Chapter 1
Introduction

> Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. Education is concerned with changes in human beings; a change is a difference between two conditions; each of these conditions is known to us only by the products produced by it—things made, words spoken, acts performed, and the like. To measure any of these products means to define its amount in some way so that competent persons will know how large it is, better than they would without measurement. To measure a product well means so to define its amount that competent persons will know how large it is, with some precision, and that this knowledge may be conveniently recorded and used. This is the general credo of those who, in the last decade, have been busy trying to extend and improve measurements of educational products (Thorndike, 1918, p 16).

Psychometrics is that area of psychology that specializes in how to measure what we talk and think about. It is how to assign numbers to observations in a way that best allows us to summarize our observations in order to advance our knowledge. Although in particular it is the study of how to measure psychological constructs, the techniques of psychometrics are applicable to most problems in measurement. The measurement of intelligence, extraversion, severity of crimes, or even batting averages in baseball are all grist for the psychometric mill. Any set of observations that are not perfect exemplars of the construct of interest is open to questions of reliability and validity and to psychometric analysis.

Although it is possible to make the study of psychometrics seem dauntingly difficult, in fact the basic concepts are straightforward. This text is an attempt to introduce the fundamental concepts in psychometric theory so that the reader will be able to understand how to apply them to real data sets of interest. It is not meant to make one an expert, but merely to instill confidence and an understanding of the fundamentals of measurement so that the reader can better understand and contribute to the research enterprise.

With the advent of powerful computer languages that have been developed with the specific aim of doing statistics (R and S+), the process of doing psychometrics has become more approachable. It is no longer necessary to write long programs in Fortran (Backus, 1998), nor is it necessary to rely on sets of proprietary computer packages that have been developed to do particular analyses. It is now possible to use R for almost all of one's basic (and even advanced) psychometric requirements.

R is an open source implementation of the computer language S. As such it is available free of charge under the *General Public License* (*GPL*) of the GNU Project of the Free Software

Foundation.[1] The source code of all R programs and packages are open to inspection and change. The core of R has been developed over the past 20 years by a dedicated group (the R Core Team) of about 15-20 members which includes some of the original authors of S. In addition, there are at least 1000 packages that are written in R and contributed to the overall R project by many different authors. In the psychometrics community, at least 10 to 20 packages have been developed and made available to the psychometrics user in particular and the R community in general.

Like psychometrics, R is initially daunting. Also like psychometrics, while it takes years to master, the basics can be learned fairly easily and expertise comes with practice. Moreover, combining examples written and analyzed in R with psychometric problems allows one to learn both at the same time. This text is thus both an introduction to psychometric theory as well as to R .

1.1 An overview of the book

The structure of this book is best represented in the form of a "structural model" showing a number of boxes and circles with a set of connecting paths (Figure 1.1). This symbolic notation is a way of showing the relationship between a set of *observed* variables (the boxes on the far left and right of the figure) in terms of a smaller set of unobserved, or *latent*, variables said to account for the observed variables (the circles in the middle of the figure). Paths in the figure represent relationships. In the following chapters we will use this figure to help locate where we are.

Part I: Basic issues

1.1.1 Constructs and measures (Chapter 2)

A basic distinction in science may be made between theoretical constructs and observed measures thought to represent these constructs. This distinction is perhaps best understood in terms of Plato's Allegory of the Cave in the Republic (Book VII). Consider a group of prisoners confined in the darkness of a cave. They are chained so that they face away from the mouth of cave and can not observe anything behind them. Behind them there is a fire and people are walking back and forth in front of the fire carrying a variety of objects. To the prisoners, all that is observable are the shadows cast on the wall of the cave of the walking people and of the objects that they carry. From the patterns of these shadows the prisoners need to infer the reality of the people and objects. These shadows are analogous to

[1] Under the Free Software Foundations's GPL, all programs are required to have the following statement. "This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details." The `GNU.org` website has an extensive discussion of the meaning of "free" software.

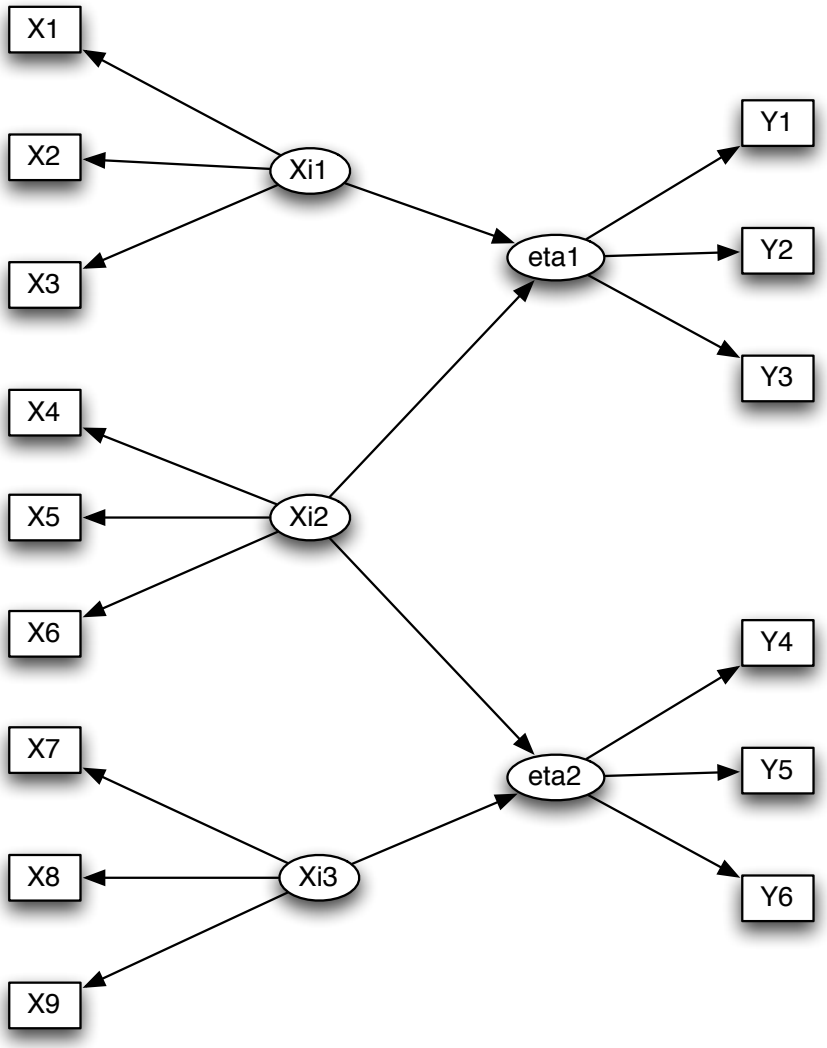

Fig. 1.1 A conceptual overview of the book. Psychometrics integrates the relationships between observed variables (rectangles) in terms of latent or unobserved constructs (ellipses). Chapter 2 considers what goes into a observable (e.g., the box X_1) while Chapter 3 addresses the shape of the mapping function between a latent variable and an observed variable (path Ξ_1 to X_1). Chapter 4 considers how to assess relationships between two variables (e.g., the simple correlation or regression of X_1 and Y_1), how to combine two or more variables to predict a third variable (multiple regression of Y_1 on X_1 and X_2), and how to assess the relationship between two variables while holding the effect of a third variable constant (partial correlation of X_1 and Y_1 with X_2 partialed out). Chapter 6 considers how to estimate the correlation between an observed variable and a latent variable (X_1 with Ξ_1) to estimate the reliability of X_1. Chapter 9 addresses the question of how many latent variables are needed. Chapter 8 considers how the structure of relationships allows alternative conceptualizations of validity. Chapter 10 addresses how to model the entire figure.

the observed variables that study, while the "real" but unobservable people and objects are the latent variables about which we make theoretical inferences. The prisoners make their inferences based upon the patterning of the shadows.

While behaviorism reigned supreme in psychology until the late 1950's, the emphasis was upon measurement of observed variables. With a greater appreciation of the process of theory building, the use of latent variables and hypothetical constructs gradually became more accepted in psychology in general, and in psychometrics in particular. For rarely are we interested in specific observations of twitches and utterances. Psychological theories are concerned with higher level constructs such as love, intelligence, or happiness. Although these constructs can not be measured directly, good theory relates them through a process of measurement to specific observed behaviors and physiological markers.

A very deep question, and one that will not be addressed in the detail it requires, is what does it to mean to say we "measure" something? The assigning of numbers to observations does not necessarily (and indeed probably does not) imply that the data are isomorphic to the real numbers. The types of inferences that can be made from observations and the types of analysis that are or are not appropriate for the observed data is a deep question that goes beyond the scope of an introductory text.

1.1.1.1 Observational and Experimental Psychology

A recurring debate in psychology ever since Wundt (1874) and Galton (1865, 1884) has been the merits of experimental versus observational approaches to the study of psychology. These two approaches tend to represent different sub fields within psychology and to emphasize different types of training. Experimental psychology tends to emphasize central tendencies and strives to formulate general laws of behavior and cognition. Observational psychologists, on the other hand, emphasize variability and covariation and study individual differences in ability, character, and temperament.

For many years these two approaches also used different statistical procedures to analyze data, with experimentalists comparing means using the *t-test* and its generalization, the Analysis of Variance (ANOVA). This was in contrast to observationalists who would study variability and particularly covariation with the correlation coefficient and multivariate procedures. However, with the recognition that ANOVA and correlations are just special cases of the general linear model, the statistical distincition is less relevant than the distinction of research methodologies.

Despite eloquent pleas for the reunification of these two disciplines Cronbach (1957, 1975), Eysenck (1966, 1997) and Vale and Vale (1969) there is surprisingly little emphasis upon individual differences within experimental (but see Underwood, 1975) for why individual differences are necessary for theory building in cognitive psychology) or experimental techniques in personality research (Revelle and Oehleberg, 2008). However, both research approaches need to understand the quality of measurements in order to make experimental or correlational inferences.

With an emphasis upon the type of data we collect, and the problems of inference associated with the metric properties of our data, chapters 2 & 3 are particularly relevant for the experimentalist who wants to interpret differences between observed means in terms of differences at a latent level. These two chapters are also important for the student of correlations who wants to interpret scale score values as if they have more than ordinal meaning.

1.1.1.2 Data = Model + Error

Perhaps the most important concept to realize in psychometrics in particular and statistics in general is that we are *modeling* data, not merely reporting it. Complex data can be partly understood in terms of simpler theories or models. But our models are incomplete and do not completely capture the data. The data we collect, no matter how carefully we do it, nor how well we understand the process that generates the data, are never quite what we expect. The world is complex and the observations that we take are multiply determined. At the most basic atomic level of physics, quantum randomness occurs. At the level of the neuron, coding is a statistical frequency of firing, not a binary outcome. At the level of human behavior although our theories might be powerful (which they tend not to be) there are causes that have not been observed.

Throughout the book we will propose models of the data and try to evaluate those models in terms of how well they fit. Conceptually, we use the equations

$$Data = Model + Error \tag{1.1}$$

to represent the problem of inference. We evaluate how well our models fit by examining error as defined as

$$Error = Data - Model \tag{1.2}$$

and evaluate the magnitude of some function of the error (Judd and McClelland, 1989). These equations would seem to imply a greater quantitative level of measurement precision than is generally the case, and should be treated as abstractions to remind us that we are *evaluating models* of data and need to continuously ask how appropriate is the model. The distinction between model and error is not new, for it dates at least to Plato. As discussed by Stigler (1999) the whole of nineteenth century statistical theory was based upon this distinction between physical truth as modeled by Newton and actual observations taken to extend the theories.

1.1.2 A theory of data (Chapter 2)

Clyde Coombs introduced a taxonomy of the kinds of data that we can observe that allows us to abstractly organize what goes into each observation (Coombs, 1964). Although many of the examples in this book are drawn from a small portion of the kinds of data that could be collected, by thinking about the basic distinctions made by Coombs we see the range of possibilities for psychometric applications.

1.1.3 Basic summary statistics – problems of scale (Chapter 3)

The problem of how to summarize data reflects some deep issues in measurement. The naive assumption that our measures are *linearly* related to our constructs leads to many a misinterpretation. Indeed, to some, this assumption reflects a pathological thought disorder (Michell, 1997). Although not taking such a strong position, examples of misinterpretation of findings because of faulty assumptions of interval or ratio levels of measurement are easy to find (3.5).

Alternative ways of estimating central tendencies can lead to very different conclusions about sets of data (3.4).

1.1.4 Covariance, regression, and correlation (Chapter 4)

Perhaps the most fundamental concept in psychometrics is the correlation coefficient. How to best represent the relationship between two or more variables is a fundamental problem that, if understood, may be generalized to most of psychometrics. Correlation takes many forms and understanding when to use which type of correlation is important. Understanding how multiple and partial correlation are generalizations of the simple zero order correlation allows for an understanding of classic reliability theory.

Part II: Classical and modern reliability theory

1.1.5 Classical theory and the Measurement of Reliability (Chapter 6)

How well does a scale measure whatever it is measuring? Do alternative measures give the same or similar values? Are measures of the construct the same over time, over items, over situations? These are the basic questions of both classical test theory as well as its generalization to Latent Trait and Item Response Theory.

1.1.5.1 Parallel tests and their generalizations

If one observed scale is thought to be composed of True score and Error, then the correlation of the test with true score may be calculated in terms of the proportion of variance that is True score. But how to estimate this? Parallel tests, tau equivalent tests, and congeneric test models make progressively fewer assumptions of the data and all yield ways of estimating true scores.

1.1.5.2 Domain sampling theory

An alternative approach, that yields the same solution as congeneric test theory is to think of tests as representing larger and larger sets of items sampled from an infinite domain of items. By thinking in terms of domain sampling, the meaning of alternative estimates of reliability (α, β, ω) is easily understood. Hierarchical structures of tests in terms of group and general factors emphasizes the need to understand the test structure.

1.1.5.3 The many sources of reliability: Generalizability theory

Reliability needs to be considered in terms of the dimensions across which we want to generalize our measures. A measure of a single trait should be a good indicator of a single domain and should be consistent across time and situation. A measure of a mood, however, should have high internal consistency (be a good measure of a single domain) but should not show consistency over time. In a prediction setting, it is possible that a test need not have high internal consistency, but it should be stable over time, situations, and perhaps forms. Reliability is not just a concept of the item, but also is concerned with the source of the data. Do multiple raters agree with each other, do various forms of the test give similar answers.

1.1.6 Latent Trait Theory - The "New Psychometrics" (Chapter 7)

Although Classical Test Theory treats items as random replicates, it is possible to consider item parameters as well. This leads to a more efficient means of estimating person parameters and also emphasizes issues of scaling shape. Extensions to two and three parameter models, ability and unfolding models, and dichotomous versus polytomous models will be considered.

1.1.7 Validity (Chapter 8)

Does a test measure what it supposed to test? How do we know? The most direct (and perhaps least accurate) way is to simply examine the item content. Does the content appear related to the construct of interest. *Face* (which is sometimes known as "Faith") validity, address the question of obvious relevance. Do questions about psychometric knowledge ask about matrices or do they ask about general knowledge of English? The former item would seem to be more valid than the latter.

If tests are used for selection or diagnosis, merely looking good is not enough. It is also necessary to assess how well the measure correlates with current status of known criterion groups or how well the test predicts future outcomes. *Concurrent* and *predictive* validity assess how well tests correlate with alternative measures of the same construct right now and do they allow future predictions?

For theory testing and development, validity is a process more than a particular value. In assessing the *construct* validity of a measure, it is necessary to examine the location of the test in the complete nomological network of the theory. Assessing *convergent* validity asks whether measures correlate with what they should correlate with given the theory? Equally important is *discriminant* validity: Do measures not correlate with what the theory says they should not correlate with? A final part of construct validity is emphincremental validity: does it make any difference if we add a test to a battery?

1.1.7.1 Decision Theory

The practical use of tests also involves knowing how to combine data to make decisions. Although the measures used to predict and to validate tend to be continuous, decisions and

outcomes are frequenenty discrete. Students are admitted to graduate school or they are not. They finish their Ph.D. or they do not. People are offered jobs, are promoted, are accused of crimes, and are found guilty or innocent. These are binary decisions and binary outcomes based upon linear and non-linear models of predictors. In addition to considering the base rates of outcomes and the selectivity of the choice process, the utility of test reflects the value applied to the various types of outcomes as well as the cost of developing and giving the test.

Part III: Latent variables

1.1.8 Factor, Principal Components and Cluster Analysis (Chapter 9)

How many latent constructs are represented in a data set of N variables? Is it possible to reduce the complexity of the data without a great loss of information? How much information is lost?

1.1.9 Exploratory versus Confirmatory models

Early in the development of a particular subarea, exploratory data reduction techniques are most appropriate. These range techniques can include cluster analysis and principal components analysis as well as exploratory factor analysis. All of the procedures are faced with the problem of what is the appropriate amount of data reduction and how to evaluate alternative models. Confirmatory models, primarily confirmatory factor models, are special cases of structural equation modeling procedures.

1.1.10 Structural Equation Modeling (Chapter 10)

Structural Equation Modeling = Reliability + Validity. How to evaluate the measurement model and the structural model at the same time. There are severe limitations on the type of inferences that can be drawn, even from the best fitting structural equation. Through the use of simulated data representing various threats to measurement, it is possible to better understand how to properly interpret results of standard sem packages.

Part IV: The construction of tests and the analysis of data

1.1.11 Scale construction (Chapter 15)

Practical suggestions about how to construct scales based upon basic item statistics. For students and practitioners with limited resources, some procedures are much more useful

than others. What are the tradeoffs involved in making particular decisions when developing tests to use in research and applied settings. Knowing a few simple rules of test construction and evaluation helps speed up the cycle of test development.

Appendices

1.1.12 Appendix – Basic R

The basic commands and methods for using R . How to get, install, and use the basic R packages.

1.1.13 Appendix - Review of Matrix algebra

Although understanding matrix algebra is not completely necessary to understand psychometrics, it makes it much easier. Because it is more abstract, matrix notation is far more compact than the alternative notation using the summation of cross products. In addition, programming operations in R using matrices produces much cleaner and faster code. This appendix is what you should have learned in college but have probably forgotten.

1.2 General comments

This book is aimed for beginners in psychometrics (and perhaps in R) who want to use the basic principals of psychometric theory in their substantive research. As an introduction to psychometrics some major philosophical issues about the meaning of measurement (e.g, Barrett, 2005, Borsboom, 2004, and Michell, 1997) will not be discussed in the detail they deserve, nor will many of the basic models be derived from first principles in the manner of Guilford (1954), McDonald (1999), or Nunnally (1967). It is hoped, however, that the reader will become interested enough in the theory and practice of psychometrics to delve into those much deeper texts.

Most scientists read books backwards. That is, we start at the later chapters and if understand them, we are finished. If we don't , we go to an earlier chapter and test ourselves with that. For that reason, the appendix on R is meant to allow the eager reader to start running programs in R without reading anything else. However, the introductory chapters are meant to be useful as they consider the meaning of our observations, the inferences we are able to draw from observations and the inferences we can not make.