CHAPTER 1

An Introduction to Multivariate Design

1.1 The Use of Multivariate Designs

The use of multivariate research designs has grown very rapidly in the behavioral and social sciences throughout the past quarter century. This has been made possible in no small part by increased availability of sophisticated statistical software packages, such as IBM SPSS, SAS, and Stata, that can be installed on personal computers. But even with such increased availability of such software, behavioral and social science researchers have been using some multivariate techniques (e.g., factor analysis, multiple regression) for a very long time.

Multivariate designs can be distinguished from the univariate and bivariate designs with which readers are likely already familiar. Experimental designs that are analyzed with t tests or analysis of variance (ANOVA) are univariate designs, so named because there is only a single dependent variable in the design and analysis of the data (Gamst, Meyers, & Guarino, 2008). A t ratio or an F ratio is generated to test whether the group means are significantly different.

A bivariate design derives its name from the fact that there are only two variables that are analyzed together; it is exemplified by a simple correlation design. The variables in such a design are often signified as X and Y and, unless we are predicting one (the Y variable) from the other (the X variable), which variable is assigned which letter is arbitrary. The degree to which the measures are correlated is assessed with a correlation statistic, the most commonly cited one being the Pearson correlation coefficient (Pearson r).

1.2 The Definition of the Multivariate Domain

To be considered a multivariate research design, the study must have more variables than are contained in either a univariate or bivariate design. Furthermore, some subset of these variables must be analyzed together, that is, they must be combined in some manner to form a composite variable or *variate*. The most common way to combine variables is by forming a *weighted linear composite* where each variable is weighted in a manner determined by the analysis. This resulting weighted linear composite is known as a variate. There are several contexts where we form such variates, three examples of which are as follows:

- In an experimental design in which we wished to compare the performance of three types of memory training, we could measure two or more variables as indicators of performance. These variables could then be combined into a single weighted composite measure. For example, we could assess both number of correct responses and speed of responding in a memory task that taken together might be interpreted as reflecting performance efficiency.
- In a prediction (regression) design, we might use self-esteem, extraversion, and product knowledge to predict dollars of sales for a set of salespeople. The variate in this instance might be thought of as sales effectiveness.
- To determine which items on a personality inventory might comprise separate subscales that measure aspects of a more global construct, we might perform a factor analysis on the responses to those items. Each factor would be a weighted linear combination of the inventory items.

1.3 The Importance of Multivariate Designs

The importance of multivariate designs is becoming increasingly well recognized. It also appears that the judged utility of these designs seems to be growing as well. Here are two of the advantages of multivariate research designs over univariate research designs based on those offered by Stevens (2009):

- Many experimental treatments are likely to affect the study participants in more than one way.
- Using multiple criterion measures can paint a more complete and detailed description of the phenomenon under investigation.

A similar argument is made by Harris (2001):

> However, for very excellent reasons, researchers in all of the sciences—behavioral, biological, or physical—have long since abandoned sole reliance on the classic univariate design. It has become abundantly clear that a given experimental manipulation . . . will affect many somewhat different but partially correlated aspects of the organism's behavior. Similarly, many different pieces of information

> about an applicant . . . may be of value in predicting his or her . . . [behavior], and it is necessary to consider how to combine all of these pieces of information into a single "best" prediction. (p. 11)

In summary, there is general consensus about the value of multivariate designs for two very general reasons. First, we all seem to agree that individuals generate many behaviors and respond in many different although related ways to the situations they encounter in their lives. Univariate analyses by definition, are, able to address this level of complexity in only a piecemeal fashion because they can examine only one aspect at a time. Multivariate analysis allows us to do this as well, but it also affords us the opportunity to examine the phenomenon under study by determining how the multiple variables interface.

The second reason why the field appears to have reached consensus on the importance of multivariate design is that we hold the causes of behavior to be complex and multivariate. Thus, predicting behavior is best done with more rather than less information. Most of us believe that several reasons explain why we feel or act as we do. For example, the degree to which we strive to achieve a particular goal, the amount of empathy we exhibit in our relationships, and the likelihood of following a medical regime may depend on a host of factors rather than just a single predictor variable. Only when we take into account a set of relevant variables—that is, when we take a multivariate approach—have we any realistic hope of reasonably accurately predicting the level—or understanding the nature—of a given construct. This, again, is the realm of multivariate design.

1.4 The General Form of a Variate

The general form of a variate—a weighted composite—is an equation or function. In the weighted linear composite shown below, each variable in the variate is symbolized by the letter X with subscripts used to differentiate one variable from another. A weight is assigned to each variable by multiplying the variable by this value; this weight is referred to as a *coefficient* in many multivariate applications. Thus, in the expression w_2X_2, the term w_2 is the weight that X_2 is assigned (multiplied by) in the weighted composite, that is, w_2 is the coefficient associated with X_2. A weighted composite of three variables would take this general form:

$$\text{weighted composite} = w_1X_1 + w_2X_2 + w_3X_3$$

These weighted composites are given a variety of names, including *variates*, *composite variables*, and *synthetic variables* (Grimm & Yarnold, 2000). Variates are therefore not directly measured by the researchers in the process of data collection but are created or computed as part of or as the result of the multivariate data analysis. Because they are not directly measured, what they assess is often referred to as a *latent construct*, and the variate is often referred to as a *latent variable*. We will have quite a bit to say about variates (weighted linear composites or latent variables) throughout this book.

1.5 The Type of Variables Combined to Form a Variate

Variates may be weighted composites of either independent variables (i.e., manipulated or predictor variables; see Section 2.3.1) or dependent variables (variables representing the outcome of the research; see Section 2.3.2), or they may be weighted composites of variables playing neither role in the analysis. Examples where the analysis creates a variate composed of independent variables are multiple regression and logistic regression designs. In these designs, two or more independent variables are combined together to predict the value of a dependent variable. For example, the number of delinquent acts performed by teenagers might be found to be predictable from the number of hours per week they play violent video games, the number of hours per week they spend doing homework (this would be negatively weighted because more homework time would presumably predict fewer delinquent acts), and the number of hours per week they spend with other teens who have committed at least one delinquent act in the past year.

Multivariate analyses can also create composites of dependent variables. The classic example of this is multivariate analysis of variance (MANOVA). This general type of design can contain one or more independent variables, but there must be at least two dependent variables in the analysis. These dependent variables are combined together into a composite, and an ANOVA is performed on this computed variate as in the case of combining number of correct responses and speed of responding mentioned above. The statistical significance of group differences on this variate is then tested by a multivariate F statistic (in contrast to the univariate F ratio that readers have presumably studied in prior coursework).

Sometimes variables do not need to play the explicit role of either independent or dependent variable and yet will be absorbed into a weighted linear composite in the statistical analysis. This occurs in principal components and factor analysis, where we attempt to identify which variables (e.g., items on an inventory) are associated with a particular underlying dimension, component, or factor. These components or factors are weighted linear composites of the variables in the analysis.

It is possible that the prior experience of readers is such that great emphasis has been placed on the differences between dependent and independent variables. If so, it might be somewhat disconcerting to learn that variates can be composed of either class of variables. But it turns out that, in the analysis of data, dependent and independent are roles that are assigned to variables by the researchers rather than absolute attributes of the variables themselves. And just as actresses in the theater can play different roles in different productions, so too can variables play different roles in different analyses. We will discuss this matter in more detail in Section 2.3.

1.6 The General Organization of the Book

The domain of multivariate research design is quite large, and selecting which topics to include and which to omit is a difficult task for authors. Most of the multivariate procedures we cover in this book are very much related to each other in that they are different surface ways of expressing the same underlying model: the general linear model (see Section 7A.5.1). For example, ANOVA, MANOVA, multiple regression, discriminant function analysis, principal components and factor analysis, and canonical correlation analysis are all members of the general linear model family. Some of the procedures just represent different ways to conceptualize the same analysis. For example, MANOVA focuses on the differences between the groups in the analysis based on a set of quantitative variables whereas discriminant function analysis focuses on (a) the dimensions along which the groups differ on the quantitative variables and (b) the prediction of group membership based on those quantitative variables.

Separating the chapters into groupings is therefore done as a convenience for the readers. The groupings that we use, and even the ordering of the chapters within the groupings, is more of a matter of personal expression than a true classification system. Other authors would likely choose a somewhat different structuring of the topics.

1.6.1 The Chapters Are in Pairs

Beginning with the third chapter, each topic is presented by a pair of chapters labeled "A" and "B." The "A" chapter of the pair treats the topic at a relatively broad, conceptual level, focusing on the uses to which the design is often put, the rationale underlying the procedure, a description of how the procedure works, some of the decisions that are likely to be encountered in performing the analysis, and some issues of controversy when they are germane to the discussion. The "B" chapter of the pair describes in a step-by-step way how to perform the analysis in IBM SPSS (or, in the Part V chapters, IBM SPSS Amos), how to interpret the output of the analysis, and how researchers might report the results of the analysis. Most of the data sets that we use for our examples are modified versions (sometimes very substantially) of ones our students have collected in their research, and we use them with the permission of those students.

We often refer to the examples as being based on fictional studies just to reinforce the idea that the conclusions we draw from them may have little to do with the empirical world in which we live. For each procedure that we perform in our "B" chapters, we present an example of how the results might be reported. It should be emphasized that there is no one best way to report results—we just wanted to illustrate one (hopefully) acceptable way to accomplish this. Readers are encouraged to consult Cooper (2010) for his suggestions on preparing results sections for dissemination.

Sage has established a place for our materials on their website (www.sagepub.com/meyers). The following materials can be found there:

- Exercises with data files for each of the "B" chapters.
- Data files for the analyses demonstrated in each of the "B" chapters.

1.6.2 Part I: The Basics of Multivariate Design

The chapters in this part of the book introduce readers to the foundations or cornerstones of designing research and analyzing data. Our first chapter—the one that you are reading—discusses the idea of multivariate design and addresses the structure of this book.

The second chapter on fundamental research concepts covers both some basics that readers have learned about in prior courses and possibly some new concepts and terms that will be explicated in much greater detail throughout this book.

Chapters 3A and 3B cover data screening. The issues covered here are applicable to all the procedures we cover later, and so we cover them once in this pair of early chapters. We discuss ways to correct data entry mistakes, how to evaluate statistical assumptions underlying the data analysis, and how to handle missing data and outliers.

1.6.3 Part II: Comparison of Means

Part II addresses comparison of means. We begin with a description of some commonly used univariate ANOVA designs in Chapters 4A and 4B and then transition to multivariate ANOVA (MANOVA) designs in Chapters 5A and 5B. The intent of researchers using these designs is to determine which groups or conditions in a study are significantly different on the one or more dependent variables that were measured. Univariate designs assess a single dependent variable in each analysis; multivariate designs assess two or more dependent variables simultaneously in each analysis.

1.6.4 Part III: Prediction of the Value of a Single Variable

Regression procedures are used to predict the value of a single variable. Pearson correlation and simple linear regression, the soul mate of Pearson correlation, are covered in Chapters 6A and 6B. Pearson correlation is used to describe the degree of linear relationship that is observed between two measures (e.g., X and Y). In simple linear regression, one variable (e.g., X) is used as a predictor of the other (e.g., Y). Multiple regression, frequently referred to as ordinary least squares regression, is an extension of simple linear regression when we use multiple measures to predict the Y variable. The basics of this procedure are covered in Chapters 7A and 7B, and some variations of it are discussed in Chapters 8A and 8B.

When the limitations of ordinary least squares regression are exceeded, alternative regression techniques need to be called into play. Two such alternatives are presented in the next two pairs of chapters. Ordinary least squares regression assumes that the cases in the analysis are independent of each other, an assumption that is violated where cases are nested, that is, hierarchically organized. Examples of such organization are students within separate classrooms and clients of particular mental health clinics in a larger health system. In predicting an outcome variable, such as standardized test scores of the students, the children within a given classroom may be more related to each other on the outcome variable than they are to other students selected at random from the entire school or school district. To the extent that the children within a classroom are more alike than students selected at random, that is, to the extent that nesting is important, the assumption of independence is violated and we must use multilevel modeling in predicting the outcome variable. This topic is presented in Chapters 9A and 9B.

Ordinary least squares regression also assumes that the variable being predicted is measured on a quantitative scale of measurement. Yet it is often the case that we wish to predict to which group cases in the data file belong; here, group assignment is represented as a categorical variable. For example, we might want to predict whether an individual is likely to succeed or not succeed in a given program based on a set of variables. This type of prediction can be performed using binary or multinomial logistic regression, topics discussed in Chapters 10A and 10B. Prediction of a binary variable entails setting a decision point so that cases are classified or predicted as belonging to either one group or the other. One powerful and commonly used procedure used to facilitate that decision making is receiver operating characteristic (ROC) curve analysis, and this topic is included within the logistic regression chapters.

1.6.5 Part IV: Analysis of Structure

We very generally mean by structure some underlying relationships among the variables that can be brought to the surface by the statistical analysis. Often, but not always, these underlying relationships are organized into themes or dimensions. The chapters in this portion of the book meet that general criterion, but they comprise a relatively diverse set of procedures.

Discriminant function analysis, covered in Chapters 11A and 11B, is the flip side of MANOVA. It can be used to predict membership in a categorical variable, and so might easily fit in Part III; in fact, there are applications where logistic regression and discriminant function analysis are both considered possible procedures to use to analyze a given data set. But we placed discriminant function analysis here because one of its uses is to describe the dimensions along which groups differ, and with this focus, the interpretations of the obtained discriminant functions take on a decidedly structural orientation.

Principal components analysis and exploratory factor analysis, discussed in Chapters 12A and 12B, both describe the dimensions underlying a set of variables. For example, although a paper-and-pencil inventory may contain two or three dozen items, these items may tap into only three or four latent main themes or dimensions. Principal components analysis and exploratory factor analysis can be used to identify which items relate to each dimension.

Canonical correlation analysis, presented in Chapters 13A and 13B, is an extension of ordinary least squares regression in which a set of quantitative independent variables is used to predict the values of a set of quantitative dependent variables. Yet we conceive of it as structural in that the two sets of variables can be related to each other along several dimensions represented by canonical functions, and so in many ways the process of interpreting the results strongly resembles what we do in principal components and factor analysis.

Chapters 14A and 14B are devoted to multidimensional scaling. Objects or stimuli (e.g., brands of cars, retail stores) are assessed in a pairwise manner to determine the degree to which they are dissimilar. These dissimilarities are analyzed in terms of the distance between the objects. In turn, the distances between the objects are arrayed or represented in a space defined by the number of dimensions specified by the researchers who then attempt to interpret these dimensions along which the objects appear to differ.

Cluster analysis is described in Chapters 15A and 15B. Rather than using common demographic variables to define groups (e.g., females and males), we group the cases (e.g., participants in a research study, presidents of the United States, brands of beer) on the basis of how they relate based on a set of quantitative variables. These groupings are called clusters. Two different approaches to such an analysis are described in the chapter.

1.6.6 Part V: Fitting Models to Data

The chapters in this section deal with fitting (causal or predictive) models to data and determining the quality of the fit. Again, our classification schema is not perfect, as multidimensional scaling does involve fitting a dimensional solution to the data set. Nonetheless, the models referred to in the Part V chapters represent more explicit hypotheses that are tested by researchers based on their best understanding of the phenomenon under study. Most of the statistical work must be accomplished by a specialized piece of software called IBM SPSS Amos which is only available at this time with the Windows (but not the Mac) version of IBM SPSS.

Principal components analysis and, to a large extent, exploratory factor analysis (both are discussed in Chapters 12A and 12B) are analogous to an inductive approach in that researchers employ a bottom-up strategy by developing a conclusion from specific observations. That is, the researchers determine the interpretation of the factor by examining the variate that emerged from the analysis. Confirmatory factor analysis, presented in Chapters 16A and 16B, represents a deductive approach in that researchers are predicting an outcome from a theoretical framework; this strategy can be thought of as a top-down approach. Confirmatory factor analysis seeks to determine if the number of factors and their respective measured variables as specified in the model hypothesized by the researchers is supported by the data set—that is, they determine the extent to which the proposed model fits the data.

Path (sometimes called causal) structures are presented in the next two sets of chapters. When the variables in the hypothesized structure are all measured variables, we speak of path analysis, which can be analyzed through either ordinary least squares regression or IBM SPSS

Amos; this topic is treated in Chapters 17A and 17B in the context of multiple regression and Chapters 18A and 18B in the context of IBM SPSS Amos. We also discuss procedures to trim some unneeded paths from models. When we have included "factors" in the path structure—unobserved or latent variables or constructs—the analysis becomes one of structural equation modeling (SEM) and must be done in IBM SPSS Amos (or comparable specialized software); this topic is treated in Chapters 19A and 19B.

Chapters 20A and 20B deal with the issue of a given model being applicable to two or more groups. When this is the case, the model is said to be invariant with respect to those groups. For example, a specific model may fit the data obtained from both males and females or from Asian American, White American, and Latino/Latina American students. A couple of different applications are covered in the chapter.

1.7 Recommended Readings

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45*, 721–734.

Grimm, L. G., & Yarnold, P. R. (2000). Introduction to multivariate statistics. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 3–21). Washington, DC: American Psychological Association.

Harlow, L. L. (2005). *The essence of multivariate thinking*. Mahwah, NJ: Lawrence Erlbaum.