6

Factor Analysis

This primer focuses on factor analysis. We assume you have read the section on factor analysis in Chapter 11, but we repeat parts of it here to set context. We also assume you are familiar with multiple regression and have had at least superficial exposure to the method of factor analysis. We first present the basics of factor analysis and distinguish it from principal components analysis, the latter of which is often confused with factor analysis. We then distinguish exploratory from confirmatory factor analysis and discuss the concepts of communalities and uniqueness. We next describe different fit functions that are used when performing factor analysis followed by a discussion of the problem of indeterminacy. After touching upon factor rotation, we discuss how to determine how well a factor model accounts for the correlational pattern among the target variables in question. Finally, we discuss a range of issues relevant to factor analysis, including how to interpret factor loadings, the use of factor scores, major and minor factors, sample size, and the use of dichotomous and ordinal variables in factor analysis.

THE BASICS OF FACTOR ANALYSIS

Factor analysis is an analytic method that seeks to explain the correlations between variables by reference to some unmeasured variable (called a "factor") that is presumed to be a common cause of the measured variables. Figure 6.1 presents an example where the correlations between six variables (X1 though X6) are thought to be due to a single underlying factor, also called a *latent variable*, that impacts each observed variable. Each variable also has unique variance associated with it that is independent of the factor – see the *u*s in Figure 6.1. A classic example of factor analysis is where the observed measures reflect different types of mental abilities of children (such as spatial ability, verbal ability, math ability, and reading ability) and the correlations between them are thought to be due to the common influence of general intelligence (see F in Figure 6.1). A way of thinking about factor analysis is that there exists a "mystery variable" that if it was measured and partialled out of each observed variable, the correlations between the observed variables would all be reduced to zero. In this sense, the factor, aka the mystery variable, explains the variable correlations.

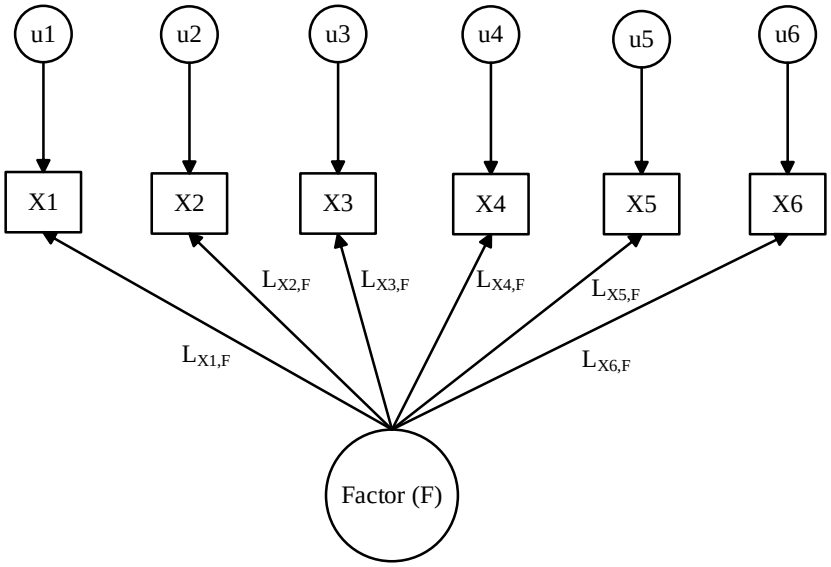

FIGURE 6.1. Single Factor Model

The Ls in Figure 6.1 are called *factor loadings*. They are regression coefficients that reflect the magnitude of the impact of F on each X. They typically are reported in standardized form, so that a loading equal to 0.50 means that for every one standard deviation that F increases, X is predicted to increase, on average, by 0.50 standard deviations.[1] A challenge in factor analysis is to estimate the values of the L given that we have no direct measure of F; without a direct measure of F, we can't actually regress X1 onto F, nor can we regress X2 onto F, and so on. Indeed, we often do not even know what F is substantively. Here, we presumed it was general intelligence.

If a single factor model holds, there should be certain regularities in the correlation matrix for the observed variables and the factor loadings. It can be shown mathematically, for example, that the correlation between any two variables (e.g., X1 and X2) must equal the product of their standardized factor loadings:

$$r_{Xi,Xj} = (L_{Xi,F})\,(L_{Xj,F})$$

Given this, it should be possible to specify a set of values for each of the factor loadings that will reproduce the observed correlations between all the variables considered as a collective. Suppose we specify the factor loading for the first variable is 0.50, for the second variable it is 0.60, and for the third variable it is 0.70. If the single factor model holds, the product of any pair of loadings should reproduce the correlations between that

[1] In this primer, we adopt notation for loadings using conventions for causal paths; the first variable in the subscript for L is the presumed "effect" and the second variable is the presumed "cause."

pair of variables. For example, for the first two variables the product of the two factor loadings, (0.50)(0.60) = 0.30. The product of the loadings for variables 1 and 3 is (0.50)(0.70) = 0.35; and the product of the loadings for variables 2 and 3 equals (0.60)(0.70) = 0.42. The product of the relevant factor loadings for a pair of variables is called a *predicted correlation* for those variables and the actual calculated correlation is called an *observed correlation*. The task of the factor analyst is to find a set of factor loading values that do a good job of reproducing the observed correlations as reflected by the predicted correlations. As long as a one factor model truly operates per Figure 6.1, this should be possible to do, though the task is not necessarily a simple one. If it is not possible to find good fitting values for the factor loadings, then perhaps the one factor model may not be viable. We might need an alternative explanatory model for the correlations, such as a two factor model or a three factor model.

For the case of a two factor model, we again seek to explain the observed correlations between the variables by making reference to underlying factors but now we state there are two latent common causes rather than one. An example appears in Figure 6.2. Although the mathematics are more complicated, the basic logic is similar to that described above: The two factor model is used as a framework to generate a set of predicted correlations between all possible pairs of variables based upon mathematically derived L values. These predicted correlations are then compared with the observed correlations. If there is correspondence between the predicted and observed correlations, this suggests the two factor model is viable. If there are large discrepancies, than a two factor model is called into question and the researcher might then consider a three factor model as a potential descriptor of the underlying theory accounting for the correlations.

Suppose we find a good model fit for a two factor model. We might then conclude that a single factor or “mystery variable” is not sufficient to reduce the correlations among the observed measures to zero if it is partialled out of them. Rather, two such variables are required. That is, if we were able to obtain measures of the two factors and we partialled both of them out of the observed variables, all of the observed correlations would reduce to zero. If a two factor model also failed to account well for the correlations, we might resort to a three factor model to account for them and if that failed, we might then consider a four factor model. A central task of factor analysis is to determine the number of factors that are necessary to adequately account for the correlations among the variables.

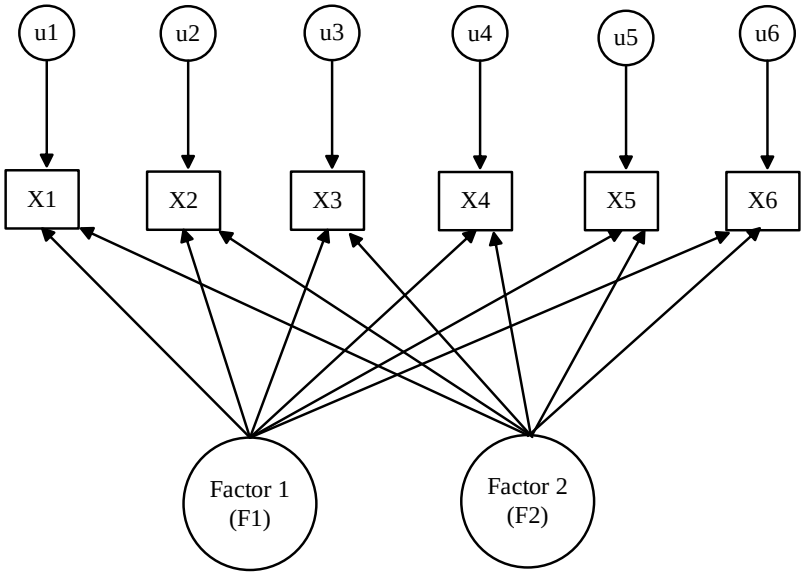

FIGURE 6.3. Two Factor Model

When statisticians try to find values of L that maximize the correspondence between predicted and observed correlations, there are different strategies they use. One strategy seeks to find L values that minimize what is known as an *unweighted least squares* criterion. Values of L are sought that minimize the sum of the squared discrepancies between the predicted and observed correlations. This is analogous to the least squares criterion in multiple regression, but it is applied to predicted and observed correlations rather than predicted and observed scores for individuals. The general term for a minimization criterion is a *fit function*, or alternatively, a *discrepancy function*. The unweighted least squares criterion is one of many fit functions. We discuss others below

The process of finding optimal L values typically uses a trial-and-error approach. An initial set of L values is tried in an attempt to minimize the fit function (in this case, unweighted least squares). Then, another set of L values is tried to see if it improves upon the first set. Then yet another set is tried. At each iteration, one examines the sum of the squared discrepancies between the predicted and observed correlations that the particular L values produced. The search continues until one finds a set of L values that provide the lowest value of the fit function one is likely to find it, i.e., they yield the best reproduction of the observed correlations. These are the set of loadings one uses. The entire search process is called an *iterative process*.

As an example, suppose we fit a one factor model to a 6 variable correlation matrix and the fit function yields the loading estimates in Figure 6.3. Each loading has a value of 0.50 (in practice, the loadings need not be equal). According to the fitted model and the

estimated loadings, the correlation between X1 and X2 should be (0.50)(0.50) = 0.25. The correlation between X1 and X6 should also be 0.25. Indeed, the model predicts that the correlation between every pair of variables should be 0.25. Suppose the actual correlation matrix is that shown in Table 1. You can see that the observed correlations are close in value to the predicted correlations. Stated another way, the data seem to be reasonably consistent with a one factor model. The *residual matrix* is calculated by subtracting each element of the predicted matrix from that of the observed matrix (see Table 1). The entries of the matrix all are near zero, suggesting a good model fit. An important part of evaluating a factor model is to examine the residual matrix

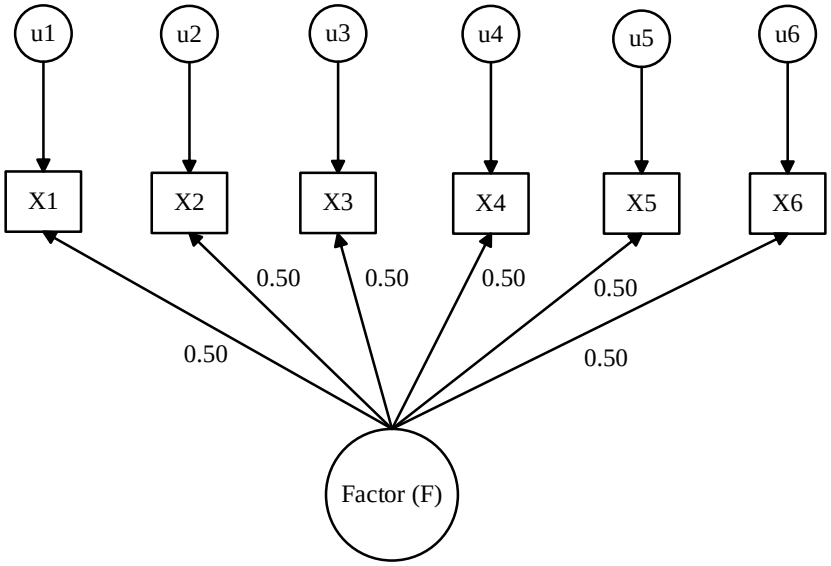

FIGURE 6.3. Results for Single Factor Model

Table 1: Observed and Predicted Correlations

	Observed						Predicted					
	X1	X2	X3	X4	X5	X6	X1	X2	X3	X4	X5	X6
X1	-	0.26	0.25	0.24	0.26	0.25	-	0.25	0.25	0.25	0.25	0.25
X2	0.26	-	0.24	0.26	0.25	0.24	0.25	-	0.25	0.25	0.25	0.25
X3	0.25	0.24	-	0.27	0.25	0.23	0.25	0.25	-	0.25	0.25	0.25
X4	0.24	0.26	0.27	-	0.28	0.22	0.25	0.25	0.25	-	0.25	0.25
X5	0.26	0.25	0.25	0.28	-	0.25	0.25	0.25	0.25	0.25	-	0.26
X6	0.25	0.24	0.23	0.22	0.25	-	0.25	0.25	0.25	0.25	0.25	-

Residual

	X1	X2	X3	X4	X5	X6
X1	-	.01	.00	-.01	.01	0.00
X2	.01	-	-.01	.01	.00	-.01
X3	.00	-.01	-	.02	.00	-.02
X4	-.01	.01	.02	-	.03	-.03
X5	.01	.00	.00	.03	-	.00
X6	.00	-.01	-.02	-.03	.00	-

If we calculate the average value in the residual matrix to index model fit, the result will always be zero or near zero because the positive residuals cancel the negative residuals. A better index takes the form of a *root mean square residual*. This index squares each residual, averages these squared residuals, and then returns the squared average to its original correlation metric by taking the square root of that average. A root mean square residual of 0.05 means that the "average" disparity between the predicted and observed correlations was 0.05. It is just that the average is a different type of average than the arithmetic average you are more familiar with (specifically, it is a positive root mean square average). The root means square residual maps roughly onto the logic of the unweighted least squares fit function. A model with a large root mean square residual is judged problematic and rejected as a good descriptor of the data. Standards differ but root mean square residuals less than 0.05 or so often are deemed reasonable (we return to this issue later). For the above matrix, it is 0.015.

Principal components analysis (PCA) is different from factor analysis in that it does *not* seek to explain correlations between variables. Rather the goal is to reduce the observed measures to a smaller number of linear combinations of them so that the linear combinations or "summaries" can then be used in this reduced form for other analyses. It is a data reduction method whereas factor analysis is a method designed to explain correlations. These goals are different. A diagram for PCA appears in Figure 6.4. In this case, we seek a linear combination of X1 through X6 that captures as much variation in the six measures considered collectively as possible so that we can represent that variation in a more parsimonious way. The linear combination has the form

$$C = L_{C,X1}\,(X1) + L_{C,X2}\,(X2) + L_{C,X3}\,(X3) + L_{C,X4}\,(X4) + L_{C,X5}\,(X5) + L_{C,X6}\,(X6)$$

where C is the summary score for the observed measures. The task is to define weights (the various L) that capture the multivariate variability across individuals in the six Xs.

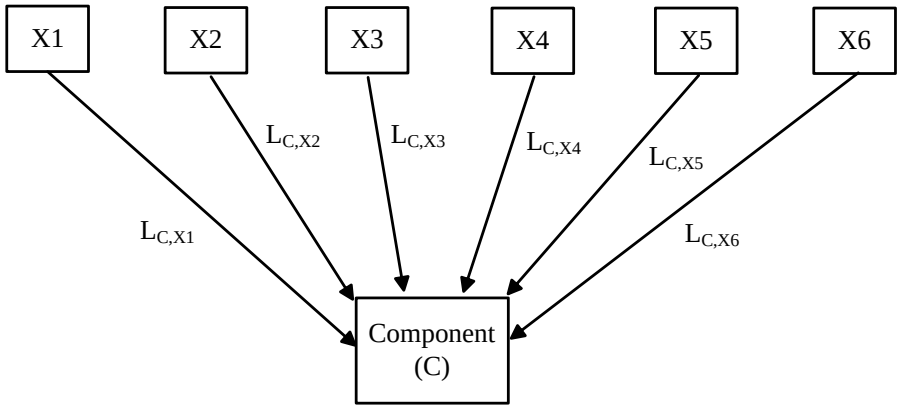

FIGURE 6.4. Principal Components Model with Single Component

Sometimes, capturing the variation in the Xs cannot be well accomplished using a single linear combination of them; too little of that variation is represented by a single component. In such cases, we define a second linear combination of the variables in addition to the first linear combination but this time using a different set of weights. The new weights are derived under the constraint that the scores on the second component be uncorrelated with scores on the second component so as to maximize information gain. This represents a "two component" principal component model rather than a "single component" principal component model. If two components do not capture the multivariate variability sufficiently, a third component might be sought. And so on.

As an example, we might want to develop computerized algorithms for facial recognition based on the measurement of 30 different features of a face, such as the width of the mouth, the width of the eyes, and the pupil to pupil distance. The idea is to match faces of new customers recorded on a camera to a prior data base of pictures of people who have previously entered the store. Rather than working with all 30 scores for each new person who enters the store, it might be more efficient to form three or four linear combinations of those features that take into account their correlational structure and variability. By then working with just these "principal components" of the 30 features, the time it takes to process a face and to match it to a reference list of faces with names can be reduced substantially. We might find that it requires four (uncorrelated) principal components to adequately capture the multivariate representations of the 30 features. We then work with those four derived variables in our computer algorithm.

The focus of this primer is on factor analysis, not principal components analysis.

EXPLORATORY AND CONFIRMATORY FACTOR ANALYSIS

Methodologists distinguish two types of factor analysis, exploratory and confirmatory. In *exploratory factor analysis* (EFA), one seeks to understand the correlations between variables by determining the number and the nature of common-cause variables (factors). The researcher lacks a theory about the number of factors and which variables in the set of X variables the factors influence. The focus is on finding this out. By contrast, in *confirmatory factor analysis* (CFA), researchers have a strong theory that dictates the number of factors that can account for the correlations among the X and whether each factor impacts each observed variable. In CFA, theorists specify a specific pattern of zero and nonzero loadings or they might specify competing models to compare that make such specifications. The goal in CFA is to apply factor analysis to formally test theory.

One way of thinking about EFA is that it is a useful method when the number of plausible factor models for explaining the correlations among the X is so large that it is impractical to test each one in a CFA framework. EFA helps to narrow the field of plausible factor models. It also may be that researchers pursuing CFA could fail to specify plausible alternative models that would be identified in an initial EFA. In this sense, EFA might be applied as a preliminary step to a more formal CFA-like effort to identify plausible models to evaluate (although care must be taken in doing so because such preliminary analyses can affect the statistical theory underlying significance tests when both the EFA and CFA are applied to the same data).

EFA and CFA have been merged into an analytic framework known as *exploratory structural equation modeling* (Asparouhov & Muthén, 2009; Marsh et al., 2010). In this framework, researchers are reasonably confident of the number of factors that underlie a set of X but are uncertain about how each factor impacts the various X. Given the latter uncertainty, paths/loadings are allowed to each X from each factor (like EFA, per Figure 6.2). The factor model is often embedded in a broader theoretical network that includes other covariates, outcomes, and mediators to allow the traditional EFA to be more fully informed by its larger nomological context. Model estimation and evaluation is pursued in the context of this larger, richer network of variables.

POPULATION AND SAMPLE MODELS

Suppose in a population a one factor model perfectly accounts for the correlations between a set of variables. When we select a sample from that population and test a one factor model, the data likely will deviate from perfect model fit because of sampling error, hopefully not by much. Brown and Cudek (1993) and MaCullum (2007) have noted that sometimes a population model will technically be wrong when applied to

population data but the approximation will be quite close. Although population data may not conform exactly to a one factor model, for example, it may be close enough that one can safely use it as a reasonable representation of the true population model. In the words of the statistician George Box "all models are wrong, but some are useful." We return to this point in some depth later, but we want to establish the basic idea here. The use of such approximate population models is somewhat controversial but many argue that working with approximate models in populations is both reasonable and realistic.

COMMUNALITIES AND UNIQUENESS IN FACTOR ANALYSIS

In factor analysis, distinctions are made between common variance and unique variance for each X variable. Common variance (also called *communality*) refers to the variation in a given X that is due to the underlying factor(s). Unique variance (or *uniqueness*) is the variation in a given X that is not due to the underlying factor(s). It represents variance that is "unique" to that variable. The communality and uniqueness indices sum to 1.0 when expressed in standardized form. Their values represent, respectively, the proportion of variation due to the factor(s) and the proportion of variation that is unique. In general, a variable that is uncorrelated or weakly correlated with other variables in the set of X will have high uniqueness and low communality. A variable that is highly correlated with other variables in the set of X will have low uniqueness and high communality.

Factor analysis when applied to sample data estimates the population values for communality and uniqueness for each X. Most factor analysis algorithms require working with an initial estimate of communality for each variable during the iteration process. The initial estimate usually is the squared multiple correlation that predicts X from all the other X in the set. Technically, this squared multiple correlation is not the communality of X per se, but it is a "ballpark" initial estimate that is used during the iterative process. Parenthetically, whereas factor analysis seeks to partition the variance of a variable into unique and common variance, this is not a goal of principal components analysis. PCA makes no distinction between common and unique variance.

In our initial presentation, we stated that a goal of factor analysis is to "explain" correlations between variables. Actually, it has multiple goals with a second goal being to document the communality and uniqueness of each variable in the population. When deriving values of L that best reproduce the correlation matrix, the algorithms also seek to define the L so that they best reflect the true communality of each variable in the population. To accomplish this, the algorithms do not analyze correlation matrices per se. Rather, they analyze correlation matrices with estimates of communalities in place of the ones in the diagonal of the correlation matrix. We do not delve into the underlying mathematics as to why (see Mulaik, 2009). The correlation matrix with communalities in

the diagonals is often referred to as the *reduced correlation matrix*.

Knowing the communality and uniqueness of a variable is important for reasons we develop later. Suffice it to say here that one will orient towards a variable differently on a substantive level if it is relatively distinct from other variables in the variable set (high uniqueness) than if it is dominated by the factors that underlie it (high communality)

When exploratory factor analysis is applied to variables that have low communalities, then analytic complications can result that lead to mischaracterizations of the true population factor model (MacCallum et al., 1999). You should try not to analyze variables that are too dominated by unique variance. In ways, very large amounts of unique variance for a variable suggest the variable may be unrelated to the domain of interest because it has little in common with other variables in that domain. Perhaps such a variable should not be in the theoretical mix. In this sense, it is important for theory and common sense to guide the selection of variables to focus a factor analytic study on.[2]

FIT FUNCTIONS

When deriving values for the factor loadings that best reproduce the communalities and correlations, many different analytic algorithms can be used. For algorithms that use iterations, the choice of loading values on a given iteration is not random. Rather, it is a very intelligent search process that quickly zeros in on the values that will produce the best possible fit to the data given the model. Once it is determined that the best possible fit has been achieved, the iterative process stops. The solution is said to have *converged*. Many computer programs set a limit on the number of iterations that a program can try. Sometimes it is necessary to override this default and set the number of iterations higher. On the other hand, if a program is having difficulty converging on a solution, it may signify that something is amiss with the model being tested.

There are several fit functions one can use to define the factor loadings and evaluate model fit. As noted, the unweighted least squares approach defines values of L that minimize the sum of the squared deviations between predicted and observed correlations. Specifically, we seek to minimize

$$\Sigma\ (r_{ij} - \hat{r}_{ij})^2$$

where r is the observed correlation and $\hat{r}$ is the predicted correlation. This index can be generalized to what is known as a *weighted least squares* formulation that adds a weight to each residual, as follows:

[2] Another source of unwanted unique variance is measure unreliability. Measures with low reliability should generally be avoided in factor analyses.

$$\Sigma\, w_{ij}\,(r_{ij} - \hat{r}_{ij})^2$$

Note that if the weight is set to 1.0 for all residuals, the expression reduces to the unweighted least squares criterion (also called *ordinary least squares*). By introducing weights, one can allow for some residuals to have more influence than others in the calculation of the L. There are different ways of defining the weights but the most common one assigns weights in a way that gives less weight to variables with low communalities. Another weighting scheme, called *generalized least squares*, assigns weights in a way that gives even less weight to variables with low communalities than traditional weighted least squares. For details, see Muthén et al. (1997) and Joreskog and Goldberger (1972). An unweighted least squares approach called *minres* focuses only on the off-diagonal elements of the correlation matrix when deriving estimates of L, i.e., it ignores communalities during the estimation process. To be sure, it ultimately estimates communalities but it does not use communality estimates in the derivation of L. Some methodologists view this as a disadvantage but others see it as an advantage as it circumvents convergence and analytic problems that can arise with more traditional methods (see Revelle, 2015). Minres tends to yield L estimates that are close to the next fit function I discuss, called the *maximum likelihood* fit function.

A very popular fit function is maximum likelihood. In this approach, one minimizes the following:

$$F_{ML} = \ln |S'| - \ln |\Sigma| + \text{trace}[(S)(\Sigma'^{-1})] - k$$

where S is the covariance/correlation matrix, Σ' is the predicted covariance/correlation matrix by the model, *k* is the number of variables in the matrix, ln is the natural log function and | | signifies the determinant of the matrix between the two bars.[3] Although it appears formidable, the maximum likelihood fit function is straightforward if one knows matrix algebra. Consider, for example, the case where there is perfect model fit and S' equals Σ. In this case, the determinant of S will equal the determinant of Σ' and the difference between the logs of these determinants in the first part of the right hand side of the equation will equal 0. If S = Σ', then $(S)(\Sigma'^{-1})$ in the equation equals $(S)(\Sigma^{-1})$. In matrix algebra, any matrix multiplied by its inverse equals an identity matrix, which is a matrix that has the same number of rows and columns of the matrix being operated on but with all 1s in the diagonal and all zeros in the off-diagonal. The trace function sums the diagonal elements of a matrix, the result of which, for a perfect fitting model, will be the sum of the diagonal elements of an identity matrix, the value of which must be *k*.

[3] The determinant of a covariance matrix is a complex set of operations that yield an overall index of multivariate variability. See Namboodiri (1984).

Subtracting *k* from this value yields zero. Thus, when there is perfect model fit, F_{ML} equals zero. As model fit becomes worse, values of F_{ML} become larger.

One reason to minimize F_{ML} is that doing so yields results with many useful statistical properties. For example, with the assumption of multivariate normality among the X, one can calculate significance tests for model fit and standard errors for factor loadings that otherwise are mathematically intractable. We discuss these tests below. Despite its desirable properties, there are cases where maximum likelihood is ill-behaved (Revelle, 2015; Mulaik, 2008) and other methods (e.g., minres) are preferable.

A final method of factor analysis we want to mention is called *principal axis factor analysis*. This approach applies the same statistical algorithms as principal components analysis but uses communality estimates in the diagonal of the correlation matrix (hence, it is formally a factor analytic model). The approach is similar to unweighted least squares (ULS) and minres, but its results are differentially conditional on prior estimates of communalities during the iterative process. Minres generally outperforms it. MacCallum (2009) goes so far as to declare principal axis factor analysis obsolete

Which fit function to use is open to debate. Current wisdom favors the maximum likelihood method because it has a well-developed underlying statistical theory. Maximum likelihood loses some of its advantages when data are non-trivially non-normally distributed, but the method has been found to be reasonably robust to violations of the assumption (Joreskog, 2007). In a simulation study, Olsson et al., (2000) compared maximum likelihood, generalized least squares, and weighted least squares for accurately characterizing model fit and population parameters for different sample sizes, specification error, and non-normality. They found that maximum likelihood as compared to generalized least squares provided more accurate indices of overall fit and less biased parameter values. Weighted least squares, despite recommendations in the literature to use it with non-normal data, never outperformed maximum likelihood or generalized least squares. Similar results were reported by Olson, Foss and Breivik (2004). Some simulation studies have found that ULS/minres recovers population factor structures better than maximum likelihood when the population model is dominated by weak common factors (MaCullum, 2009). However, the accuracy of standard errors and p values for ULS and minres is on weaker grounds compared to maximum likelihood (but see Bentler & Savalei, 2010). All things considered, maximum likelihood usually is the extraction method of choice, but there are exceptions.

MULTIPLE FACTORS AND INDETERMINANCY

In scenarios where one tests a factor model with more than one factor and each factor has a causal path to each X, the factor model is *indeterminant* (or, using different jargon, it is

underidentified). This means there is more than one set of factor loadings, L, that will produce the smallest residual discrepancy index when reproducing the correlation matrix. The issue then becomes which alternative solution to use because they all perform equally well. Statisticians have developed criteria for choosing among the different solutions that are referred to as *identification conditions*. The conditions focus on how loadings relate to the eigenvalues and eigenvectors of the reduced correlation matrix and are designed to maximize certain desirable statistical properties.[4] The conditions include (a) the first common factor must account for the most variance in the X, (b) the second factor must be orthogonal to the first factor, with this property of orthogonality applying to all subsequent factors as well, and (c) the largest eigenvalue must equal the variance explained by the first factor, the next largest eigenvalue must equal the variance explained by the second factor, and so on. Even with such restrictions, there still exist more than one set of loadings that produce the smallest residuals in the residual matrix. At this point, loading interpretability becomes an additional criterion for selecting the final loadings as does the possibility of relaxing the requirement of factor orthogonality, both of which are accomplished using factor rotation, my next topic. For statistical details about identification conditions, see Bollen (1989), Gorsuch (1983) and Harmen (1976).

The indeterminant nature of exploratory factor analysis is an issue that has generated heated debate about the utility of factor analysis more generally. We defer discussion of this issue until after we discuss factor rotation.

FACTOR ROTATION

Suppose we fit a two factor model to a set of data and conclude that it provides good data fit. We can conclude from this that the correlational pattern among the X can be accounted for by two latent factors that serve as common causes of the X. Put another way, there are two unknown variables “out there” that are common causes of the various X and that explain the correlations among them. Our task as a theorist is to identify those two variables.

One approach to guessing the content of the factors is to use substantive considerations. For example, in the case of a single factor model of different types of abilities (verbal, spatial, math,, reading), we might posit the underlying factor that accounts for the correlational pattern among the abilities is general intelligence based on the content of the variables that have been analyzed. Another strategy is to study the pattern of the factor loadings and then deduce what the latent factors must be given the loading pattern. For example, suppose for a two factor solution of symptoms associated

[4] Eigenvalues and eigenvectors are numerical values with certain properties that result from performing operations on a square matrix. It is beyond the scope of this primer to elaborate them – see Namboodiri (1984)

with post-traumatic stress the loadings are such that the first factor strongly impacts symptoms related to reliving the traumatic event (e.g., nightmares, flashbacks) and the second factor impacts symptoms related to avoidance of social situations that might remind one of the traumatic event. The conclusion might be that the two factors are generalized re-experiencing propensities and generalized avoidance propensities, respectively. Such interpretations are easier to make if the pattern of loadings that emerges from the factor analysis is distinctive, such as if each factor strongly impacts a few variables but not others within the various X.

When we "rotate" factors, we do so to try to make the pattern of loadings distinct and more interpretable. In essence, we settle upon a final set of loadings that produce the smallest residuals between the predicted and observed correlation matrices *and* that are easy to interpret and make conceptual sense of. Factor rotation is usually explained geometrically, but the process fundamentally amounts to little more than transforming the initial set of loadings derived under the identification conditions described above so as to now make the loadings more interpretable. A transformation matrix, T, is specified and then multiplied by the loading matrix. The resulting transformed loadings preserve fundamental properties of the initial loading pattern but lend themselves to better interpretation.[5] The challenge for statisticians who evolved factor rotation was to specify criteria for making loadings more interpretable without any knowledge of the substantive area to which the method is applied. Not an easy task! Thurstone (1947) made one of the first such attempts using the idea of *simple structure*. He offered five desiderata:

1. Each variable should have a zero or near zero loading on at least one factor.

2. Each factor should have at least *k* variables with near-zero loadings associated with it, where *k* is the number of factors in the rotation. If there are three factors, each factor should have at least three variables that have large loadings on it

3. There should be variables with non-trivial loadings on at least one factor but near-zero loadings on the other factors.

4. For a given factor, a large proportion of the loadings should be near-zero, at least when the number of factors is large.

5. There should be only a few variables, if any, with non-trivial loadings on more than one factor.

[5] One property not preserved is the percent of shared variance accounted for by a factor. The total amount of shared variance accounted for by the factors remains the same, but it is redistributed among the factors, usually towards greater equality in shared variance accounted for by each factor

These desiderata typically result in a solution where each factor has a few variables that have high loadings associated with the factor. Thurstone argued that factor analyses that achieved such simple structure likely would be more interpretable. Cattell (1978) has elaborated on simple structure and noted that an advantage of seeking it is that factor solutions are more likely to replicate across studies.

Orthogonal and Oblique Rotations

Rotation methods are generally divided into two classes. The first class, called *orthogonal rotations*, maintain the orthogonality of the extracted factors in the initial solution as they seek various desiderata, such as simple structure. The second class, called *oblique rotations*, allows for relaxation of the orthogonality requirement and permits the underlying factors to be correlated when maximizing desiderata. It turns out that under orthogonal rotations, the factor loading for a given variable-factor combination will equal the correlation between the factor and the variable. For example, if the factor loading for variable X_1 on the first factor is 0.30, this is the estimated correlation between X_1 and that factor. With oblique rotations, this is not the case. Because of this property, for oblique rotations, most computer programs report the loadings in the form of path coefficients (that regress the observed variable onto the latent factors), called the *pattern matrix*, as well as the estimated zero-order correlation of each variable with the factor, called the *structure matrix*. In orthogonal rotations, these two matrices are identical. In oblique rotations, they are not. In addition, programs report the estimated correlations between the factors for oblique rotations. If these correlations are very high, then the solution might be called into question.

It is important to keep in mind that oblique rotations do not force the transformed factors to be correlated. If the solution with the best simple structure results from the factors being uncorrelated, then oblique rotations will yield factor correlations that are near-zero and will produce results close to those obtained by orthogonal rotation. In this sense, oblique rotation methods are more general than orthogonal rotation methods and are "data driven." Some have argued that forcing factors to be uncorrelated is theoretically unrealistic; that latent factors are likely correlated in the real world and that analyses should respect this. Studies suggest that when orthogonal rotations are applied to population models where the factors are indeed correlated, distortions in loadings can result (MacCallum, 2009). We recommend you use oblique rotations unless there are strong theoretical reasons not to.

Blind and Targeted Rotations

Another general distinction between factor rotations is that of *blind* versus *targeted*

rotation methods. Blind methods are those in which the objective is to attain simple structure in the spirit of Thurstone but with no specific expectations beyond that. Targeted rotation methods allow the analyst to incorporate into the rotation process *a priori* specifications about the value of loadings for certain targeted loadings. For example, we might target a loading between a given factor and a measured variable to be zero. Most rotation methods offered in popular computer programs are blind.

Popular Methods of Factor Rotation

There are a dizzying number of rotation/transformation methods that have been proposed. We mention a number of them here with the idea you may encounter them in the literature. After doing so, we make some recommendations. Arguments can be made for each rotation/transformation method and detailed consideration of the methods is well beyond the scope of this primer (see Fabrigar et al., 1999; Gorsuch, 1983; Mulaik, 2009). For orthogonal rotations, varimax and quartimax strategies are among the more popular. To explain these methods, consider a matrix of factor loadings where the columns are factors, the rows are the measured variables, and the entries are the loadings (note: we use simplified notation to indicate loadings):

	F1	F2	F3
X1	L_1	L_7	L_{13}
X2	L_2	L_8	L_{14}
X3	L_3	L_9	L_{15}
X4	L_4	L_{10}	L_{16}
X5	L_5	L_{11}	L_{17}
X6	L_6	L_{12}	L_{18}

Varimax strategies emphasize simple structure by seeking a set of loadings that increase initially large loadings within a given column (e.g., L_1 through L_6; L_7 through L_{12}; and L_{13} through L_{18}) and decrease initially small loadings within that column so that each factor has fewer variables with large loadings. *Quartimax* strategies seek to increase initially large loadings within a given row (L_1, L_7 and L_{13}; L_2, L_8 and L_{14}; and so on) and decrease initially small loadings within that row so that each variable will have larger loadings on fewer factors. Varimax rotation seems to be the more popular of the two.

There is no single method of oblique rotation that dominates current usage. Asparouhov and Muthén (2009) report favorable results for the *geomin* rotation method, especially for simple and moderately complicated loading structures. However, they find

geomin has limitations when trying to recover complex loading structures that have three or more factors with many variables that have sizeable loadings on three or more factors.[6] Such scenarios are not common, however, so geomin is often a good choice.

The classic *promax* rotation method is a derivative of varimax algorithms as applied to oblique rotations. On a technical level, it raises factor loadings to different powers for rotation purposes, with higher powers leading to more simplistic rotated structures. A power of 3 often works well. Promax has the advantage of being computationally efficient, but it is generally inferior to newer oblique rotation methods (Browne, 2001). The *simplimax* rotation method (Kiers, 1994) is related to the promax method but uses what is known as *partially specified targeting*. It allows the user to specify *a priori* the number of loadings that should be forced to be near-zero but does not specify which loadings to make near-zero. This is determined empirically during the rotation process. If the specification of the number of near-zero loadings is not realistic relative to the true model, the solution will not converge.

Bentler (1977) proposed a metric invariant *pattern simplicity* rotation criterion that has many desirable properties and that if minimized promotes many simple structure desiderata. It can be applied to both orthogonal and oblique rotations. There have not been many simulation studies to evaluate the approach.

The *oblimin* oblique rotation method is a popular method. A parameter called gamma (sometimes it is called delta) associated with the oblimin method places restrictions on the magnitude of the factor correlations that are allowed.[7] When gamma is set to 0, the method is equivalent to an oblique rotation method called *quartimin*. When set to 0.50 it is equivalent to a rotation method called *biquartimin*. Harmen (1976) recommends setting gamma to zero, which is the typical default in computer programs. Although gamma can be set to numbers greater than 0.80, doing so can introduce analytic complications by allowing factors to be so highly correlated that they are indistinguishable. Of course, this will not always happen because the factor correlations are data driven. If gamma is set to 1.0, the method is equivalent to an oblique rotation approach called *covarimin*.

In an influential paper, Crawford and Ferguson (1970) elaborated a family of rotation methods that are commonly referred to as the *Crawford-Ferguson family* or CF-methods. There are two sub-classes, one for orthogonal rotations and one for oblique rotations. We concentrate here on the CF-oblique family. Crawford and Ferguson derived a numerical index of loading pattern complexity, called *kappa*, that includes a term that

[6] There exists an orthogonal variant of geomin rotation but its performance is not well documented. Geomin was developed primarily as an oblique rotation method.

[7] Gamma does not reflect a correlation coefficient nor is it in correlation units. However, its value governs the magnitude of correlations allowed between factors.

reflects the size of loadings within variables but across factors and a term reflecting the magnitude of loadings within factors but across variables (see the above matrix). By setting different values of kappa prior to rotating the factors, one emphasizes different facets of simple structure during the rotation process. For example, setting kappa to 1/k (where k is the number of measured variables) and then executing the CF algorithm, one obtains an oblique version of varimax criteria. Setting kappa to 0 produces a quartimin rotation. For details of the approach, see Browne (2001).

Recommendations and Additional Comments

Of the many rotation method available, we tend to prefer the geomin method and recommend it coupled with maximum likelihood extraction. The CF-varimax (oblique) and CF-quartimin (oblique) rotation methods, coupled with maximum likelihood extraction, also have much to recommend. The advantage of using maximum likelihood and any of these three rotation methods is that they are well grounded in a statistical theory that permits one to generate standard errors, confidence intervals, and margins of error for the loadings and factor correlations as well as providing indices of model fit (described below).

Factor rotation is common in exploratory factor analysis but it is not used in confirmatory factor analysis. This is because exploratory factor analytic models are inherently under-identified (i.e., indeterminant) hence rotation is possible. Confirmatory factor analytic models generally are not under-identified. Factor rotation makes no sense in confirmatory factor analysis as there is one and only one solution for reproducing the correlations. Factor rotation also does not apply to single factor models. This is because for single factor models, there is a unique solution for loadings to reproduce the correlation matrix, i.e., one factor models are not indeterminant

Many methodologists are uncomfortable with the indeterminacy of factor analysis and do not find rotation strategies to be satisfactory resolutions to the problem. They argue that the ultimate choice of a final set of loadings is arbitrary because it is driven by arbitrary statistical criteria, such as simple structure. In this sense, classic exploratory factor analysis is a controversial method that has been the subject of considerable debate about whether it is ever an appropriate approach to data (see, for example, Heim, 1975; Steiger & Schonemann, 1999; Lovie & Lovie, 1995).

MODEL FIT

Once a model with a given number of factors has been fit to the data, we need to evaluate how well it fits, i.e., how well it reproduced the correlation matrix. We discuss in the next

section the problem of how to choose the number of factors for a factor model. However, we need to discuss evaluating model fit first because, ultimately, choosing the number of factors is a question of evaluating model fit for models with differing numbers of factors and then comparing them. We describe a wide range of model fit indices to give you a sense of ones you might encounter. We then make recommendations, accordingly.

Root Mean Square Residual

One obvious way to evaluate model fit is to examine the residual correlation matrix to ensure that its values are near zero. A rough index of fit is to calculate the average (root mean square) residual of the off diagonals of the residual matrix. As a rough rule of thumb, statisticians suggest it should not be larger than 0.05 to 0.08, but this standard can shift depending on context. We mentioned this index earlier as the root mean squared residual. In the CFA literature, it is called the standardized root mean squared error and is calculated a bit differently.

Chi Square Test of Fit

When maximum likelihood extraction is used and if the population X scores are approximately multivariately normally distributed, then under the assumption of a zero residual correlation matrix in the population, the value of F_{ML} times N-1 is chi square distributed with degrees of freedom equal to

$$df = ((0.5)(k)(k + 1) + k) - t$$

where *k* is the number of measured variables in the input matrix and *t* is the number of parameters estimated in the model. This property allows us to compute a p value for the chi square statistic to evaluate the following null and alternative hypotheses:

H_0: The population residual correlation matrix is a zero matrix

H_1: The population residual correlation matrix is not a zero matrix.

If the chi square test is statistically significant ($p < 0.05$), then the null hypothesis of perfect model fit in the population is rejected and the model is called into question. If the chi square is statistically non-significant, then there may or may not be perfect model fit in the population. Given this, a model is said to be viable if it yields a statistically non-significant chi square test.

Parenthetically, whereas F_{ML} often is multiplied by N-1 to produce the chi square statistic, a small sample correction has been suggested that multiplies F_{ML} instead by N -

(2k + 5)/6 – 2m/5, where *k* is the number of measured variables and *m* is the number of factors (Bartlett, 1950). The latter is commonly used in software for exploratory factor analysis whereas software for confirmatory factor analysis tends to use N-1 as the multiplier. With large N, the difference between the approaches is minimal

The chi square test statistic and its associated test of significance have been criticized on several grounds. First, the statistic is not always chi square distributed for purposes of testing statistical significance, especially for small sample sizes and non-normal data (but see Joreskog, 2007, for its robustness properties). In such cases, the p values associated with it may not be accurate. Second, like most statistical tests, the p value is influenced by sample size; larger sample sizes produce smaller p values, everything else being equal. As such, model evaluations with large sample sizes will tend to lead to model rejection given it is inevitable the population residual matrix will not be exactly zero. Kenny (2015) suggests the test is most meaningful for sample sizes between 75 and 250, but there is controversy surrounding this. Third, the value of chi square is affected by the size of correlations between variables: The larger the correlations, the larger the chi square tends to be, everything else being equal (Kenny et al., 2015).

The Root Mean Square Error of Approximation

Another fit index for the factor model is called the *root mean square error of approximation* (RMSEA) developed by Steiger and Lind (1980) and elaborated by Browne and Cudek (1993). It is formally defined in the population as

$$\varepsilon = (F_0 / df)^{1/2} \qquad [1]$$

where F_0 is a generic discrepancy function reflecting disparities between the predicted and observed correlations and df is the associated degrees of freedom. F_{ML} is typically used as F_0 in the above definition in conjunction with the formula for df given above. Conceptually, ε is an index of the lack of model fit (as reflected by F_0) per model degree of freedom, because F_0 is divided by df.

The above formula is the population representation of the RMSEA. If we work with F_{ML} as our discrepancy function, statisticians have shown that the sample value of F_{ML} is a positively biased estimator of the corresponding population value of F_{ML}. A correction factor has been suggested to eliminate the bias as follows:

$$\hat{F}_{ML} = F_{ML} - m$$

where *m* is the correction factor and equals df/(N-1) and $\hat{F}_{ML}$ is the sample estimate of the population F_{ML}. If $\hat{F}_{ML}$ is negative, it is set to 0.

The smallest value RMSEA can take is 0 and the largest value it can have is infinity, although it rarely exceeds 1.00. The smaller the value of RMSEA, the better the model fit. More parsimonious models have larger degrees of freedom, so the presence of df in the equation acts as a penalty function for lack of parsimony. Browne and Cudek (1993) suggest that, as a rule of thumb, RMSEA values less than 0.08 imply adequate model fit and values less than 0.05 imply good model fit.

The Test of Close Fit

Browne and Cudek (1993) also argue that the p value for the traditional chi square test of model fit is too stringent because, as noted, it tests for perfect model fit. Perfect model fit is often seen as unrealistic. Browne and Cudek devised an inferential test for a "close" fitting population model, where “close” is defined as a population RMSEA value of 0.05 or less. Thus, the null and alternative hypotheses for the test are

H_0: $\varepsilon \leq 0.05$

H_1: $\varepsilon > 0.05$

where ε is the population RMSEA. By contrast, the chi square test translates into a test of

H_0: $\varepsilon = 0.00$

H_1: $\varepsilon > 0.00$

Focusing on the first set of null and alternative hypotheses, if the p value for the *test for close fit* is non-significant ($p > 0.05$), then this is consistent with the presence of a close fitting model in the population. By contrast, a statistically significant p value for close fit ($p < 0.05$) leads one to reject the null hypothesis and conclude the model as applied in the population does not yield a close fit.

The test of close fit also can be implemented by calculating a 90% confidence interval for RMSEA. If the lower limit of the interval for the estimate of ε is less than 0.05, than this is the same as obtaining a non-significant test of close fit. If the lower limit is greater than 0.05, then this is the same as obtaining a statistically significant p value for close fit. A 90% confidence interval is used because of the one-sided nature of the test. An even more demanding test is to evaluate if the upper confidence interval of the RMSEA is less than 0.05. This would lead one to be quite confident that the model is a close fitting model in the population. For details about this latter test, see MacCallum et al. (1996). MacCallum et al. (1996) refer to this latter test as the *test of not close fit.*

The Comparative Fit Index

Another fit index is the *comparative fit index* (CFI), or its close counterpart, the Tucker-Lewis Index (TLI). The CFI compares the fit of the target model with the fit of a competing model known as the "independence" or "null" model. The "null model" is one that posits no correlation between any of the observed variables. Such a model is not very viable in most research situations, so one expects that a “good” model will fit quite a bit better than it. The CFI ranges from 0 to 1.0, with larger values implying the target model fits better than the null model. A CFI of 0.90 means the target model fits 90% better than the null model. A CFI of 0.70 means the target model fits 70% better than the null model. A CFI of 1.00 means the target model fits at least 100% better than the null model. And so on. A rule of thumb is that models with a CFI less than 0.95 are suspect.

The formal definition of the CFI is as follows: Let d_M = the χ^2 value for the target model minus its degrees of freedom and d_I = the χ^2 value for the independence model minus its degrees of freedom. Then

$$CFI = (d_M - d_I) / d_I$$

If the index is greater than one, it is set to one. Note that the CFI scales the improvement in fit of the model relative to the independence model (the numerator) against the lack of fit of the independence model (the denominator).

It turns out that the CFI is affected by the size of the correlations between variables in the population correlation matrix, which is a bothersome property. If all of the population correlations are, in fact, low, then the null/independence model will fit the data well and the target model cannot improve much on it. Kenny et al. (2015) recommend not using the CFI (or TLI) if the RMSEA for the null model (a model that presumes all the correlations are zero) is less than 0.16, because the CFI will tend to be artificially low.

Recommendations

In sum, there are a wide range of model fit indices to choose from. No single index of global fit is perfect. All have strengths and weaknesses. We like to examine model fit from multiple perspectives before making a judgment about model fit. As such, we examine the residual correlation matrix, the root mean square of the off-diagonals of the residual matrix, the chi square test of the residual matrix, the RMSEA (and its associated confidence intervals for the close fit test), and the CFI. Examination of these fit indices usually gives one a good sense of the fit of the model. See the worked example associated with this primer for a substantive application.

CHOOSING THE NUMBER OF FACTORS

A critical decision when conducting an exploratory factor analysis is to decide how many factors are needed to adequately account for the correlations between variables. One of the more popular rules for making this decision is to retain factors that have eigenvalues greater than 1.0 (Fabrigar et al., 1999). Statisticians have found the rule to be sub-optimal despite its common use by applied researchers. One problem is that the criterion was proposed for the case of principal components analysis not factor analysis, the latter for which it is not appropriate (Fabrigar et al., 1999). Another problem is that it can lead to arbitrary decisions, such as when one factor has an eigenvalue of 1.01 (hence it is retained) and another has an eigenvalue of 0.99 (hence it is not retained). Finally, the rule has not performed well at detecting the true number of factors in a large number of simulation studies (Zwick & Velicer, 1986). We recommend against this approach.

In this section, we review a range of approaches you may encounter for determining the number of factors to adopt in a factor analysis. After doing so, we make some recommendations from among them.

Percent of Variance Accounted For

One approach is to base the decision on the percent of variance that a factor accounts for, retaining only those factors that account for a reasonable percent of the variance (which often is considered to be about 10%). There are two indices of percent of variance accounted for one can use in this approach. The first is the percent of variance accounted for as derived from the eigenvalues of the correlation matrix. This provides perspectives on the percent of *total* variance of the various X accounted for by each possible factor. The second is the percent of variance accounted for as derived from the eigenvalues of the reduced correlation matrix. This provides perspectives on the percent of *common variance* accounted for by each possible factor, ignoring the unique variance. Most computer software does not report the latter percent. This is because the reduced correlation matrix can occasionally be non-positive definite, meaning it will have negative variances, which are nonsensical. Nevertheless, some researchers find it useful to examine the eigenvalues of the reduced correlation matrix with the idea of retaining factors that account for nontrivial amounts of the shared variance as opposed to the total variance.

The Scree Test

Related to the above is an approach known as the *scree test*. This test focuses on the pattern of eigenvalues associated with the factors using the fact that the eigenvalues of

each successive factor will be lower than the factor before it. The researcher makes a subjective judgment about where there is a large “break” in the magnitude of the eigenvalues from one factor to the next. The test is often implemented using a scree plot with the factors on the X axis and the corresponding eigenvalues on the Y-axis. As one moves to the right the eigenvalues decrease in value and often form an “elbow” reflecting a less steep decline. As the values “flatten out” after the elbow, one chooses the factor number just before the beginning of the flattening. Figure 6.5 presents an example that would lead to a two factor choice. This plot provides both the eigenvalues associated with the correlation matrix (labeled PC) and the eigenvalues associated with the reduced correlation matrix (labeled FA). Both show a clear elbow at factor 3, so the choice is for 2 factors. This test has been criticized because of its subjective nature and because the graph often fails to show an obvious “elbow.”

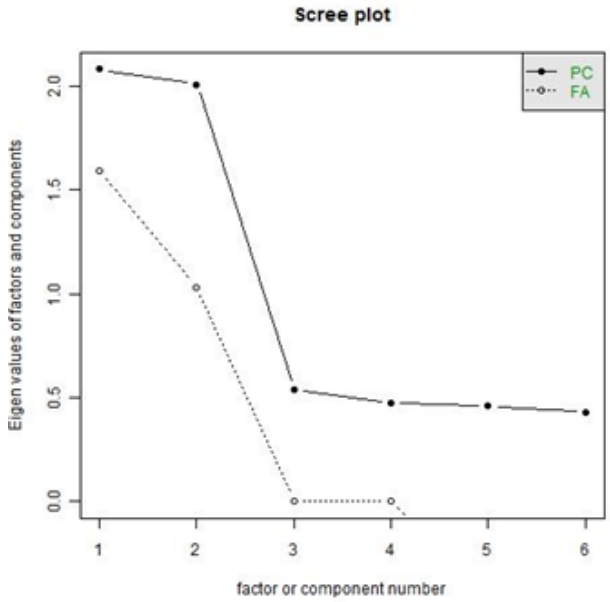

FIGURE 6.5. A Scree Test

The Minimum Average Partial Method

Another method for choosing the number of factors that is often used is called the *minimum average partial* (MAP) method (Velicer, 1976). It estimates the average squared partial correlation between the measured variables after holding constant successive factors and chooses the number of factors by identifying the point where the partial correlation is lowest (Recall that in a good fitting factor model, if one partials out

all the factors, the correlations between the X variables should be near zero). The test makes use of the fact that the average partial correlation ultimately will start to increase as one continues to partial out factors/components because of characteristics of statistical suppression inherent within the strategy. Unfortunately, this method only works for principal components analysis; for factor analysis, the average partial correlation will continually decrease (see Velicer, 1976, for details) and searching for the point where it begins to increase is futile. As such, we do not consider it further here. It is not really a factor analytic diagnostic, in a strict sense – although some researchers erroneously use it as such.

Parallel Analysis

Yet another method used to identify the number of factors is called *parallel analysis* (PA). Parallel analysis compares the eigenvalues of the observed correlation matrix to eigenvalues of randomly generated samples from a comparable population correlation matrix based on normal distributions. The rationale is that if nonrandom factors exist, then eigenvalues generated from the real data will be larger than those from the randomly generated eigenvalues. Only factors associated with empirical eigenvalues consistently larger than eigenvalues based on the randomly generated data are retained. There is a variant of PA for principal components analysis and for factor analysis, so one must be careful to use the proper method (Dinno, 2010). Simulation studies have generally found positive support for the approach (Zwick & Velicer, 1986).

Chi Square Difference Tests

In contrast to the above strategies, Preacher, Zhang, Kim and Mels (2013) argue for a model selection approach that formally compares the fit of competing models that vary the number of factors, an approach we think is best. The idea is to conduct an exploratory factor analysis for, say, a one factor solution, then a two factor solution, then a three factor solution, and so on, followed by formal tests of comparative fit between them.

One strategy for comparing models when using maximum likelihood extraction is to conduct chi square difference tests between models that differ in their number of factors. In general, the model with fewer factors is called the *constrained model* and the model with more factors is called the *unconstrained model*. We want to choose between them and determine if the unconstrained model improves on the fit of the constrained model in the population. We can frame the null hypothesis as follows:

H_0: $Fit_A - Fit_B = 0$

where Fit_A is the population fit for the constrained model and Fit_B is the population fit for unconstrained model. The hypothesis is that there is no difference in fit; the model with more factors does not increase the fit at all. The alternative hypothesis has the form:

$H_1: Fit_A - Fit_B > 0$

The formal significance test can be executed by forming the difference in the two chi squares associated with each model, as follows:

$\chi^2_{Diff} = \chi^2_{Constrained} - \chi^2_{Unconstrained}$

which itself is chi square distributed with degrees of freedom equal to $df_{Constrained} - df_{Unconstrained}$. We can then calculate a p value to determine if we reject the null hypothesis of no incremental fit. If the p value is less than 0.05 (or some other *a priori* specified alpha), the null hypothesis of equal model fit is rejected. This test has been criticized because it can be sample size sensitive and because the null hypothesis of exact model equivalence is often unrealistic. It also can yield invalid conclusions with small sample sizes and can lack statistical power when the unconstrained model is itself a poor fitting model (Yuan & Bentler, 2004).

RMSEA Tests

MacCallum, Browne and Cai (2006; see also MacCallum, Lee & Browne, 2010) describe a more general method for nested model comparisons. Their approach makes use of the RMSEA fit index and compares the RMSEAs for the constrained and unconstrained models in the spirit of the chi square difference test described above. However, whereas the chi square difference test evaluates the null hypothesis of equal model fit versus the alternative hypothesis of non-equal model fit, McCallum et al.'s framework allows one to evaluate a null hypothesis of a small difference in model fit (rather than equivalent fit) against an alternative hypothesis of a fit difference greater than that small difference. The underlying premise of the test is that asking if competing models have exactly the same fit per the traditional chi square difference test will almost always be false, so a test of it is moot (and driven, in part, by sample size). It is more realistic, they argue, to work from a null hypothesis that there is a small difference between models but it is so small that the two models can be viewed as being *functionally equivalent.* When the traditional chi square test is expressed using RMSEAs, the null and alternative hypotheses are

$H_0: \varepsilon_A - \varepsilon_B = 0$

$H_1: \varepsilon_A - \varepsilon_B > 0$

where ε_A is the population RMSEA for the constrained model and ε_B is the population RMSEA for the unconstrained model. McCallum et al. suggest instead testing

$H_0: \varepsilon_A - \varepsilon_B \leq d$

$H_1: \varepsilon_A - \varepsilon_B > d$

where *d* is a non-negative constant. We can set *d* to any value, but a typical value is 0.05, which, as noted above, is the standard criterion for "close fit."

It turns out that McCallum et al. approach requires we specify the individual values of ε_A and ε_B in the null hypothesis rather than a single value of *d*. This is because an RMSEA difference of, say, 0.01 when the values of the RMSEA are 0.03 and 0.04 implies a different amount of fit disparity between models than, say, a difference of 0.05 and 0.06. Liu and Bentler (2011) propose a modification to the MacCallum et al. method that allows one to specify just *d* and not the individual RMSEAs that define *d*.

Related to the RMSEA approach just described is a somewhat easier method suggested by Preacher et al. (2013). Based on an extensive simulation study, these authors recommend using the RMSEA and its confidence interval to make decisions about the number of factors to retain. Starting with, say, a one factor solution, the analyst successively increases the number of factors until one finds the first model that has a lower bound RMSEA 90% confidence interval value less than the close fit RMSEA standard of 0.05. This model then reflects the number of factors to retain. The logic is grounded in the test of close fit logic described earlier.

Information Theory Approaches

Yet another model comparison approach that has merit is based on information theory using either Akaike's Information Criterion (AIC) or a Bayesian Information Criterion (BIC). These indices are discussed in the primer on regression mixture modeling. The section from that primer with minor edits is appended to this primer for easier access. The current appendix includes a discussion of a variant of the BIC that is particularly relevant to factor analysis, called Haughton's BIC.

CFI Based Approaches

A final approach to choosing between models with differing numbers of factors is to use the CFI index discussed earlier. Recall that the CFI compares two models, the target model and the independence model. It scales the proportion of improvement in fit of the target model relative to the independence model. The index can be used for the two

models with differing number of factors, but where the chi square for the constrained model (the one with fewer factors) and its associated degrees of freedom is used in place of the independence model and the unconstrained model (the one with more factors) and its associated degrees of freedom is used as the target model. The CFI will then index the degree of improvement in fit that the model with more factors provides relative to the model with fewer factors. For example, a CFI of 0.25 indicates that the model with more factors improved fit by 25% relative to the model with fewer factors. In this scenario, one obviously would not use the standard rule-of-thumb of 0.95 to declare one model fits better than another because this standard presumes comparison with the independence model, which is unrealistic to begin with

Recommendations

With such a large array of approaches to determine the number of factors, which one should a researcher use? Each method has strengths and weaknesses. No single index is best. Our preference is to evaluate the question from multiple perspectives and hope that the conclusions converge across different methods. As such, we use multiple tests to make decisions and are most confident in decisions where convergence is evident. When different tests lead to different conclusions, we tend to give greater weight to the less subjective methods that are accommodating of sampling error. We lean towards use of the Preacher et al. (2013) strategy based on RMSEAs for its simplicity and parallel analysis, but, as you will see by consulting the worked example for factor analysis, we often take a broader perspective than just focusing on these two approaches. We also place a premium on factor interpretability, preferring solutions that make the most conceptual sense.

INTERPRETING THE LOADINGS AND COMMUNALITIES

A factor loading is a standardized regression coefficient. It is the number of standard deviations that a measured variable is predicted to change given a one standard deviation increase in the underlying factor, holding constant other factors in the model. As noted, one task of a researcher is to identify what the unmeasured factors that explain the correlations between measures represent. Of particular interest are factor loadings that are large for a given factor because such loadings serve as clues about what the factor might be. The question then becomes, what is a “large” loading?

Many rules of thumb have been offered. One rule of thumb often advocated is if the absolute value of the standardized loading is greater than 0.30, then the variable is “relevant” for that factor. Another rule-of-thumb terms loadings as "weak" if they are less

than 0.40, “moderate” if they are between 0.40 and 0.60, and "strong" if they are more than 0.60. These guidelines, of course, are arbitrary and should be treated as such. They also ignore the fact that there is sampling error in loading estimates and that large margins of error (MOE) or wide confidence intervals may be associated with them. Ideally one will take this into account. Unfortunately, the calculation of MOEs and confidence intervals has been challenging for factor loadings in exploratory factor analysis because they often are mathematically intractable. We discuss this more below.

A property of factor analysis that is underappreciated is that a good fitting model can be found even when the measured variables are only modestly correlated and the factor loadings are small. To use a simplistic example, suppose a single factor model is applicable in a population and has standardized factor loadings for each measured variable equal to 0.40. As noted, the correlation between two variables in a valid one factor model equals the product of the loadings of the two variables. In this case, all of the variables will be correlated (0.40)(0.40) = 0.16 in the population. If the sample data reasonably reflect this pattern (which should be the case with a sufficiently large N), the fit indices described above will point to good model fit. There will be a low chi square, a low RMSEA, and the residual correlation matrix will have entries close to zero. This is because the goal of factor analysis is to *explain the correlations* between measured variables irrespective of the magnitude of those correlations. Note also that the rule of thumb that a factor loading of 0.30 or greater is meaningful takes on a somewhat different meaning when one appreciates that it can be associated with measured variable correlations as low as (0.30)(0.30) = 0.09 for two variables “loading” on the same factor. Such variables have very little in common by virtue of their low correlation.

A statistic of interest in factor analyses is the uniqueness associated with each measured variable because this indicates the unique variance contained in the variable independent of the underlying factors. In a single factor model, for example, a variable with a factor loading of 0.70 has over 50% unique variance.[8] We believe many researchers are too quick to dispense with unique variance contained in measured variables by focusing attention only on common factors or factor scores designed to represent those factors. When we focus our theorizing on common factors (e.g., by averaging scores across variables that load highly on a given factor), we are essentially trivializing the unique variance associated with those variables, which may be substantively counterproductive. In research that we conduct, we often measure what people view as the advantages and disadvantages of engaging in a behavior using single item measures of each perception (e.g., the perceived likelihood that voting for a given candidate will lead to advantage A, that it will lead to advantage B, that it will lead to

[8] In a one factor model, (one minus the squared factor loading times 100), which is considerable.

disadvantage C, and so on). We seek to understand which particular advantages and which particular disadvantages account for most of the variation in decisions to engage in the behavior. Time and again, reviewers of our research are critical that we do not factor analyze the perception ratings and focus on the underlying common factors that emerge from the analysis. This practice ignores the unique variance in the variables and it is often the unique variance that is most predictive of behavior, not the common variance. If the data suggest the measures are dominated by common variance, then it might make sense to focus on it; but if there are non-trivial amounts of unique variance operating, then we should not automatically ignore it for the sake of factor analysis.

On computer output, the way to determine the contribution of unique variance to a measured variable is to examine its standardized uniqueness value. One minus it will equal the contribution of common variance (i.e., the factors) to the measure. For example, if the standardized uniqueness is 0.65, then 65% of the variation is unique to the measure and 35% is due to the underlying factors

We close out this section by returning to the problem of calculating margins of errors of factor loadings to describe technical issues in doing so. Non-interested readers can skip to the next section. As noted, for many forms of factor analysis, the calculation of standard errors for factor loadings is difficult because they are mathematically intractable. Faced with intractable standard errors, statisticians often resort to bootstrapping to estimate them. Bootstrap applications to factor loadings in exploratory factor analysis (not confirmatory factor analysis) are problematic because depending on sample-to-sample fluctuations, factors can change their ordinal positions and loadings can change signs as these ordinal positions shift. This will artificially inflate bootstrapped estimates of the standard errors. To apply bootstrapping, the correlation matrices from each subsample must be aligned before standard errors are computed (Zhang, 2014). Creating such alignment is a non-trivial analytic challenge. Zhang (2014) reviewed approaches for estimating standard errors of factor loadings. A general but somewhat ad hoc method was offered by Cattell (1988, p. 192), but it is approximate. The preferred method is the *infinitesimal jackknife method* (Zhang, Preacher & Jennrich, 2012), but it is challenging to implement. Some statisticians use a stabilizing transformation when converting standard errors to confidence intervals for factor correlations (SAS Institute, 2010). In general, one should focus on MOEs and confidence intervals only after a good fitting model has been settled upon. It makes little sense to calculate MOEs and confidence intervals for a model that is inconsistent with the data.

FACTOR SCORES

Given a satisfactory factor model, many researchers seek to calculate scores on each

factor for each individual, usually with the idea of then relating those scores to some other construct, be it a hypothesized outcome of the factor or a hypothesized predictor of the factor. Technically, we cannot know individuals' scores on a factor, we can only estimate them. Some scoring strategies yield better estimates of the true factor scores than others. When deriving factor score estimates, we want them to be highly correlated with the true factor scores and we want them to reflect the essential properties of the factor model on which they are based. The estimation of factor scores has been debated for decades and is quite controversial. In this section, we first describe the nature of the controversy and then address strategies that have been used to estimate factor scores in practice. We then suggest resolutions to the estimation dilemma.

Factor Score Indeterminacy

The central issue that has plagued factor score estimation in exploratory factor analysis is that of indeterminacy, i.e., there typically are more than one set of factor scores that satisfy the required properties of the factor model and we do not know which scores to use unless we make additional assumptions - assumptions that critics argue are ad hoc (Schonemann & Steiger, 1978; Mulaik, 2005). If one of the factors has been determined to be a latent construct of anxiety, for example, the same individual who has a large anxiety factor score estimate for one set of factor scores might have a low anxiety factor score estimate for another equally valid set of factor scores that satisfy the requirements of the factor model as well as the first set. One such requirement is that the estimated scores for the different factors be uncorrelated if an orthogonal rotation method is used or reproduce well the correlations between factors if an oblique rotation method is used. As well, if we regress a given observed measure, X, onto the factor scores, we should obtain factor loadings for the observed measures that are consistent with those of the adopted factor model. Finally, within the context of such regression analyses, we should obtain estimates of common and unique variance for a given X that reflect those derived from the factor model.

Here is a way of thinking about factor score indeterminacy: Suppose we learn from a factor analysis that an observed variable, X, is correlated 0.80 with an underlying factor. We know what each individual's score is on X because X is directly measured. We do not, however, know what the individual's score on F because F is unmeasured. There are many different ways we could arrange possible scores of individuals on F to produce a correlation of 0.80 with X. Which of the many possible patterns is the correct one? This is the essence of factor score indeterminacy.

The degree of factor score indeterminacy in a given factor analysis can vary; in some cases there exist many reasonable alternative sets of factor score estimates that

satisfy the factor model and in other cases, the number is more limited. Statisticians have proposed indices of the degree of indeterminacy that likely exists in order to help researchers appreciate the magnitude of the problem in a given application (Grice, 2001). However, these indices have been questioned and controversial at best (Mulaik, 2005).

Methods for Estimating Factor Scores

Two general approaches have evolved for estimating factor scores. Grice (2001) refers to them as the refined and coarse approaches to estimation. The *refined approaches* make use of all of the observed measures (in standardized form) and all of the factor loadings associated with each measure. Three such methods are popular, the least squares regression method, the Bartlett method, and the Anderson-Rubin method. Each calculates a set of factor score coefficients, one for each observed variable, that are then multiplied by the standardized observed scores for an individual on the X. These products are then summed to yield the individual's factor score estimate. The methods differ in how they define the factor score coefficients and the information used to calculate those coefficients. Each approach has strengths and weaknesses and each differs in which properties of the factor model are preserved. For statistical details, see Grice (2001).

The *coarse methods* do not seek to preserve the formal properties of the factor model as rigorously as the refined methods. One commonly used coarse strategy is to identify the X measures that have a factor loading above a given threshold (e.g., 0.40) for a given factor and then to sum or average the standardized scores for all Xs that exceed that threshold in order to define the factor score (using reverse scoring for Xs with loadings of opposite sign). If all the X are on a common metric (e.g., all are on a 0 to 10 scale), some researchers sum or average the raw scores rather than using the standardized scores. Averaging raw scores preserves the differences in standard deviations across the Xs whereas standardization makes the standard deviations for each X uniform (and equal to 1.0). Variations of this strategy include (1) allowing a given X to be used for only one factor, even if it exceeds the inclusion threshold on several factors, (2) dropping an X that exceeds the inclusion threshold on more than one factor, and (c) using the factor score coefficient matrix to determine inclusion of X instead of the structure matrix of loadings. Coarse factor score estimates often fail to preserve the properties of the factor model used to derive them (e.g., they often mischaracterize the correlations between the factors). Nevertheless, researchers find them attractive because of their simplicity and intuitive nature. Simulation studies suggest that coarse factor score estimates based on the structure matrix often have low correlations with true factor scores (Grice & Harris, 1998); using the factor score coefficient matrix to make decisions about which X "load" on a factor in place of the structure matrix tends to yield better estimates (Grice, 2001).

Dealing with Indeterminacy of Factor Scores

One obvious strategy for dealing with factor score indeterminacy is not to use factor scores. If you are interested in assessing the relationship between a factor and some other variable (either a cause or a determinant of the factor), this can be accomplished in structural equation modeling (SEM) without recourse to factor scores. Interested readers are referred to Bollen (1989), Brown (2015) and Kline (2015)

Having said that, practitioners may find themselves in situations where they need a brief assessment tool that is grounded in psychometric theory and that is based in empirics. In such cases, factor scores typically have been computed using the coarse method described above. There is nothing inherently wrong with this practice as long as one keeps in mind that the tool is now removed from the core logic of the factor analysis that motivated the tool in the first place. Rather, it is an assessment device in its own right whose properties need to be empirically established.

Mulaik (2005) is a strong advocate of embedding exploratory factor analysis into a larger theoretical network to better inform what the underlying factors mean and whether the factor scores derived from the factor model are theoretically meaningful. He uses the following analogy: Suppose you measure a variable, say an anxiety symptom (which we call X1), and someone tells you that there is another variable, F, that is correlated 0.80 with it. Can you state from this information what F is? The answer is, of course, no. If we also learn that F is highly correlated with X2 and X3 but not X4, X5 and X6 within the context of a factor analysis, then this provides us with more "clues" to discern what F might be. X1, X2 and X3 all might focus on social phobia/anxiety and X4, X5, and X6 might each focus on panic anxiety. Taking this a step further, if we also can identify predictors of F and outcomes of F in a broader nomological network, we have even more information to base an inference on. For example, we might find that F is related to a childhood history of parents being critical of their children in social situations and that F predicts future avoidance of social situations. Mulaik (2005) argues that the conduct of exploratory factor analysis in such broader nomological networks and the evaluation of the theoretical coherence of factor scores in those networks is key to addressing factor score indeterminacy and factor identification. Recent advancements in exploratory structural equation modeling have provided researchers with the tools to do this (Asparouhov & Muthén, 2009; Marsh et al., 2009).

The problem of factor indeterminacy has generated heated debate among psychometricians, a point driven home by papers titled like "The Green-MacDonald Proof of the Non-Existence of Factor Analysis" by Louis Guttman (1944) in response to work by Green and MacDonald that supposedly affirmed the utility of factor analysis. All three of these individuals (Green, MacDonald, and Guttman) were esteemed statisticians

with impeccable credentials. The underlying issues for indeterminacy are complex and a primer such as this is no place to delve into them in depth. We suggest Mulaik (2005) as a useful resource on the controversy more broadly

MAJOR AND MINOR FACTORS

Sometimes a factor analysis will result in a poor fit because in addition to a few major factors there also exist some minor factors that are more localized. Consider Figure 6.6 that presents a one factor model where a generalized communication quality factor between parent and child is thought to influence adolescent perceptions of parent-adolescent communication quality. The loadings are large and the correlations between the variables are obviously strong. Suppose that X1 and X2 are both in topical domains that pertain to school. The predicted correlation between them is the product of the two factor loadings, or (0.85)(0.85) = 0.72. Suppose, the observed correlation between them was 0.85 and, further, that all of the other observed correlations in the model were well reproduced by the loadings - it is just this one particular pair of variables that the correlation was under-predicted by too much. This suggests that there might be a "minor factor" that needs to be taken into account - a factor that serves as an additional, weaker common cause of just X1 and X2. This minor factor can be represented in different ways, but one approach represents it in the form of correlated residuals or correlated uniqueness components, per Figure 6.7. For example, suppose that both of the items in question dealt with conflict in school and that there exists a latent factor pertaining to school conflict that influences just these two items. This localized latent school conflict factor resides in *u1* and it also resides in *u2* because these uniqueness components reflect all determinants of X1 and X2 that are not captured by the underlying latent factor of generalized communication quality. Because both *u1* and *u2* contain this variable, they should be correlated with one another. In essence, the addition of correlated uniqueness components recognizes the presence of this minor factor. Technically, this is no longer a one factor model because we have posited a minor factor that influences X1 and X2. But the generalized communication quality factor can still be thought of as the primary source of the correlations between the X. Most software for exploratory factor analysis does not permit correlated residuals. One must use instead SEM software.

SAMPLE SIZE CONSIDERATIONS

A common question is what sample size is necessary to conduct a factor analysis. Most guidelines base sample size recommendations on the number of variables in the analysis, with more variables demanding larger sample sizes. The recommendations vary

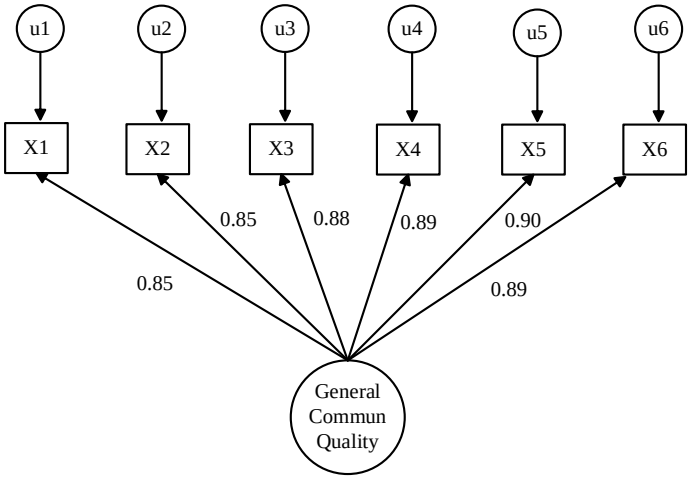

FIGURE 6.6. Single Factor Model of Communication

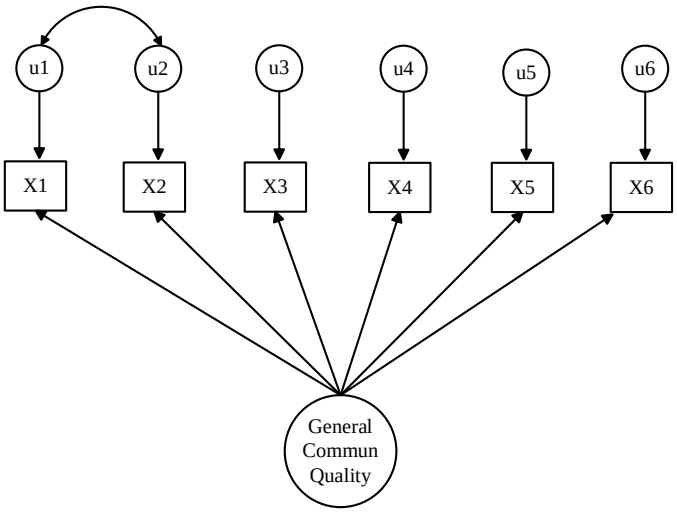

FIGURE 6.7. Single Factor Model of Communication with Correlated Error

dramatically, with some methodologists stating a ratio of 5 participants per measured variable is necessary; others recommend a ratio as high as 100 to 1. In simulation studies, most of these rules of thumb have been found to be unsatisfactory (Wolf, Harrington, Clark & Miller, 2013). The main limitation is that adequate sample size is not a simple function of the number of measured variables. Sample size needs are impacted by at least three factors, (1) the stability of the sample correlation matrix, (2) the use of asymptotic theory, and (3) statistical power/magnitude of margins of error. We discuss each, in turn.

Correlation Matrix Stability

The concept of sample correlation matrix stability is complex but it has to do with how well the sample correlation matrix represents the population correlation matrix and the extent to which deviations between the two based on sampling error can lead the analyst astray during model testing and evaluation. In simple terms, a reasonably stable sample correlation matrix is one that preserves the rank ordering of population correlations among all possible pairs of variables. As well, the magnitude of a given sample correlation should not be too discrepant from the corresponding population correlation. As a simple example, consider a 4X4 correlation matrix. Table 2 presents the population correlations and two examples of sample correlation matrices derived from that population but using different sample sizes.

Table 2: Population and Sample Correlations

	Population				Sample 1				Sample 2			
	X1	X2	X3	X4	X1	X2	X3	X4	X1	X2	X3	X4
X1	-				-				-			
X2	0.36	-			0.37	-			0.24	-		
X3	0.21	0.20	-		0.24	0.19	-		0.28	0.27	-	
X4	0.14	0.16	0.38	-	0.13	0.17	0.39	-	0.22	0.20	0.25	-

Note that population matrix shows tendencies towards a two factor model (X1 and X2 for one factor and X3 and X4 for another factor). Sample 1 correlations preserve the rank ordering of the population correlations and they are reasonably close in value to the population correlations. This is not the case for the Sample 2 correlations, which tend towards a one factor model. The Sample 1 correlation matrix is more stable than the Sample 2 correlation matrix.

Stability of a sample correlation matrix is impacted by (a) the sample size (larger sample sizes reduce sampling error), (b) the number of variables in the matrix (the greater the number of variables, the larger the correlation matrix and, in turn, the greater the opportunity for sampling error), and (c) the absolute magnitude of the correlations (correlations closer to zero tend to have more sampling error than correlations far away from zero. As well, (d) the sheer patterning of correlations in the population matrix can make a difference in the stability of the sample correlation matrix. For example, the more

variables that have non-trivial loadings on a factor in the population and the higher the communalities of those variables, the smaller the sample size usually can be to adequately capture the factor structure in the population (MacCallum et al., 1999). For all of these reasons, it should not be surprising that simple rules of thumb about sample size to number of variables ratios are of limited utility.

Asymptotic Theory

Asymptotic theory for exploratory factor analysis is an issue for the use of significance tests and confidence intervals in the analysis. It refers to scenarios where the sampling distribution of a parameter behaves in mathematically tractable ways but only as sample sizes become large. For example, sampling distributions for factor loadings in traditional maximum likelihood extraction tend towards mathematical tractability as sample size increases. The question becomes at what sample size is asymptotic theory compromised if one uses a smaller sample size. As with correlation matrix stability, you will encounter various rules of thumb about this matter that usually hover around required samples sizes of 100 to 150. These rules of thumb also tend to be oversimplified.

Statistical Power and Margins of Error

A third consideration for sample size determination is that of statistical power, again, if you are going to make use of significance tests or confidence intervals. If you use maximum likelihood based fit statistics to determine the number of factors, you will want to ensure that the sample size is sufficient to detect undesired levels of ill fit. MacCallum, Browne and colleagues (MacCallum, Browne & Sugawara, 1996; MacCallum, Browne & Cai, 2006; MacCallum, Lee & Browne, 2010) describe an approach to conducting power analysis using RMSEA fit statistics. As well, you may want to use a sample size that will produce margins of errors or confidence intervals for selected parameters that are not too large. Thus, rather than statistical power dictating sample size, a focus on sample size determination for margins of error is relevant (Kelley & Maxwell, 2003).

Determining Sample Size

In the final analysis, the only way to truly know the requisite sample size needed for a given exploratory factor analysis is to conduct a targeted simulation study specific to the model and research design you plan to employ. Such simulation strategies are discussed in Muthén and Muthén (2002). Having said that, sample sizes less than 100 or 150 typically (but not always) are a warning flag for potential problems.

DICHOTOMOUS AND ORDINAL VARIABLES

The factor models discussed thus far applies to measured variables that are continuous with interval or ratio level metrics. Researchers sometimes desire to factor analyze variables that either are measured coarsely, have ordinal properties, or are dichotomous. For such cases, special analytics may be required.

Traditional EFA analyzes Pearson's correlations. Although Pearson correlations are appropriate with continuous variables, it is often the case that we work with coarse measures of continuous constructs. Continuous variables, technically, have an infinite number of values between any two points. For example, time is a continuous variable and there is an infinite number of values that can occur between, say, 1 and 2 seconds (e.g., 1.10 seconds, 1.103 seconds, 1.1035 seconds, and so on). When we measure time or other continuous variables in research, our measures may be "coarse" and only make a limited number of discriminations between the lowest and highest scores. We might measure job satisfaction, a continuous construct, on a 7 point scale from very dissatisfied to very satisfied in a way that we have only 7 categories (1, 2, 3, 4, 5, 6, and 7). An important question is: how much trouble do we get into by using such coarse measures with statistical approaches that presume continuous measurement?

The answer depends on the question being asked and the particular statistical model being evaluated. In a classic study focused on Pearson correlations, Bollen and Barb (1981) conducted a simulation where they created data so that the true population correlations between two continuous variables were either 0.2, 0.6, 0.8, or 0.9. Bollen and Barb then created coarse measures from the continuous measures for each population by breaking the continuous measures into anywhere from 2 to 10 categories. For example, a continuous variable that ranges from -3 to +3 can be turned into a two point scale by assigning anyone with a score of 0 or less a "0" and anyone with a score greater than 0 a "1." Bollen and Barb computed the correlations using "coarse" measures and examined how close they were to the case where the correlation was computed when both variables were fully continuous. They found that the true correlations were relatively well reproduced by the coarse measures as long as the coarse measures had 5 or more categories. For example, the reproduced correlations for five category measures were within about 0.06 correlation units of the continuous-based correlations when the true correlations were at or below 0.60. Bollen and Barb (1981) concluded that five categories were probably sufficient for many applications. This recommendation has been replicated in many other simulation studies (although some research suggests seven or more categories may be best; see Green et al., 1997; Lozano, García-Cueto & Muñiz, 2008; Lubke & Muthén, 2004; Rhemtulla, Brosseau-Liard & Savalei, 2012; Taylor, West &

Aiken, 2006). Thus, coarse measurement is not necessarily problematic unless it is very coarse (less than five categories).

Coarseness of metrics is not the same as ordinality of metrics, although there is a link between the two properties. Given an underlying quantitative dimension that is continuous, a metric is an interval indicator of that dimension if it is a linear function of it. In the Bollen and Barb study, the break points of the scales were defined in a way that the quantitative categories were roughly linear functions of the true scores, although strict linearity was obviously compromised as the metrics became quite coarse. Nevertheless, the approximation to linearity was not too crude in many cases. Ordinal measures are ones that are monotonic, non-linear functions of the underlying dimension.

Some methodologists question the use of ordinal measures with Pearson correlations. To be sure, one can correlate any set of numbers with any other set of numbers and correlation analysis will provide insights about how the two sets of numbers are (linearly) related. However, in most applications, we are not interested in making statements about numbers per se; rather we want to make statements about the *constructs* that those numbers represent. We might want to estimate the correlation between the constructs of anxiety and depression (do people who worry alot also tend to be sad?), but all we have to work with are scores on anxiety and depression tests. If the measures are ordinal rather than interval, then characterizations of the magnitude of the correlation between the underlying *constructs* using the observed measures can be biased. For example, the true correlation might be 0.35, but if our measures are decidedly ordinal, we might find that the correlation between *measures* is only 0.20. In this sense, correlation analysis for continuous variables presumes interval measures.

Specialized ordinal methods for estimating correlations have been proposed to adjust for the bias that is introduced by using ordinal measures, but such approaches make their own assumptions that can lead to worse estimates of associations than just treating ordinal data as interval. The reality is that if measures are not “too ordinal,” then they likely can be meaningfully analyzed using Pearson correlations. The statement that a measure will not be problematic as long as it is not “too ordinal” will strike some readers as unusual because most students learn that scales are either ordinal or interval. Such dichotomous thinking is misleading. It is similar to saying someone has a fever because their body temperature is one tenth of one degree above the standard of 98.6ºF (37ºC). The person does indeed have a fever, but it is small and inconsequential. The same is true of measurement properties like ordinality. A measure can approach intervalness closely, but technically still not be interval.

Statisticians have developed measures of association to complement the Pearson correlation for the case where measures are too ordinal or too coarse. A *tetrachoric*

correlation is an index of association that reflects the correlation between two continuous variables, but the measurement of both of the variables occurs on coarse, dichotomous metrics. This type of correlation makes a simplifying assumption that the underlying continuous variables are bivariately normally distributed, but this assumption rarely holds in practice. A *polychoric correlation* generalizes the tetrachoric correlation to the case where one or both of the observed measures has more than two categories that are ordinal in character. Again, bivariate normality of the underlying continuous constructs is assumed. A *biserial correlation* estimates the correlation between two underlying continuous variables, but where only one of the two observed measures has been measured on a coarse, dichotomous level and the other has been measured on a continuous metric. A *polyserial correlation* is one where one of the measures is a coarse ordinal measure (of a continuous construct) with more than two categories, but the other measure is continuous. A tetrachoric correlation is a special case of a polychoric correlation and a biserial correlation is a special case of a polyserial correlation. Finally, a *point biserial correlation* is the correlation between a true dichotomous variable (e.g., has smoked marijuana versus not) and a continuous variable whose metric is continuous.

A problem with performing factor analysis on the above correlations is that many of the indices assume normality in ways that violations of the assumption are consequential. For example, a tetrachoric correlation between two dichotomous variables, X1 and X2, assumes that a given dichotomous variable (e.g., X1) is a crude index of a continuous variable that underlies it. The underlying variable is assumed to be normally distributed so that one can specify threshold values that define how individuals' true scores on the continuous variable translate into responses on the observed, dichotomous metric; values below the threshold translate to a score of 0 and values at or above the threshold translate into a score of 1. If the threshold values are misspecified, the correlation estimate is subject to bias. If the underlying variable is not normally distributed, then the threshold values will be misspecified. Tetrachoric correlations (as well as polychoric and polyserial correlations) also can be sample size demanding, show large sample-to-sample fluctuations, and can result in non-Gramian correlation matrices that produce analytic complications. Our experience is that one sometimes gets into more trouble using these methods with their attendant demands than just analyzing the data as if they were Pearson correlations in the first place.

When one works exclusively with dichotomously measured variables in factor analytic contexts, a common approach is to use as input a correlation matrix that contains tetrachoric correlations and then to apply one of the factor extraction methods described above (e.g., minres, maximum likelihood, weighted least squares). A common argument against using phi coefficients (which are simply Pearson's correlation computed on

dichotomous variables) is that factor analyses of them can yield artifactual factors based on "item difficulty," i.e., the base rates of the items. This assertion usually is made because the maximum value of a phi coefficient is not plus or minus unity but rather depends on the distribution of the dichotomous variables per se. A maximum absolute phi of 1.00 is only possible when the marginal distributions of the dichotomous variables are uniform, that is both are 50%. The argument is that if items differ substantially in their "difficulty" (base rates), then an artificial "difficulty" factor will emerge in a factor analysis. The use of tetrachoric correlations supposedly circumvents this artifact.[9]

The idea that difficulty confounds are inherent to phi coefficients has been challenged (McDonald, 1985; McDonald & Ahlawat, 1974). McDonald and Ahlawat (1974) showed that as long as the probabilities implied by dichotomous items are linearly related to the underlying factors, the analysis of phi coefficients will not result in spurious factors. Many contemporary measurement models for dichotomous variables are based in Item Response Theory (IRT) and assume the relationship between factors and dichotomous items has a logit function. This assumption usually is arbitrary. McDonald and Ahlawat also argue that the presence of a difficulty factor may be substantively meaningful rather than spurious. Factor analysis cannot distinguish this. Several simulation studies are consistent with McDonald's arguments. For example, Parry and McArdle (1991) compared methods for factor analyzing dichotomous variables where the input matrices were (1) phi correlations, (2) tetrachoric correlations estimated in two different ways, and (3) correlations derived from latent trait theory. They found that the latter methods were not markedly superior to the approach based on phi coefficients across a wide range of simulation conditions. Similar results have been reported by Collins, Cliff, McCormick and Zatkin (1986). Weng and Chen (2005) examined the performance of parallel analysis with phi versus tetrachoric correlations and found the latter to be unstable (see also Chou, Li & Bandalos, 2009); they recommended the phi coefficient (but see Garrido, Abad & Ponsoda, 2013, for qualifications). The decision to use specialized factor analytic methods for ordinal and dichotomous indicators is more complicated than many researchers appreciate.

SUMMARY AND CONCLUDING COMMENTS

Factor analysis is widely used in the social sciences. It is distinct from principal components analysis in that it attempts to explain the correlations between a set of

[9] As it turns out, the maximum value of Pearson's correlation in a population for continuous variables also is dependent on the variable marginal distributions. With bivariate normality, the maximum possible population correlations are -1.00 to 1.00. When the population X has a standard normal distribution and the population Y has a standard lognormal distribution, for example, the population correlation bounds are approximately ±0.76.

variables by identifying generalized common causes of them. By contrast, principal components analysis is a data reduction method aimed at forming orthogonal linear composites of a set of variables so as to reduce the number of variables one needs to work with in multivariate contexts.

Fundamental decision points in factor analysis include choosing a fit function, choosing the number of factors that are necessary to account for the correlation structure of the measured variables, and choosing a rotation method. Maximum likelihood fit functions have the advantage of being grounded in a strong theory of statistical inference, but the inferential aspects of the theory require assumptions of multivariate normality. Research suggests the method is relatively robust to violations of the assumption, but it does have its limits. Many methods have been suggested for choosing the number of factors necessary to account for a correlation structure, with pattern analysis being among the more effective of the ad hoc methods. Formal model comparison approaches have stronger grounding in statistical theory, with comparisons based on chi square difference tests, RMSEAs, and information fit indices (e.g., Haughton's BIC) being among the favored strategies. To be sure, each has strengths and weaknesses, so it is best to approach one's data from multiple vantage points in the spirit of sensitivity frameworks. Careful examination of the residual correlation matrix also is important. Among the many rotation methods, oblique rotations tend to be more realistic than orthogonal rotations and the geomin method has much to recommend as does the oblimin method

The indeterminacy of factor analysis and the factor scores associated with it is a limitation of the method. This fact has led some methodologists to view factor analysis as an arbitrary method that has limited practical value. At a minimum, the presence of indeterminacy should make you nervous and tentative in your conclusions. One way of dealing with indeterminacy is to reduce it by using targeted factor analysis or eliminate it altogether by using confirmatory factor analysis with just-identified or over-identified models.

Traditional methods of factor analysis focus on Pearson correlations as applied to continuous variables. Specialized methods have been developed for ordinal and dichotomous variables. It sometimes is possible to apply factor analysis to such variables without invoking the specialized methods if the metrics are not too ordinal and if the functions relating the factors to the measures are roughly linear in form.

REFERENCES

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling,* 16, 397-438.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, Statistical Section*, 3, 77-85.

Bentler, P. M. 1977. Factor simplicity index and transformations. *Psychometrika*, 42, 277–295.

Bentler, P. & Savalei, V. (2010). Analysis of correlation structures: Current status and open problems. In S. Kolenikov, L. Thombs and D. Steinley (Eds.) *Statistics in the social sciences: Current methodological developments*. New York: Wiley.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K., Harden, J. Ray, S. & Zavisca, J. (2014). BIC and alternative Bayesian Information Criteria in the selection of structural equation models. *Structural Equation Modeling*, 21, 1-19.

Brown, T. (2015). *Confirmatory factor analysis for applied research.* New York: Guilford.

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

Browne, M. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111-150.

Browne, M. & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. Clogg and M. Sobel (Eds). *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum.

Browne, M.W. & Cudek, R. (1993). Alternative ways of assessing model fit. pp. 136-162 in Bollen, K. and Long J.S (eds.) *Testing structural equation models*. Newbury Park: Sage.

Burnham, K. & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261-304.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Cattell, R. (1988). The meaning and strategic use of factor analysis. In J. Nesselroade and R. Cattell (Eds.). *Handbook of multivariate experimental psychology*. New York: Plenum

Cho, S., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, 69, 748 –759.

Collins, L., Cliff, N., McCormick, D. & Zatkin, J. (1986). Factor recovery in binary data sets: A simulation. *Multivariate Behavioral Research*, 21, 377-399.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, 17, 402-422.

Crawford, C. B. & Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, 35, 321-332.

Dinno, A. (2010). Gently clarifying the application of Horn's parallel analysis to principal component analysis versus factor analysis. From http://doyenne.com/Software/files/PA_for_PCA_vs_FA.pdf

Dunteman, G.H. (1989). *Principal components analysis*. Newbury Park: Sage.

Fabrigar, L., Wegener, D., MacCallum, R. & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Fabrigar, L. & Wegener, D. (2011). *Exploratory factor analysis: Understanding statistics*. New York: Oxford University Press.

Garrido, L., Abad, F. & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, 18, 454-474.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430-450.

Grice, J. W., & Harris, R. J. (1998). A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research*, 33, 221-247.

Guttman, L. (1944). The Green-McDonald proof of the nonexistence of factor analysis. In S. Levy (Ed.) *Louis Guttman on theory and methodology: Selected writings*. Brookfield, VT: Dartmouth Press.

Harman, H. H. (1976). *Modern factor analysis.* Chicago: The University of Chicago Press.

Heim, A. W. (1975). *Psychological testing*. London: Oxford University Press.

Jollife, I. (2002). *Principal component analysis*. New York: Springer.

Joreskog, K. (2007) Factor analysis and its extensions. In R. Cudeck and R. MacCallum (Ed.) *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum

Joreskog, K. & Goldberger, A. (1972). Factor analysis by generalized least squares. *Psychometrika*, 37, 243-260.

Kelley, K & Maxwell, S. (2003). Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, 8,3-5-321.

Kenny, D.A. (2015). SEM. Web page at http://davidakenny.net/cm/causalm.htm

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods and Research*, 44, 486-507.

Kiers, H. (1994). Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika*, 59, 567-579.

Kline, R. (2015). *Principles and practice of structural equation modeling.* New York: Guilford.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research*, 33, 188-208.

Liu, L. & Bentler, P. (2011). Quantified choice of RMSEAs for evaluation and power analysis of small differences between structural equation models. *Psychological Methods*, 16, 116–126.

Lovie, P. & Lovie, A. (1995). The cold equations: Spearman and Wilson on factor indeterminacy. *British Journal of Mathematical and Statistical Psychology*, 48, 237–253.

MacCallum, R. (2007) Factor analysis models as approximations. In R. Cudeck and R. MacCallum (Ed.) *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum

MacCallum, R. (2009). Factor analysis. In Milsap, R. and Maydeu-Olivares, A. (Eds.) *The SAGE handbook of quantitative methods in psychology*. Thousand Oaks, CA: Sage.

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.

MacCallum, R. C., Lee, T. & Browne, M. W. (2010). The issue of isopower in power analysis for tests of structural equation models. *Structural Equation Modeling*, 17, 23-41.

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130-149.

MacCallum, R. C., Widman, K. F., Zhang, S. & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-89.

McArdle, J. J. (1990). Principles versus principals of structural factor analyses. *Multivariate Behavioral Research*, 25, 81-87.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale NJ: Erlbaum.

McDonald, R. P. & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82-99.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U. & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment*, 22, 471-491.

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing non-nested structural equation models. *Psychological Methods*, 21,,151-163.

Muthén, B., du Toit, S. & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. https://www.statmodel.com/download/Article_075.pdf

Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age Publishing.

Mulaik, S. (2005) Looking back on the indeterminacy controversies in factor analysis. In A. Maydeu-Olivares and J. McArdle (Eds.) *Contemporary psychometrics*. Hillsdale, New Jersey: Erlbaum

Mulaik, S. (2009). *Foundations of factor analysis*. New York: Chapman and Hall.

Namboodiri, K. (1984). *Matrix algebra: An introduction*. Thousand oaks, CA: Sage.

Olsson, U., Foss, T., Troye, S. & Howell, R. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7, 557–595.

Olsson, U., Foss, T. & Brevil, 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 Research*, 32, 453-500.

Parry, C. & McArdle, J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement*, 15, 35-46.

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17, 1-14.

Preacher, K., Zhang, G., Kim, C. & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48, 28–56.

Pornprasertmanit, S., Wu, W. & Little, T. (2013). Taking into account sampling variability of model selection indices: A parametric bootstrap approach. *Multivariate Behavioral Research*, 48, 168 -169.

Raftery, A.E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, 25, 111-195.

Revelle, W. (2015). An introduction to psychometric theory with applications in R. Chapter 6: Constructs, components, and factor models. Downloaded on 12.10.2015 from http://www.personality-project.org/r/book/#chapter6.

Rhemtulla, M., Brosseau-Liard, P. & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354–373.

SAS Institute (2010). SAS/STAT 9.22 Users guide. Confidence intervals and salience of factor loadings. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_factor_sect016.htm

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 31, 699-714.

Schonemann, P. H. & Steiger, J. H. (1978). On the validity of indeterminate factor scores. *Bulletin of the Psychonomic Society*, 12, 287–290.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 2, 333-343.

Steiger, J. & Lind, J. (1980). Statistically based tests for the number of factors. Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, Iowa.

Steiger, J. H., & Schonemann, P. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the social sciences* (pp.136-178). San Francisco: Jossey-Bass.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321-327.

Wasserman, L. (1997). Bayesian model selection and model averaging (Working Paper No. 666). Carnegie Mellon University, Department of Statistics.

Weng, L. J. & Cheng, Ch. P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, 65, 697-716.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, 1, 354-365.

Yuan, K. & Bentler, P. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, 64, 737-757.

Zhang, G. (2014). Estimating standard errors in exploratory factor analysis. *Multivariate Behavioral Research*, 49, 339–353.

Zhang, G., Preacher, K. J., & Jennrich, R. I. (2012). The infinitesimal jackknife with exploratory factor analysis. *Psychometrika*, 77, 634-648.

Zwick, W.R. & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432-442.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, 92, 937-950.

APPENDIX: INFORMATION INDICES FOR MODEL CHOICE

When choosing between the different models to determine the number of classes, a commonly used set of comparative fit indices is based in a statistical theory known as *information theory*. Two such indices are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In general, researchers calculate an AIC index and/or a BIC index for the different models and then choose the model that has the best BIC or AIC value. In this appendix, we develop the logic of these indices, taking a few liberties in the interest of pedagogy. We first develop the concept of a log likelihood, a concept that is central to both the AIC and BIC. We then describe the model comparison process for the AIC, followed by consideration of that process for the BIC.

Log Likelihoods

Suppose we have a very large population and half the population is male and half the population is female. The probability of a randomly selected case being a male is 0.50 and this also is true for being a female. Stated more formally:

p(male) = 0.50 p(female) = 0.50

If we randomly select two cases, the probability of a given joint result across the two selections or "trials" is the product of their probabilities. As such, the probability of observing two males is

p(male)*p(male) = (0.50)(0.50) = 0.25

This is known as the multiplication rule for independent trials. Stated more formally, let p(A) = the probability of event A on a trial and p(B) = the probability of event B on a second (independent) trial. The joint probability of both events A and B is the product of the individual probabilities p(A) p(B). To be more concrete, there are four combinations that can result, each with a probability of 0.25:

Probability of a male on the first trial followed by a male on the second trial: 0.25
Probability of a male on the first trial followed by a female on the second trial: 0.25
Probability of a female on the first trial followed by a male on the second trial: 0.25
Probability of a female on the first trial followed by a female on the second trial: 0.25

and if we do not care about the order of appearance in the trials,

Probability of two males: 0.25
Probability of a male and a female: 0.50
Probability of two females: 0.25

We now shift gears review another facet of statistical theory that we will make use of. If we know that a very large set of scores is normally distributed with a certain mean and standard deviation, then we can use knowledge of the probability density function for a normal distribution to compute the probability of obtaining any given value when we randomly select a case from that distribution. The density formula is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{\frac{-.5(x-\mu)^2}{\sigma^2}}$$

where x is the score value in question, μ is the mean of the distribution, σ is the standard deviation of the distribution, π is the mathematical constant pi, *e* is the constant associated with the Naperian logarithm, and the density describes the height of the normal curve at the value of x. We can use this density in conjunction with calculus to calculate the probability of observing the score in question. As an example, if scores are normally distributed with a mean of 100 and a standard deviation of 13.77, then, using

the above formula, we find that the likelihood of a score of 99 is 0.0289. For a score of 87, it is 0.0186.[10]

Suppose we randomly select two scores from an extremely large population where scores are normally distributed with a mean of 100 and a standard deviation of 13.77. The probability that the scores will be 87 and 99, using the joint probability theorem described above, is (0.0289)(0.0186) = 0.00053754. Stated another way, the probability of observing these two data points given that the mean is 100 and the standard deviation is 13.77 (and assuming a normal distribution) is 0.00053754, with further adjustments to account for disinterest in the order of selection.

Suppose we randomly sample 100 data points from the population and calculate the likelihood of those 100 data points occurring using a strategy similar to the above method. The strategy would involve multiplying each probability by one another, with the result being a very, very small number. To make things more manageable and so as not to work with such small numbers, statisticians transform the final result by calculating the log of it, yielding what is called a *log likelihood*. The log likelihood is indicative of (but not equal to) the probability of obtaining the sample data given a "model" that states (a) the scores are normally distributed, (b) the mean is 100, and (c) the standard deviation is 13.77.

Log likelihoods are negative because the log of numbers less than 1.00 is always negative. For example, the natural log of 1.00 is zero, the natural log of 0.50 is -0.69, the natural log of 0.25 is -1.39, and the natural log of .01 is -4.61.[11]

Now, let's turn the above situation on its head. Suppose we have a set of 100 data points but we do not know the mean and standard deviation of the (assumed normal) distribution from which they come. We might, based on theory or logic, decide to "test" a model that states the mean is 95 and the standard deviation is 15. Using the probability density function from above and the strategies described, we can calculate the log likelihood for this model. The closer the log likelihood value is to zero (i.e., the less negative it is), the more likely the data came from the postulated model. We might formulate a second (competing) model that the mean is 100 and the standard deviation is 13.75 and calculate the log likelihood for it. Again, the closer the value of the log likelihood for this model is to zero, the more likely it is the data came from the model positing a mean of 100 and a standard deviation of 13.75.

[10] Technically, the probability of observing an exact value for a continuous variable is zero. We compute the likelihoods here by focusing on the interval defined by the real limits of the number (e.g., 98.5 to 99.5) in conjunction with the integral that scales the area under the curve to 1.00.

[11] Actually, some operationalizations of log likelihoods can yield positive numbers, but discussion of this point is beyond the scope of this primer.

We can compare the log likelihood values for the two models and we might find that one model results in a log likelihood closer to 0 than the other model. The model with the log likelihood closer to zero is more likely to have produced the data, hence we would prefer it to the model with the more negative log likelihood. Such is the fundamental logic of choosing between models based on their relative log likelihoods: We calculate the log likelihood of competing models and then choose the model with the log likelihood that is closest to zero. To be sure, the above explanation is simplistic and glosses over technicalities, but hopefully it conveys the general idea of comparing log likelihoods for two models.

As an aside, the above logic also is central to the well-known method of estimation called *maximum likelihood estimation*. In this approach, to estimate the mean of a distribution, one conceptually posits different models each representing a possible population mean value, calculates the likelihood of observing the data given the "model," and then selects the value/model that has the maximum likelihood.

Model Comparisons using the AIC

The AIC is an index of model likelihood or "model fit" based on a log likelihood. A common representation of it is

$$AIC = (-2)\,(LL) + 2k \qquad [1]$$

where LL is the log likelihood associated with the model in question and k is the number of estimable parameters in the model (such as when we estimate an intercept and the various regression coefficients). By multiplying the log likelihood by -2, the AIC essentially becomes a positive number, with larger numbers indicating lower likelihoods of the model. The AIC also includes what is often referred to as a penalty function for lack of parsimony, namely 2k. If the model has many parameters in it that must be estimated, then the AIC will be larger, everything else being equal. With the AIC, model parsimony is rewarded.[12] In general, the smaller the value of AIC, the better the "fit" of the model to the data. To make this intuitive, if the probability of the data given the model is 0.25, the log likelihood will be -1.39 and multiplying this by -2 yields 2.78. If the probability of the data given the model is much higher, say 0.50, the log likelihood is -0.69 and multiplying this by -2 yields 1.38. So, the smaller the value, the better the model. To this term, a penalty function is added that inflates the value of AIC for models that estimate more parameters

[12] Technically, the 2k term is part of the mathematical theory underlying the derivation of AIC. Also, choosing the value of -2 to multiply the LL by is not arbitrary. This value has a clear rationale. See Burnham and Anderson (2004).

There are many variations of the AIC. For example, some researchers use the above formula but with a small sample bias correction incorporated into it. This is sometimes referred to as AIC_c. The nuances of the different versions of the AIC are described in Burnham and Anderson (2004). Do not be surprised if for some software you observe AIC indices that are quite different in magnitude from other software. The important idea for all them is that we can compare different models using their respective AICs and then choose models that have "better" AICs when compared to other models.

Sometimes we compare more than two models, i.e., we might compare three, four or five models. When comparing more than two models, it is common to first identify the model with the lowest AIC value (which is the best fitting model of all the models being considered). One then calculates the difference in AIC values between each of the models and this best fitting model (subtracting the latter from the former). For the best fitting model, the difference will be zero and for all other models, it will be positive in value, with the larger the disparity, the worse the fit of the target model relative to the best fitting model.

General rules of thumb have been proposed to contextualize the magnitude of the difference in AICs between models (see Burnham & Anderson, 2004). The most common rules of thumb are as follows:

1. If the disparity in AICs is < 2, then the two models have about the same support

2. If the disparity in AICs is > 2 and < 4, then the better fitting model has positive support relative to the model it is compared with

3. If the disparity in AICs is > 4 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the disparity in AICs is > 10, then the better fitting model has very strong support relative to the model it is compared with.

Of course, one must be careful when applying rules of thumb like this because they may not apply in all contexts. Indeed, some analysts object to their specification, arguing that they can result in the same rigid and counterproductive use of a criterion like "$p < 0.05$" that plagues null hypothesis testing frameworks.

Another standard for comparing two models vis-a-vis the AIC is to examine what is called the *evidence ratio*. Let D = the AIC for the worse fitting model of the two models

minus the AIC for the better fitting model of the two models (and let *e* be the traditional Naperian constant). The evidence ratio is defined as

$$ER = 1 / e^{(-D / 2)}$$

where ER stands for "evidence ratio." It indicates how much more likely the better fitting model is (given the data) than the worse fitting model (given the data). For example, if the AIC for the better fitting model is 100 and for the worse fitting model it is 102, then the evidence ratio is

$$1 / e^{-(102-100) / 2)} = 2.63$$

The better fitting model is 2.63 times more likely to have yielded the data than the model it is being compared with.

Finally, some researchers normalize AIC differences relative to all models being compared so that they sum to 1. These are called *Akaike weights* and indicate the "weight of evidence" in favor of a model relative to *all* models in the comparison set. Akaike weights are distinct from evidence ratios because Akaike weights are impacted by the particular set of models being compared when the number of models is greater than two. Let us first describe how Akaike weights are calculated and then we will make them more concrete with an example.

To calculate the Akaike weight, each model is assigned an index of its likelihood relative to that of the best fitting model using the value from the denominator of the evidence ratio, $e^{(-D/2)}$, as the index. Let T = the sum of the $e^{(-D/2)}$ values across all the models being considered. Then the Akaike weight for a given model is defined as

$$e^{(-D/2)} / T$$

The weight ranges from 0 to 1.00, with higher values favoring the model in question.

To make this concrete, suppose we fit five different models to a set of data. Here is a table with the AICs, the differences between the model AIC versus the model with the lowest AIC, and the Akaike weights (w):

Model	AIC	D	$e^{(-D/2)}$	$w = e^{(-D/2)}/T$
1	204	2	0.3678	0.2242
2	202	0	1.0000	0.6094
3	206	4	0.1353	0.0824
4	206	4	0.1353	0.0824

5	214	12	0.0024	0.0015
Sum			T = 1.6408	1.0000

The sum of the weights across all five models is 1.00. The weights represent a continuous measure of relative strength of evidence for each model. Each weight can be crudely interpreted as the probability that the model is the best model among the set. In the present case, the data support Model 2.

The basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in AICs, the evidence ratios, the Akaike weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

Model Comparisons using the BIC

We describe the logic of the BIC using the Schwartz BIC, which is formally defined as

$$BIC = -2\ LL + \ln(N)\ k \qquad [2]$$

where k = the number of estimable parameters in the model, N = the sample size, and LL = the model log likelihood. Like the AIC, the smaller the BIC, the better the model fit, everything else being equal. Like the AIC, there is a penalty function for lack of parsimony, but the penalty is different than the AIC. The penalty is somewhat harsher for the BIC as opposed to the AIC. There are other instantiations of the BIC, and we discuss these below. For current purposes, we use the Schwartz formulation.

Like the AIC, it is not uncommon for the model with the smallest BIC to be used as a reference point for comparing models, with a common practice being to calculate the difference between each model in the model set and the model with the best BIC, like we did for the AIC. For the best fitting model, this difference will be zero.

To evaluate models in terms of BIC differences, general rules of thumb are (see Raftery, 1995):

1. If the BIC disparity < 2.2, then the better fitting model and the model it is compared with have about the same support

2. If the BIC disparity > 2.2 and < 6, then the better fitting model has positive support relative to the model it is compared with

3. If the BIC disparity > 6 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the BIC disparity > 10 then the better fitting model has very strong support relative to the model it is compared with

For similar but slightly different standards, see Wasserman (1997).

One also can calculate what is called a *Bayes Factor* (BF) for each model relative to the best fitting model. It is defined as

$BF = e^{(D'/2)}$

where D' is the BIC difference between the target model and the best fitting model. The Bayes factor is the probability that the model with the lower BIC produced the data divided by the probability the model in question produced the data. For example, a BF = 10 means it is 10 times more likely the model with the minimum BIC produced the data than the model in question.

Finally, a relative model weight, analogous to the Akaike weight, can be computed by normalizing model likelihoods relative to *all* models in the comparison set so that they sum to 1. Let D = the difference in the BIC for the model in question minus the value of the BIC for the best fitting model, T = the sum of the index $e^{(-D/2)}$ across each model. The relative weight for a model is

$e^{(-D/2)} / T$

The weight ranges from 0 to 1.00, with higher values favoring the model. Again, the sum of the weights across models is 1.00.

As with the AIC, the basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in BICs, the Bayes factors, the relative weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

You will encounter variants of the BIC, but the basic logic in applying them is the same. For example, like the AIC_c, there is a sample size adjusted BIC that is similar to Schwartz' BIC, but it applies a somewhat milder penalty function (Sclove, 1987). There also are variants of both the AIC and BIC to deal with dispersion issues in count regression models (called QAIC and QBIC).

Which Method is Better, AIC or BIC?

A debated topic in statistics is which approach to model comparison is better, one based on AICs or one based on BICs. There are advocates on both sides of the matter and we dare not venture into this controversy here. The BIC tends to favor simpler models more so than the AIC. This can be both a strength and a weakness. Interested readers are referred to Burnham and Anderson (2004), Yang (2005), and Kuha (2004). Kuha argues for the use of both indices.

An issue with both approaches is that researchers can be lulled into thinking that the best fitting model within a set of models is the true model. This is not necessarily the case. Researchers can choose the best of a set of wrong models, which is not our goal.

Haughton's BIC

In factor analysis, a particularly useful index of fit based on the BIC is known as *Haughton's BIC*, which we abbreviate as HBIC. Let χ^2_{ML} be the chi square value associated with the maximum likelihood test of the residual matrix discussed above. The formula for the HBIC index is

$$HBIC = \chi^2_{ML} - (df)\,(\ln(N/(2\pi))$$

where df is the degrees of freedom associated with the chi square statistic. Recall that the χ^2_{ML} ranges from 0 (perfect fit) to large positive numbers, with lower values indicating better model fit, everything else being equal. As such, HBIC can be thought of as an index of model fit with lower values indicating better fit, but with an "adjustment factor" for the chi square, namely, the term (df) (ln(N/(2π)). The statistical theory used to define this adjustment factor is beyond the scope of this primer (see Bollen et al., 2014), but ultimately, HBIC reflects model fit. The values of HBIC range from large negative values to large positive values and, like other forms of the BIC, are difficult to interpret for a single model in isolation. Instead, HBIC is useful for comparing two (or more) models in terms of their relative fit, such as a two factor model versus a one factor model. The better fitting model will have the lower value of HBIC. Thus, another strategy for comparing models with differing numbers of factors is to choose the model with the lowest HBIC.

Suppose two models are compared using HBIC based on a random sample from a population. One likely would obtain different values of the HBIC disparity between the two models if a different random sample of the same size was selected from the population. Some methodologists have argued that such sample-to-sample fluctuations in HBICs should be taken into account when choosing models (Preacher & Merkel, 2012).

The fluctuations typically will be greater for smaller as opposed to larger sample sizes. Preacher and Merkel (2012) developed a method for estimating margins of error for HBIC, allowing uncertainty to be taken into account when making modeling choices. For a description of formal methods for comparing BIC differences using a confidence interval approach, see Merkle, You and Preacher (2016) and Pornprasertmanit, Wu and Little (2013).