factors are not independent of each other. The path diagrams for the bifactor model and the second-order factor model are included in Figures 5.2 and 5.3, respectively.

As we have discussed, and shown in Empirical Demonstration 1, the sample size for a bifactor model or second-order model can be planned from four different perspectives, and all the functions and methods we discussed previously are readily applicable in the present context. Therefore, in the present example we do not demonstrate sample size planning from all four perspectives, but rather focus on two interesting questions commonly raised in applications of bifactor and second-order models. First, the researcher often wants to understand whether the bifactor model or the second-order factor model can better explain the data. To address this question, we consider sample size calculations in the context of model comparisons. Second, after the researcher adopts one from these two competing models, some factor loadings within the model will likely be of interest. To address this question, we plan the sample size from the AIPE for model parameter perspective.

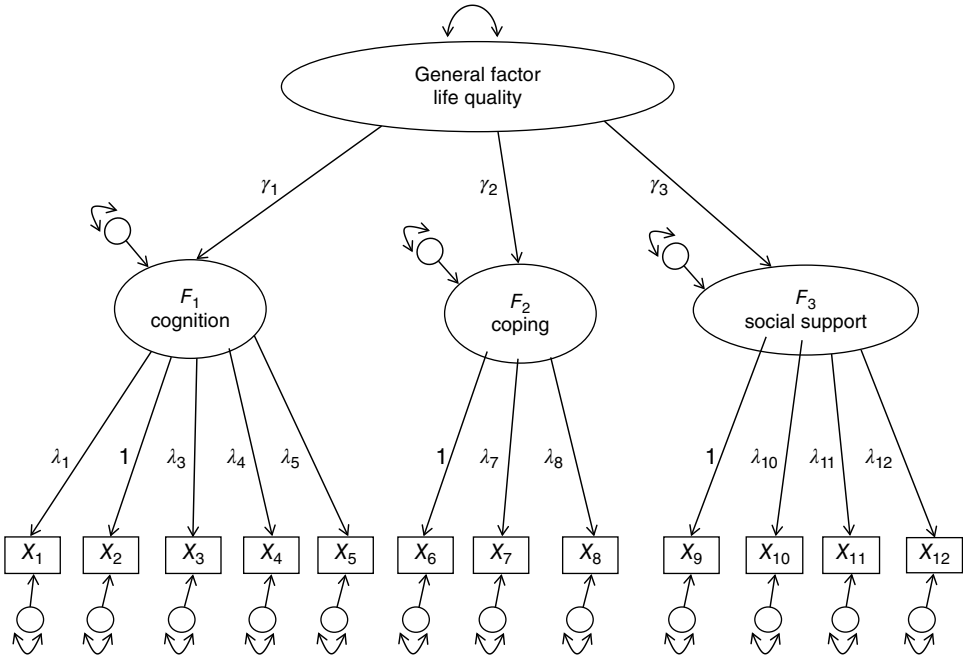

Figure 5.3 Second-order factor model representation of items, where the items are measured from specific factors (Cognition, Coping, or Social Support), and the specific factors are in turn measured from a second-order, higher level factor (Life Quality).

Approach 1: Power analysis comparing a bifactor model and a second-order factor model
Yung, Thissen, and McLeod (1999) proved that the second-order factor model is nested within the bifactor model, and thus will almost always fit less well compared to the bifactor model. Accordingly, one can test the null hypothesis that the second-order factor model fits equally as well as does the bifactor model. Rejecting this null hypothesis leads to the conclusion that the bifactor model has better fit. If one considers the bifactor model as the full model and the second-order factor model as the restricted model, sample size in the current study can be planned with Approach 3 in Empirical Demonstration 1. More specifically, it requires as input information the (1) full and restricted models, (2) Type I error rate, (3) desired power, and (4) population covariance matrix of the manifest variables, $\boldsymbol{\Sigma}$. Again, all the input information except (d) can be specified easily. We obtain (4) by first specifying the model parameters in the full model (i.e., the bifactor model).

Following the scheme used in Approach 3 in Empirical Demonstration 1, we specify $\boldsymbol{\Sigma}$ by first specifying all the standardized model parameter values in the bifactor model (i.e., the full model). Compared to the previous one-factor CFA model with only five indicators, the present bifactor model has a large number of parameters to be specified. To simplify this task, we utilize an exchangeable pattern of factor loadings within the same measurement cluster; that is, indicators that measure the same latent factor have the same standardized factor loadings. The rationale for using the exchangeable pattern in the SEM context to simplify the input information specification is explained in greater detail in Lai and Kelley (2011). In essence, an exchangeable pattern is reasonable because the necessary sample size estimate usually does not differ from the idealized sample size by much. We maintain the input factor loadings for measuring cognition

Table 5.5 Specifying standardized model parameters for the bifactor model in Empirical Demonstration 2.

	General factor	*Cognition*	*Coping*	*Social support*	*Error variance*
X_1	.500	.550			.4475
X_2	.500	.730			.2171
X_3	.500	.680			.2876
X_4	.500	.750			.1875
X_5	.500	.600			.390
X_6	.600		.600		.280
X_7	.600		.600		.280
X_8	.600		.600		.280
X_9	.700			.450	.3075
X_{10}	.700			.450	.3075
X_{11}	.700			.450	.3075
X_{12}	.700			.450	.3075

in the previous example, and specify all the new factor loadings in an exchangeable manner. The resulting input information for the standardized model parameters are included in Table 5.5. Based on the input information in Table 5.5, the bifactor model will reproduce a correlation matrix of X_1 through X_{12}. Then the researcher should make educated guesses or informed from other studies about the standard deviation of the 12 manifest variables and thereby will obtain an input covariance matrix for sample size planning. We use $\boldsymbol{\Sigma}_2$ to denote this input information. All of the syntax that leads to this input covariance matrix is included in Appendix Code 2.

Now that the input information is complete, one can plan the sample size. Note the bifactor model will fit $\boldsymbol{\Sigma}_2$ perfectly, but the second-order factor model will have some misfit to $\boldsymbol{\Sigma}_2$. More specifically, fitting the second-order factor model to $\boldsymbol{\Sigma}_2$ will yield $F_{sec-order} = .01891$ and $\varepsilon_{sec-order} = .01926$. To calculate the necessary sample size so as to have .80 power to reject the null hypothesis that the second-order factor model has the same population fit as does bifactor model, one can use the function `ss.power.sem()` as follows:

```
> ss.power.sem(F.full=0, F.res=.01891, df.full=42, df.res=51,
alpha=.05, power=.80).
```

The resulting necessary sample size estimate is $N = 829$. Note this is an application of the Satorra–Saris method to power analysis. Now suppose one does not have enough information to specify $\boldsymbol{\Sigma}_2$ and chooses to plan the sample size in terms of RMSEA (an application of the MacCallum et al. method). If the researcher could correctly specify $\varepsilon_{sec-order} = .01926$ as the input information, the sample size returned will be equivalent to the method that uses $\boldsymbol{\Sigma}_2$ as input information:

```
> ss.power.sem(RMSEA.full=0, RMSEA.res=.01926, df.full=42,
df.res=51, alpha=.05, power=.80),
```

which also returns a sample size estimate $N = 829$. Therefore, to demonstrate that the bifactor model subsumes the second-order factor model and has better fit, it requires a sample of size 829, provided that the present input information is correct and all the assumptions are satisfied. Due to the complexity of bifactor and second-order factor

models, it is usually not possible for the input RMSEA to exactly equal the population RMSEA. Suppose the researcher believes the difference in fit between the bifactor and second-order factor models is relatively small and uses .02 as the input value for RMSEA:

```
> ss.power.sem(RMSEA.full=0, RMSEA.res=.02, df.full=42,
df.res=51, alpha=.05, power=.80).
```

In this case, the function returns $N = 769$, slightly different from the idealized sample size 829.

Approach 2: AIPE for model parameters in the bifactor model In Approach 4 of Empirical Demonstration 1, the researcher was interested in the value of X_1's factor loading when measuring the latent factor cognition. Suppose in the present bifactor model, this factor loading is still of interest, but there is an extra parameter of interest, namely the loading of X_1 on the general factor, a loading denoted as κ_1. To plan sample size so as to achieve sufficient certainty (i.e., narrow enough CI) in estimating both λ_1 and κ_1, the process is exactly the same as was shown in Approach 4 of Empirical Demonstration 1, except that the researcher needs to perform the process twice, once for λ_1 and once for κ_1. The necessary sample size for the study is simply the larger of the two.

Recall that the input information for AIPE for model parameters is as follows: the (1) the model, (2) model parameter of interest, (3) confidence level, (4) desired expectation value of the CI width, and (5) population covariance matrix of the manifest variables. We continue to use $\boldsymbol{\Sigma}_2$ as the input covariance matrix, and all of the other input information can be readily specified by this point. We plan the sample size with respect to λ_1 first. Suppose the desired CI width is still .15. Then the function `ss.aipe.sem.path()` can be specified as follows:

```
> ss.aipe.sem.path(model=AIPE.bifactor, Sigma=Sigma.2,
desired.width=.15, which.path="l1", conf.level=0.95),
```

where `model` refers to the model's specification, `Sigma` refers to the input population covariance matrix $\boldsymbol{\Sigma}_2$, and `which.path` refers to the model parameter of interest. Due to space limit, we did not present the values of $\boldsymbol{\Sigma}_2$ in the text, but have included the syntax that leads to $\boldsymbol{\Sigma}_2$ in Appendix Code 2. All of the intermediate syntax that is used to specify the bifactor model `AIPE.bifactor` is included in Appendix Code 2 as well. Note that, although the factor loading of interest and the desired CI width remain the same as the example in Approach 4 of Empirical Demonstration 1, the model and input covariance matrix are different, and the resulting sample size estimate will likely be different as well. Executing this function `ss.aipe.sem.path` yields $N = 1011$. Next, we plan sample size with respect to κ_1 as follows and set the desired CI width to be .20:

```
> ss.aipe.sem.path(model=AIPE.bifactor, Sigma=Sigma.2,
desired.width=.20, which.path="k1", conf.level=0.95).
```

The necessary sample size estimate is $N = 1491$. Therefore, in order to obtain estimates of both λ_1 and for κ_1 with confidence intervals of desired width, the study needs a sample of size 1,491.

In addition to planning the sample size with respect to the expected CI width, one can include the assurance parameter in the process, so as to ensure a higher probability of observing a sufficiently narrow CI in a particular study. For example, we can use $\gamma = .90$ and plan the sample size again for for λ_1 and for κ_1:

```
> ss.aipe.sem.path(model=model.bifactor, Sigma=Sigma.2,
desired.width=.15, which.path="l1", conf.level=0.95,
assurance=.90)
> ss.aipe.sem.path(model=model.bifactor, Sigma=Sigma.2,
desired.width=.20, which.path="k1", conf.level=0.95,
assurance=.90).
```

The resulting sample size estimates are 1,068 and 1,560, respectively, and thus the study's planned sample size is 1,560.

Conclusion

In this chapter, we have discussed sample size planning in the context of CFA for omnibus and targeted effects from both the power analytic as well as the AIPE approaches. These ideas are general, but we illustrated our ideas specifically in the context of the CFA model, a model that is extremely important in psychometrics and in related fields that use psychometric principles.

We hope that our chapter will be useful for researchers considering a CFA (or even an SEM) study, because the chapter lays out a general framework for considering effect sizes and research goals. Our chapter shows that there is no simple answer to planning sample size for a CFA (or SEM) analysis, as the appropriate sample size necessarily depends on the effect size of interest (e.g., fit index or path coefficient), the goals of the study (to show the existence of an effect or its magnitude), the characteristics of the population under study (e.g., estimated/hypothesized values of covariance), and the desired likelihood of satisfying the goals (e.g., the degree of power or the level of assurance in AIPE). Taken together, we hope that the general framework to sample size planning we have discussed, and the software provided, will facilitate an understanding of the variety of issues involved in sample size planning for CFA.

References

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, *105*, 317–327.

Kelley, K., & Lai, K. (2014). MBESS (Version 3.0.0 and more recent) [computer software and manual], Accessible from http://cran.r-project.org/web/packages/MBESS/index.html.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*, 305–321.

Kelley, K., & Maxwell, S. E. (2008). Power and accuracy for omnibus and targeted effects: Issues of sample size planning with applications to multiple regression. In P. Alasuuta, L. Bickman, & J. Brannen (Eds.), *Handbook of social research methods* (pp. 166–192). Newbury Park, CA: Sage.

Kelley, K., & Maxwell, S. E. (2012). Sample size planning. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (pp. 181–202). Washington, DC: American Psychological Association.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152.

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, *16*, 127–148.

Lin, T. Y., & Weng, L. J. (2014). Graphical extension of sample size planning with AIPE on RMSEA. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 482–490.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19–35.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90.

Venables, W. N., Smith, D. M., & The R Core Team (2014). *An introduction to R*. Available at: https://cran.r-project.org/doc/manuals/R-intro.pdf (accessed September 6, 2017).

Yung. Y., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128.

Code Appendix

Code 1 R Code for Approaches 3 and 4 in Empirical Demonstration 1.

Step 1: Create the input covariance matrix.

We first load the packages MBESS and sem.

```
> library(MBESS)
> library(sem)
```

We start specifying the full model using the function `specifyModel()` in the sem package as follows. Each line of the model specification refers to one model parameter and consists of three parts, separated by comma: (1) to identify the model parameter, (2) parameter name, and (3) start value. For example, the argument "`F1 -> X1, l1, .580`" below means "the model parameter from F_1 to X_1; parameter name l1; start value .580." The argument "`F1 <-> F1, NA, 1`" below means "covariance between F_1 and F_1; not a free parameter; value fixed at 1." The

argument "`X1 <-> X1, e1, .6636`" below means "covariance between error of X_1 and error of X_1 (i.e., error variance); parameter name `e1`; start value .6636."

```
> model.full <- specifyModel()
F1 -> X1, l1, .580
F1 -> X2, l2, 0.720
F1 -> X3, l3, 0.680
F1 -> X4, l4, 0.750
F1 -> X5, l5, .600
F1 <-> F1, NA, 1
X1 <-> X1, e1, .6636
X2 <-> X2, e2, .4816
X3 <-> X3, e3, .5376
X4 <-> X4, e4, .4375
X5 <-> X5, e5, .640
```

We then specify the vector of standardized model parameters $\boldsymbol{\theta}$ as follows.

```
> theta <- c( .580, .720, .680, .750, .600,
.6636, .4816, .5376, .4375, .640)

> names(theta) <- c("l1","l2", "l3","l4","l5",
"e1", "e2","e3","e4", "e5")
```

We then obtain the model-implied correlation matrix $\mathbf{P}(\boldsymbol{\theta})$ given $\boldsymbol{\theta}$ and the full model as follows, using the `theta.2.Sigma.theta()` function in the MBESS package. The argument latent.vars indicates which variables in the model specification model.full are latent variables.

```
> res <- theta.2.Sigma.theta(model=model.full, theta=theta,
latent.vars=c("F1"))
> P.theta <- res$Sigma.theta
```

Given $\mathbf{P}(\boldsymbol{\theta})$ and the standard deviations of manifest variables, we calculate the model-implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The result Sigma.1 below is $\boldsymbol{\Sigma}_1$ we referred to in the text.

```
> Sigma.1 <- cor2cov(P.theta, sd=c(1.2, 1.3, 0.7, 1.2, 1.1))
```

Step 2: Fit the input covariance matrix to the restricted model and obtain the ML discrepancy function value.

Note in the model specification below the factor loadings for X_1 and X_2 have the same name. This places the equality constraint on the two parameters, as the restricted model proposed.

```
> model.res <- specifyModel()
F1 -> X1, l1, .580
F1 -> X2, l1, 0.580
F1 -> X3, l3, 0.680
F1 -> X4, l4, 0.750
F1 -> X5, l5, .600
```

```
F1 <-> F1, NA, 1
X1 <-> X1, e1, .6636
X2 <-> X2, e2, .4816
X3 <-> X3, e3, .5376
X4 <-> X4, e4, .4375
X5 <-> X5, e5, .640
```

Next we fit the restricted model to $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and obtain F^*_{Res} (see Equation 8). Because the model estimating function `sem()` within the sem package requires a sample size but we are interested in the population, we can simply assign an arbitrarily large value to the argument `N` as follows. The object `F.res` contains the value of F^*_{Res}.

```
> res2 <- sem(model=model.res, S=Sigma.1, N=1000001)
> F.res <- res2$criterion
```

Step 3a: Plan the sample size for sufficient statistical power.

Now that we have F^*_{Res}, we can plan the sample size from the power analytic perspective, using Approach 3 explained in Empirical Demonstration 1. We employ the `ss.power.sem()` function in the MBESS package as follows.

```
> ss.power.sem(F.full=0, F.res=F.res, df.full=5,
df.res=6, alpha=.05, power=.80)
```

Step 3b: Plan the sample size for narrow confidence interval.

If the interest is in finding a necessary sample that ensures a sufficiently narrow confidence interval (i.e., Approach 4 in Empirical Demonstration 1), the function `ss.aipe.sem.path` in the MBESS package can be used. The model parameter of interest is λ_1 in the path diagram in Figure 1, or equivalently `l1` in the previous specification of the full model in Step 1. The input covariance matrix $\boldsymbol{\Sigma}_1$ (i.e., `Sigma.1`) was obtained previously in Step 1 as well.

```
> ss.aipe.sem.path(model=model.full, Sigma=Sigma.1,
desired.width=.15, which.path="l1", conf.level=0.95)
```

Code 2 R Code for Approaches 1 and 2 in Empirical Demonstration 2.

This appendix code uses the same functions as does Code 1. Explanations for all the arguments and objects are available in Code 1, except for the ones newly introduced in the present code.

Step 1: Create the input covariance matrix.

```
> library(MBESS)
> library(sem)

> bifactor.std.coef <- specifyModel()
G -> X1, k1, 0.5
G -> X2, k2, 0.5
G -> X3, k3, 0.5
G -> X4, k4, 0.5
G -> X5, k5, 0.5
G -> X6, k6, 0.6
G -> X7, k7, 0.6
G -> X8, k8, 0.6
G -> X9, k9, 0.7
G -> X10, k10, 0.7
G -> X11, k11, 0.7
G -> X12, k12, 0.7
F1 -> X1, l1, .550
F1 -> X2, l2, 0.73
F1 -> X3, l3, 0.68
F1 -> X4, l4, 0.75
F1 -> X5, l5, .6
F2 -> X6, l6, 0.6
F2 -> X7, l7, 0.6
F2 -> X8, l8, 0.6
F3 -> X9, l9, .45
F3 -> X10, l10, 0.45
F3 -> X11, l11, 0.45
F3 -> X12, l12, 0.45
G <-> G, NA, 1
F1 <-> F1, NA, 1
F2 <-> F2, NA, 1
F3 <-> F3, NA, 1
X1 <-> X1, e1, .4475
X2 <-> X2, e2, .2171
X3 <-> X3, e3, .2876
X4 <-> X4, e4, .1875
X5 <-> X5, e5, .39
X6 <-> X6, e6, .28
X7 <-> X7, e7, .28
X8 <-> X8, e8, .28
X9 <-> X9, e9, .3075
```

```
X10 <-> X10, e10, .3075
X11 <-> X11, e11, .3075
X12 <-> X12, e12, .3075

> theta <- c(rep(0.5, 5), rep(.6, 3), rep(.7, 4),
.55, .73, .68, .75, .6, rep(.6,3), rep(.45,4),
.4475, .2171, .2876, .1875, .39, rep(.28,3), rep(.3075, 4))

> names(theta) <- c("k1", "k2", "k3", "k4", "k5", "k6", "k7",
"k8", "k9", "k10", "k11", "k12", "l1", "l2", "l3", "l4", "l5",
"l6", "l7", "l8", "l9", "l10", "l11", "l12", "e1", "e2", "e3",
"e4", "e5", "e6", "e7", "e8", "e9", "e10", "e11", "e12")

> res <- theta.2.Sigma.theta(model=bifactor.std.coef,
theta=theta, latent.vars=c("G", "F1","F2","F3"))

> P.theta.2 <- res$Sigma.theta

> Sigma.2 <- cor2cov(P.theta.2, sd=c(1.2, 1.2, 0.7, 1.3,
1.1, 1.5,1.6,1.7, 2.2,2.3,2.1,2.4))
```

Step 2: Fit the input covariance matrix to the second-order factor model and obtain the ML discrepancy function value.

```
> CFA.2order <- specifyModel()
G -> F1, NA, 1
G -> F2, g2, .4
G -> F3, g3, .4
F1 -> X1, NA, 1
F1 -> X2, l2, 0.6
F1 -> X3, l3, 0.6
F1 -> X4, l4, 0.6
F1 -> X5, l5, .6
F2 -> X6, NA, 1
F2 -> X7, l7, 0.6
F2 -> X8, l8, 0.6
F3 -> X9, NA, 1
F3 -> X10, l10, 0.6
F3 -> X11, l11, 0.6
F3 -> X12, l12, 0.6
G <-> G, r, 1
F1 <-> F1, p1, .5
F2 <-> F2, p2, .5
F3 <-> F3, p3, .5
X1 <-> X1, e1, .7
X2 <-> X2, e2, .7
X3 <-> X3, e3, .7
X4 <-> X4, e4, .7
```

```
X5 <-> X5, e5, .7
X6 <-> X6, e6, .7
X7 <-> X7, e7, .7
X8 <-> X8, e8, .7
X9 <-> X9, e9, .7
X10 <-> X10, e10, .7
X11 <-> X11, e11, .7
X12 <-> X12, e12, .7

> res2 <- sem(model=CFA.2order, S=Sigma.2, N=1000001)

> F.2order <- res2$criterion
```

Step 3a: Plan the sample size so as to obtain statistical power to demonstrate the bifactor model has better fit than the second-order model.

```
> ss.power.sem(F.full=0, F.res=F.2order, df.full=42,
df.res=51, alpha=.05, power=.80)
```

Step 3b: Plan the sample size so as to obtain narrow confidence intervals for bifactor model parameters.

The model specification `AIPE.bifactor` below implies the same model as does `bifactor.std.coef` in Step 1 previously, but uses a more common way to identify the model. The identification `bifactor.std.coef` makes it more convenient to specify model parameters in the standardized metric but is less common for analyzing covariance matrices.

```
> AIPE.bifactor <- specifyModel()
G -> X1, k1, 1
G -> X2, NA, 1
G -> X3, k3, 0.5
G -> X4, k4, 0.5
G -> X5, k5, 0.5
G -> X6, k6, 0.6
G -> X7, k7, 0.6
G -> X8, k8, 0.6
G -> X9, k9, 0.7
G -> X10, k10, 0.7
G -> X11, k11, 0.7
G -> X12, k12, 0.7
F1 -> X1, l1, 1
F1 -> X2, NA, 1
F1 -> X3, l3, 0.68
F1 -> X4, l4, 0.75
F1 -> X5, l5, .6
F2 -> X6, NA, 1
F2 -> X7, l7, 0.6
F2 -> X8, l8, 0.6
```

```
F3 -> X9, NA, 1
F3 -> X10, l10, 0.45
F3 -> X11, l11, 0.45
F3 -> X12, l12, 0.45
G <-> G, r, 1
F1 <-> F1, p1, 1
F2 <-> F2, p2, 1
F3 <-> F3, p3, 1
X1 <-> X1, e1, .4475
X2 <-> X2, e2, .2171
X3 <-> X3, e3, .2876
X4 <-> X4, e4, .1875
X5 <-> X5, e5, .39
X6 <-> X6, e6, .28
X7 <-> X7, e7, .28
X8 <-> X8, e8, .28
X9 <-> X9, e9, .3075
X10 <-> X10, e10, .3075
X11 <-> X11, e11, .3075
X12 <-> X12, e12, .3075

> ss.aipe.sem.path(model=AIPE.bifactor, Sigma=Sigma.2,
desired.width=.15, which.path="l1", conf.level=0.95)

> ss.aipe.sem.path(model=AIPE.bifactor, Sigma=Sigma.2,
desired.width=.20, which.path="k1", conf.level=0.95)
```