Chapter 1

INTRODUCTION TO FACTOR ANALYSIS

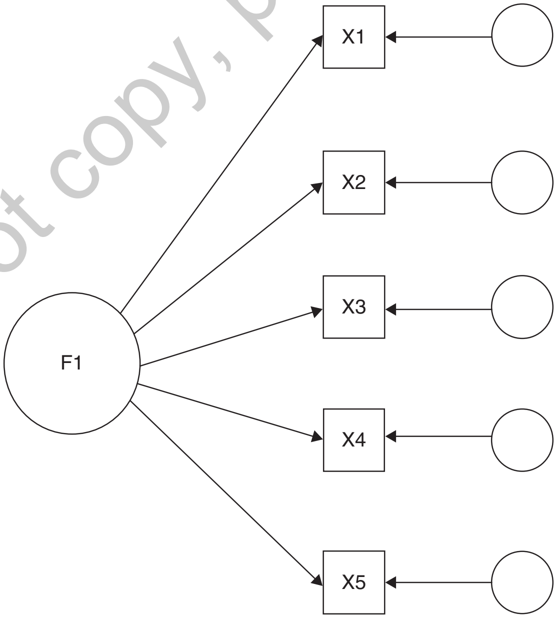
Factor analysis is perhaps one of the most widely used statistical procedures in the social sciences. An examination of the PsycINFO database for the period between January 1, 2000, and September 19, 2018, revealed a total of approximately 55,000 published journal articles indexed with the keyword *factor analysis*. Similar results can be found by examining the ERIC database for education research and JSTOR for other social sciences. Thus, it is not an exaggeration to state that understanding factor analysis is key to understanding much published research in the fields of psychology, education, sociology, political science, anthropology, and the health sciences. The purpose of this book is to provide you with a solid foundation in exploratory factor analysis, which, along with confirmatory factor analysis, represents one of the two major strands within this broad field. Indeed, a portion of this first chapter will be devoted to comparing and contrasting these two ways of conceptualizing factor analysis. However, before getting to that point, we first need to describe what, exactly, factors are and the differences between latent and observed variables. We will then turn our attention to the importance of having strong theory to underpin the successful use of factor analysis, and how this theory should serve as the basis upon which we understand the latent variables that this method is designed to describe. We will then conclude the chapter with a brief discussion of the software available for conducting factor analysis and an outline of the book itself. My hope in writing this book is to provide you, the reader, with a sufficient level of background in the area of exploratory factor analysis so that you can conduct analyses of your own, delve more deeply into topics that might interest you, and confidently read research that has used factor analysis. If this book achieves these goals, then I will count it as a success.

Latent and Observed Variables

Much research in fields such as psychology is focused on variables that cannot be directly measured. These variables are often referred to as being latent, and include such constructs as intelligence, personality, mood, affect, and aptitude. These latent variables are frequently featured in social science research and are also the focus for clinicians who want to gain insights into the psychological functioning of their clients. For example, a researcher might be interested in determining whether there is a relationship between extraversion and job satisfaction, whereas a clinician may want to know whether her client is suffering from depression. In both cases, the variables of interest (extraversion, job satisfaction, and depression) are conceived of as tangible, real constructs, though they cannot be directly measured or observed. We talk about an individual as being an extravert or we conclude that a person is suffering from depression, yet we have no direct way of observing either of those traits. However, as we will see in this book, these latent variables can be represented in the statistical model that underlies factor analysis.

If latent variables are, by their very nature, not observable, then how can we hope to measure them? We make inferences about these latent variables by using variables that we *can* measure, and which we believe are directly impacted by the latent variables themselves. These observed variables can take the form of items on a questionnaire, a test, or some other score that we can obtain directly, such as behavior ratings made by a researcher of a child's behavior on the playground. We generally conceptualize the relationship between the latent and observed variables as being causal, such that one's level on the latent variable will have a direct impact on scores that we obtain on the observed variable. This relationship can take the form of a path diagram, as in Figure 1.1.

Figure 1.1 Example Latent Model Structure



We can see that each observed variable, represented by the squares, is linked to the latent variable, denoted as F1, with unidirectional arrows. These arrows come from the latent to the observed variables, indicating that the former has a causal impact on the latter. Note also that each observed variable has an additional unique source of variation, known as error and represented by the circles at the far right of the diagram. Error represents everything that might influence scores on the observed variable, other than the latent variable that is our focus. Thus, if the latent variable is mathematics aptitude, and the observed variables are responses to five items on a math test, then the errors are all of the other things that might influence those math test responses, such as an insect buzzing past, distracting noises occurring during the test administration, and so on. Finally, latent variables (i.e., the factor and error terms) in this model are represented by circles, whereas observed variables are represented by squares. This is a standard way in which such models are diagrammed, and we will use it throughout the book.

In summary, we conceptualize many constructs of interest in the social sciences to be latent, or unobserved. These latent variables, such as intelligence or aptitude, are very important, both to the goal of understanding individual human beings as well as to understanding the broader world around us. However, these constructs are frequently not directly measurable, meaning that we must use some proxy, or set of proxies, in order to gain insights about them. These proxy measures, such as items on psychological scales, are linked to the latent variable in the form of a causal model, whereby the latent variable directly causes manifest outcomes on the observed variables. All other forces that might influence scores on these observed variables are lumped together in a latent variable that we call error, and which is unique to each individual indicator variable. Next, we will describe the importance of theory in both constructing and attempting to measure these latent variables.

The Importance of Theory in Doing Factor Analysis

As we discussed in the previous section, latent variables are not directly observable, and we only learn about them indirectly through their impact on observed indicator variables. This is a very important concept for us to keep in mind as we move forward in this book, and with factor analysis more generally. How can we know that performance or scores on the observed variables are in fact caused by the latent variable of interest? The short answer is that we cannot know for sure. Indeed, we cannot know that the latent variable does in fact exist. Is depression a concrete, real disease? Is extraversion an actual personality trait? Is there such a thing as reading aptitude? The answer to these questions is we don't know for sure. How then can we make statements about an individual suffering from depression, or that Juan is a good reader, or that Yi is an extravert? We can make such statements because we have developed a theoretical model that explains how our observed scores should be linked to these latent variables. For example, psychologists have taken prior empirical research as well as existing theories about mood to construct a theoretical explanation for a set of behaviors that connote the presence (or absence) of depression. These symptoms might include sleep disturbance (trouble sleeping or sleeping too much), a lack of interest in formerly pleasurable activities, and contemplation of suicide. Alone, these are simply behaviors that could be derived from a variety of sources unique to each. Perhaps an individual has trouble sleeping because he is excited about a coming job change. However, if there is a theoretical basis for linking all of these behaviors together through some common cause (depression), then we can use observed responses on a questionnaire asking about them to make inferences about the latent variable. Similarly, political scientists have developed conceptual models of political outlook to characterize how people view the world. Some people have views that are characterized as being conservative, others have liberal views, and still others fall somewhere in between the two. This notion of political viewpoint is based on a theoretical model and is believed to drive attitudes that individuals express regarding particular societal and economic issues, which in turn are manifested in responses to items on surveys. However, as with depression, it is not possible to say with absolute certainty that political viewpoint is a true entity. Rather, we can only develop a model and then assess the extent to which observations taken from nature (i.e., responses to survey questions) match with what our theory predicts.

Given this need to provide a rationale for any relationships that we see among observed variables, and that we believe is the result of some unobserved variable, having strong theory is crucial. In short, if we are to make claims about an unobserved variable (or variables) causing observed behaviors, then we need to have some conceptual basis for doing so. Otherwise, the claims about such latent relationships carry no weight. Given that factor analysis is the formalized statistical modeling of these latent variable structures, theory should play an essential role in its use. This means that prior to conducting factor analysis, we should have a theoretical basis for what we expect to find in terms of the number of latent variables (factors), and for how observed indicator variables will be associated with these factors. This does not mean that we cannot use factor analysis in an exploratory way. Indeed, the entire focus of this text is on exploratory factor analysis. However, it does mean that we should have some sense for what the latent variable structure is likely to be. This translates into having a general sense for the number of factors that we are likely to find (e.g., somewhere between two and four), and how the observed variables would be expected to group together (e.g., items 1, 3, 5, and 8 *should* be measuring a common construct and thus *should* group together on a common factor). Without such a preexisting theory about the likely factor structure, we will not be able to ascertain when we have an acceptable factor solution and when we do not. Remember, we are using observed data to determine whether predictions from our factor model are accurate. This means that we need to have a sufficiently well-developed factor model so as to make predictions about what the results *should* look like. For example, what does theory say about the relationship between depression and sleep disturbance? It says that individuals suffering from depression will experience what for them are unusual sleep patterns. Thus, we would expect depressed individuals to indicate that they are indeed suffering from unusual sleep patterns. In short, having a well-constructed theory about the latent structure that we are expecting to find is crucial if we are to conduct the factor analysis properly and make good sense of the results that it provides to us.

Comparison of Exploratory and Confirmatory Factor Analysis

Factor analysis models, as a whole, exist on a continuum. At one extreme is the purely exploratory model, which incorporates no a priori information, such as the possible number of factors or how indicators are associated with factors. At the other extreme lies a purely confirmatory factor model in which the number of factors, as well as the way in which the observed indicators group onto these factors, is provided by the researcher. These modeling frameworks differ both conceptually and statistically. From a conceptual standpoint, exploratory models are used when the researcher has little or no prior information regarding the expected latent structure underlying a set of observed indicators. For example, if very little prior empirical work has been done with a set of indicators, or there is not much in the way of a theoretical framework for a factor model, then by necessity the researcher would need to engage in an exploratory investigation of the underlying factor structure. In other words, without prior information on which to base the factor analysis, the researcher cannot make any presuppositions regarding what the structure might look like, even with regard to the number of factors underlying the observed indicators. In other situations, there may be a strong theoretical basis upon which a hypothesized latent structure rests, such as when a scale has been developed using well-established theories. However, if very little prior empirical work exists exploring this structure, the researcher may not be able to use a more confirmatory approach and thus would rely on exploratory factor analysis (EFA) to examine several possible factor solutions, which might be limited in terms of the number of latent variables by the theoretical framework upon which the model is based. Conceptually, a confirmatory factor analysis (CFA) approach would be used when there is both a strong theoretical expectation regarding the expected factor structure and prior empirical evidence (usually in the form of multiple EFA studies) supporting this structure. In such cases, CFA is used to (a) ascertain how well the hypothesized latent variable model fits the observed data and (b) compare a small number of models with one another in order to identify the one that yields the best fit to the data.

From a statistical perspective, EFA and CFA differ in terms of the constraints that are placed upon the factor structure prior to estimation of the model parameters. With EFA there are few, if any, constraints placed on the model parameters. Observed indicators are typically allowed to have nonzero relationships with all of the factors, and the number of factors is not constrained to be a particular number. Thus, the entire EFA enterprise is concerned with answering the question of how many factors underlie an observed set of indicators, and what structure the relationship between factors and indicators takes. In contrast, CFA models are highly constrained. In most instances, each indicator variable is allowed to be associated with only a single factor, with relationships to all other factors set to 0. Furthermore, the specific factor upon which an indicator is allowed to load is predetermined by the researcher. This is why having a strong theory and prior empirical evidence is crucial to the successful fitting of CFA models. Without such strong prior information, the researcher may have difficulty in properly defining the latent structure, potentially creating a situation in which an improper model is fit to the data. The primary difficulty with fitting an incorrect model is that it may appear to fit the data reasonably well, based on statistical indices, and yet may not be the correct model. Without earlier exploration of the likely latent structure, however, it would not be possible for the researcher to know this. CFA does have the advantage of being a fully determined model, which is not the case with EFA, as we have already discussed. Thus, it is possible to come to more definitive determinations regarding which of several CFA models provides the best fit to a set of data because they can be compared directly using familiar tools such as statistical hypothesis testing. Conversely, determining the optimal EFA model for a set of data is often not a straightforward or clear process, as we will see later in the book.

In summary, EFA and CFA sit at opposite ends of a modeling continuum, separated by the amount of prior information and theory available to the researcher. The more such information and the stronger the theory, the more appropriate CFA will be. Conversely, the less that such prior evidence is available, and the weaker the theories about the latent structure, the more appropriate will be EFA. Finally, researchers should take care not to use both EFA and CFA on the same set of data. In cases where a small set of CFA models do not fit a set of sample data well, a researcher might use EFA in order to investigate potential alternative models. This is certainly an acceptable approach; however, the same set of data used to investigate these EFA-based alternatives should not then be used with an additional CFA model to validate what exploration has suggested might be optimal models. In such cases, the researcher would need to obtain a new sample upon which the CFA would be fit in order to investigate the plausibility of the EFA findings. If the same data were used for both analyses, the CFA model would likely yield spuriously good fit to the sample for the model, given that the sample data had already yielded the factor structure that is being tested, through the EFA.

EFA and Other Multivariate Data Reduction Techniques

Factor analysis belongs to a larger family of statistical procedures known collectively as data reduction techniques. In general, all data reduction techniques are designed to take a larger set of observed variables and combine them in some way so as to yield a smaller set of variables. The differences among these methods lies in the criteria used to combine the initial set of variables. We discuss this criterion for EFA at some length in Chapter 3, namely the effort to find a factor structure that yields accurate estimates of the covariance matrix of the observed variables using a smaller set of latent variables. Another statistical analysis with the goal of reducing the number of observed variables to a smaller number of unobserved variates is discriminant analysis (DA). DA is used in situations where a researcher has two or more groups in the sample (e.g., treatment and control groups) and would like to gain insights into how the groups differ on a set of measured variables. However, rather than examining each variable separately, it is more statistically efficient to consider them collectively. In order to reduce the number of variables to consider in this case, DA can be used. As with EFA, DA uses a heuristic to combine the observed variables with one another into a smaller set of latent variables that are called discriminant functions. In this case, the algorithm finds the combination(s) that maximize the group mean difference on these functions. The number of possible discriminant functions is the minimum of p and J-1, where p is the number of observed variables, and J is the number of groups. The functions resulting from DA are orthogonal to one another, meaning that they reflect different aspects of the shared group variance associated with the observed variables. The discriminant functions in DA can be expressed as follows:

$$D_{fi} = w_{f1}\,x_{1i} + \mathrm{w}_{f2}x_{2i} + \cdots + \mathrm{w}_{fp}\,x_{pi} \qquad \text{(Equation 1.1)}$$

where

D_{fi} = Value of discriminant function f for individual i

w_{fp}= Discriminant weight relating function f and variable p

x_{pi} = Value of variable p for individual i.

For each of these discriminant functions (D_f), there is a set of weights that are akin to regression coefficients and correlations between the observed variables and the functions. Interpretation of the DA results usually involves an examination of these correlations. An observed variable having a large correlation with a discriminant function is said to be associated with that function in much the same way that indicator variables with large loadings are said to be associated with a particular factor. Quite frequently, DA is used as a follow-up procedure to a statistically significant multivariate analysis of variance (MANOVA). Variables associated with discriminant functions with statistically significantly different means among the groups can be concluded to contribute to the group mean difference associated with that function. In this way, the functions can be characterized just as factors are, by considering the variables that are most strongly associated with them.

Canonical correlation (CC) works in much the same fashion as DA, except that rather than having a set of continuous observed variables and a categorical grouping variable, CC is used when there are two sets of continuous variables for which we want to know the relationship. As an example, consider a researcher who has collected intelligence test data that yields five subtest scores. In addition, she has also measured executive functioning for each subject in the sample, using an instrument that yields four subtests. The research question to be addressed in this study is, how strongly related are the measures of intelligence and executive functioning? Certainly, individual correlation coefficients could be used to examine how pairs of these variables are related to one another. However, the research question in this case is really about the extent and nature of relationships between the two *sets* of variables. CC is designed to answer just this question, by combining each set into what are known as canonical variates. As with DA, these canonical variates are orthogonal to one another so that they extract all of the shared variance between the two sets. However, whereas DA created the discriminant function by finding the linear combinations of the observed indicators that maximized group mean differences for the functions, CC finds the linear combinations for each variable set that maximize the correlation between the resulting canonical variates. Just as with DA, each observed variable is assigned a weight that is used in creating the canonical variates. The canonical variate is expressed as in Equation 1.2.

$$C_{vi} = w_{c1}\,x_{1i} + w_{c2}\,x_{2i} + \cdots + w_{cp}\,x_{pi} \qquad \text{(Equation 1.2)}$$

where

C_{vi} = Value of canonical variate *v* for individual *i*

w_{cp} = Canonical weight relating variate *v* and variable *p*

x_{pi} = Value of variable *p* for individual *i*.

Note how similar Equation 1.1 is to Equation 1.2. In both cases, the observed variables are combined to create one or more linear combination scores. The difference in the two approaches is in the criteria used to obtain the weights. As noted above, for DA the criteria involve maximizing group separation on the means of D_f, whereas for CC the criteria is the maximization of correlation between C_v for the two sets of variables.

The final statistical model that we will contrast with EFA is partial least squares (PLS), which is similar to CC in that it seeks to find linear combinations of two sets of variables such that the relationship between the sets will be maximized. This goal stands in contrast to EFA, in which the criterion for determining factor loadings is the optimization of accuracy in reproducing the observed variable covariance/correlation matrix. PLS differs from CC in that the criterion it uses to obtain weights involves both the maximization of the relationship between the two sets of variables as well as maximizing the explanation of variance for the variables within each set. CC does not involve this latter goal. Note that PCA, which we discuss in Chapter 3, also involved the maximization of variance explained within a set of observed variables. Thus, PLS combines, in a sense, the criteria of both CC and PCA (maximizing relationships among variable sets and maximizing explained variance within variable sets) in order to obtain linear combinations of each set of variables.

A Brief Word About Software

There are a large number of computer software packages that can be used to conduct exploratory factor analysis. Many of these are general statistical software packages, such as SPSS, SAS, and R. Others are specifically designed for latent variable modeling, including Mplus and EQS. For many exploratory factor analysis problems, these various software packages are all equally useful. Therefore, you should select the one with which you are most comfortable, and to which you have access. On the other hand, when faced with a nonstandard factor analysis problem, such as having multilevel data, the use of specialized software designed for these cases might be necessary. In order to make this text as useful as possible, on the book website at **study.sagepub.com/researchmethods/qass/finch-exploratory-factor-analysis**, I have included example computer code and the annotated output for all of the examples included in the text, as well as additional examples designed to demonstrate the various analyses described here. I have attempted to avoid including computer code and output in the book itself so that we can keep our focus on the theoretical and applied aspects of exploratory factor analysis, without getting too bogged down in computer programming. However, this computer-related information does appear on the book website, and I hope that it will prove helpful to you.

Outline of the Book

The focus of this book is on the various aspects of conducting and interpreting exploratory factor analysis. It is designed to serve as an accessible introduction to this topic for readers who are wholly unfamiliar with factor analysis and as a reference to those who are familiar with it and who need a primer on some aspect of the method. In Chapter 2, we will lay out the mathematical foundations of factor analysis. This discussion will start with the correlation and covariance matrices for the observed variables, which serves as the basis upon which the parameters associated with the factor analysis model are estimated. We will then turn our attention to the common factor model, which expresses mathematically what we see in Figure 1.1. We will conclude Chapter 2 with a discussion of some important statistics that will be used throughout the book to characterize the quality of a particular factor solution, including eigenvalues, communalities, and error variances.

Chapter 3 presents the first major step in conducting a factor analysis, extraction of the factors themselves. Factor extraction involves the initial estimation of the latent variables that underlie a set of observed indicators.

We will see that there are a wide range of methods for extracting the initial factor structure, all with the goal of characterizing the latent variables in terms of the observed ones. The relationships between the observed and latent variables are expressed in the form of factor loadings, which can be interpreted as correlations between the observed and latent variables. The chapter describes various approaches for estimating these loadings, with a focus on how they differ from one another. Finally, we conclude Chapter 3 with an example. Chapter 4 picks up with the initially extracted factor loadings, with a discussion of the fact that the initially extracted loadings are rarely interpretable. In order to render them more useful in practice, we must transform them using a process known as rotation. We will see that there are two general types of rotation: one allowing factors to be correlated (oblique) and the other which restricts the correlations among the factors to be 0 (orthogonal). We will then describe how several of the more popular of these rotations work, after which we present a full example, and then conclude the chapter with a discussion of how to decide which rotation we should use.

One of the truths about exploratory factor analysis is that the model is indeterminate in nature. This means that there are an infinite number of mathematically plausible solutions, and no one of them can be taken as optimal over the others. Thus, we need to have some criteria for deciding what the optimal solution is likely to be. Making this determination is the focus of Chapter 5. First and foremost, we must be sure that the solution we ultimately decide upon is conceptually meaningful. In other words, the factor model must make sense and have a basis in theory in order for us to accept it. Practically speaking, this means that the way in which the variables group together in the factors is reasonable. In addition to this theoretically based determination, there are also a number of statistical tools available to us when deciding on the number of factors to retain. Several of these are ad hoc in nature and may not provide terribly useful information. Others, however, are based in statistical theory and can provide useful inference regarding the nature of the final factor analysis model. We will devote time to a wide array of approaches, some more proven than others, but all useful to a degree. We close the chapter with a full example and some discussion regarding how the researcher should employ these various methods together in order to make the most informed decision possible regarding the number of factors to retain.

We conclude the book with a chapter designed to deal with a variety of ancillary issues associated with factor analysis. These include the calculation and use of factor scores, which is somewhat controversial. Factor scores are simply individual estimates of the latent trait being measured by the observed indicator variables. They can be calculated for each member of the sample and then used in subsequent analyses, such as linear regression. Given the indeterminacy of the exploratory factor model, however, there is disagreement regarding the utility of factor scores. We will examine different methods for calculating them and delve a bit into the issue of whether or not they are useful in practice. We will then consider important issues such as a priori power analysis and sample size determination, as well as the problem of missing data. These are both common issues throughout statistics and are important in exploratory factor analysis as well. We will then focus our attention on two extensions of EFA, one for cases in which we would like to investigate relationships among latent variables, but where we do not have a clear sense for what the factors should be. This exploratory structural equation modeling merges the flexibility of EFA with the ability to estimate relationships among latent variables. We will then turn our attention to the case when we have multilevel data, such that individuals are nested within some collective, such as schools or nations. We will see how ignoring this structure can result in estimation problems for the factor model parameters, but that there is a multilevel factor model available to deal with such situations. We will conclude the chapter and the book with discussions on best practices for reporting factor analysis results and where exploratory factor analysis sits within the broader framework of statistical data reduction. This discussion will include tools such as discriminant analysis, canonical correlation, and partial least squares regression.

Upon completing this book, I hope that you are comfortable with the basics of exploratory factor analysis, and that you are aware of some of the exciting extensions available for use with it. Factor analysis is a powerful tool that can help us understand the latent structure underlying a set of observed data. It includes a set of statistical procedures that can be quite subtle to use and interpret. Indeed, it is not hyperbole to say that successfully using factor analysis involves as much art as it does science. Thus, it is important that when we do make use of this tool, we do so with a good sense for what it can and cannot do, and with one eye fixed firmly on the theoretical underpinnings that should serve as our foundation. With these caveats in mind, let's dive in.