

CHAPTER 1. INTRODUCTION

1.1 Latent and Observed Variables

Social and behavioral science theories often involve abstract concepts that can be difficult to measure. Rich concepts such as social capital, racial animus, depressive symptoms, or self-efficacy are poorly approximated by single survey or questionnaire items. In many cases, analysts instead rely on multiple items to capture a concept. For instance, rather than a single question such as “During the past week, how often did you feel sad?” to capture depressive symptoms, researchers may ask several questions covering different facets of depressive symptoms, such as “During the past week, how often did you feel that everything you did was an effort?” and “During the past week, how often was your sleep restless?.” These items, or observed variables, can be considered measures (or indicators) of an underlying, unobserved latent variable that represents the concept of interest.

Classical measurement theory (Lord et al., 1968) provides a framework for thinking about the relationship between observed indicators or measures and underlying latent variables. In this framework, an observed variable (x) includes two components: (1) a true component (τ) that reflects the concept of interest and (2) an error component (ε) that reflects other sources of variation in the observed variable. This relationship is represented by the simple equation $x = \tau + \varepsilon$. For instance, some component of variation in the question “During the past week, how often did you feel sad?” captures depressive symptoms and some component reflects measurement error. The key challenge, then, lies in accounting for or minimizing the presence of measurement error in the observed variable when analysts are interested in the underlying latent variable. The specification of a measurement model via confirmatory factor analysis (CFA) provides one solution to this challenge.

1.2 Reliability and Validity

Psychometricians have distinguished two measurement properties, reliability and validity, of observed indicators of latent variables (or combinations of indicators, such as scales). Drawing on the often-used darts analogy, a reliable measure is one in which repeated measurements result in quite similar responses (i.e., a cluster of darts all hitting around the same point). In contrast, a valid measure is one in which the measure captures the under-

lying concept that is the target of measurement. These two measurement properties are distinct in that a measure may be reliable but not valid or vice versa.

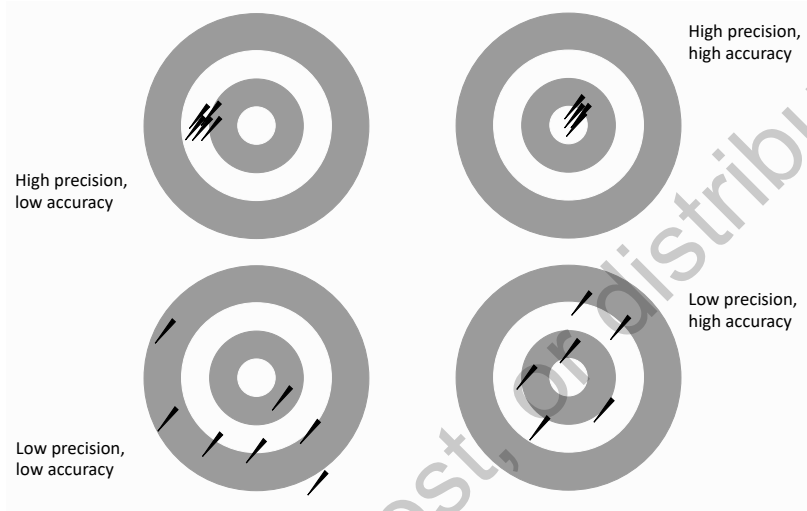


Figure 1.1: Reliability and Validity.

Figure 1.1 presents the relation between reliability and validity visually, with four bulls-eye targets struck by darts. The lower left target is low reliability, low validity—thus, the darts are not clustered well, nor is the grouping centered on the target. The upper left target is high reliability, low validity—the darts are grouped tightly but are far from the center of the target. The lower right target is the inverse, low reliability and high validity—the darts are not grouped tightly, but the grouping is centered on the target. Thus, either reliability or validity may either be high while the other is low. Last, the upper right target displays both high reliability and validity—the darts are grouped tightly and placed at the center of the target. This is the measurement ideal.

There are a number of forms of validity, including content validity, criterion validity, construct validity, and convergent and discriminant validity (Bollen, 1989). The first form, content validity, refers to whether the measures or indicators of a latent variable fully capture the domain of the latent variable. This is primarily a theoretical or conceptual concern and is not evaluated empirically. The next two forms, criterion validity and construct validity, capture the extent to which measures relate to other measures

or variables that they should relate to (e.g., whether a questionnaire item measuring depressive symptoms relates to a clinical diagnosis of depression). These two forms can be empirically evaluated, but such tests rely on obtaining additional variables to do so. The fourth form, convergent and discriminant validity, concerns whether measures of a given latent variable are related to each other in the first case and whether measures of different latent variables are not related to each other in the second case. CFA provides a means of evaluating both the reliability of measures and the convergent and discriminant validity of measures.

1.3 Confirmatory Factor Analysis

CFA is a statistical model that addresses the measurement of latent variables (or factors) through the specification of a measurement model—that is, the relationships between latent variables and the observed variables (or indicators) that are thought to measure them. The fundamental idea underlying factor analysis is that a relatively small number of latent variables (or factors) are responsible for the variation and covariation among a larger set of observed indicators. The initial goal then is to identify the number of latent variables and the nature of the relationships between the latent variables and the observed indicators. Exploratory factor analysis (EFA) makes relatively few assumptions about either the number of latent variables or the relationships between the latent variables and the observed indicators and allows this information to emerge from the data. CFA, by contrast, requires theoretical or substantive knowledge to specify a measurement model in advance and then evaluates how well the specified model fits with the means, variances, and covariances of the observed indicators.

CFA and the specification of measurement models have a number of uses in the social and behavioral sciences. First, it is an invaluable tool in the psychometric evaluation of measurement instruments. For instance, an analyst may propose new questionnaire items to measure racist attitudes and could evaluate the quality of these items using CFA. Second, relatedly, CFA can be used for construct validation, particularly through assessments of convergent and discriminant validity. Third, CFA can allow analysts to explore and address various method effects, such as the effect of similar wording in the racist attitudes model. Fourth, CFA provides a framework for evaluating measurement invariance. For example, the latent variables capturing racist attitudes and the relationships between the latent variables and the indicators may vary across subgroups of the population (e.g., perhaps the measurement model differs for people living in the western United States as compared

with those living in the northeast). Finally, CFA is an important first step in the specification of a broader structural equation model that allows for an analysis of structural relationships among latent variables as well as the specification of multiple endogenous variables.

Before we move to an empirical example, a note on types of indicators. Alwin (2007) details the difference between *multiple measures* and *multiple indicators*. In the first case, *multiple measures*, the intent is that identical or near-identical replicates of measures are administered more than once. In a survey, these might be the same item with word order slightly changed, or even with identical wording but administered at a later period of data collection or merely at a different point within the same period of data collection (i.e., later in the same survey interview). The degree to which item wording may vary while measures are considered identical is uncertain, and the less similar they are, the more they approach the condition of *multiple indicators* and the more complicated assessing reliability becomes. Additionally, it is generally assumed that *multiple measures* measure one and only one thing. In contrast with *multiple measures*, *multiple indicators* are not assumed to perfectly measure the same thing (a situation analysts working with secondary data often find themselves in), even though they are theoretically assumed to be measures of the same construct (an assumption we can test with CFA models). In the case of *multiple indicators*, the assumption that indicators relate only to the construct of interest is also relaxed, and it is possible that other variables may influence them (and thus, explain part of their variance). In the example that follows (and, indeed, most examples in this book), indicators are assumed to be *multiple indicators*.

To give an example, suppose an analyst is interested in characterizing the latent variables or factors that capture dimensions of contemporary racist attitudes in the United States. Social scientists have devised a set of six measures (see Table 1.1) that have been included in nationally representative surveys, such as recent waves of the American National Election Studies (ANES). The first two measures are generated from four raw survey items (two each about each group) where respondents are asked to rate on a scale of 1 to 7 for Whites and Blacks how hardworking versus lazy they are and how peaceful versus violent they are. These items are used to construct differences in the ratings for Whites and Blacks on laziness/hardworking and peaceful/violent. The second set of four measures ask respondents to indicate their agreement on a continuum from “agree” to “strongly disagree” with a series of statements about racist attitudes. Thus, an analyst working with these measures has six indicators to consider.

Background theoretical and substantive knowledge might lead an analyst to specify a measurement model for these indicators that involves two latent

Table 1.1: Item prompts and responses for racial attitudes indicators (American National Election Studies).

Next are some questions about different groups in our society. Please look, in the booklet, at a 7-point scale on which the characteristics of the people in a group can be rated.

| | | |
|-----------------------|--|------------------------------|
| Lazy ^a | Where would you rate [group] in general on this scale? | [1] Hard-working to [7] lazy |
| Peaceful ^a | Where would you rate [group] in general on this scale? | [1] Peaceful to [7] violent |

The following questions are concerned with racial attitudes with responses ranging from strongly disagree (SD) to strongly agree (SA).

| | | |
|----------------|--|------------------|
| Try hard | It's really a matter of some people not trying hard enough; if Blacks would only try harder they could be just as well off as Whites. | [1] SD to [5] SA |
| Prejudice | Irish, Italians, Jews, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors. | [1] SD to [5] SA |
| Discrimination | Generations of slavery and discrimination have created conditions that make it difficult for Blacks to work their way out of the lower class. | [1] SD to [5] SA |
| Deserve | Over the past few years, Blacks have gotten less than they deserve. | [1] SD to [5] SA |

^aIndicators for a measurement model were calculated as the difference in these items for Black and White groups.

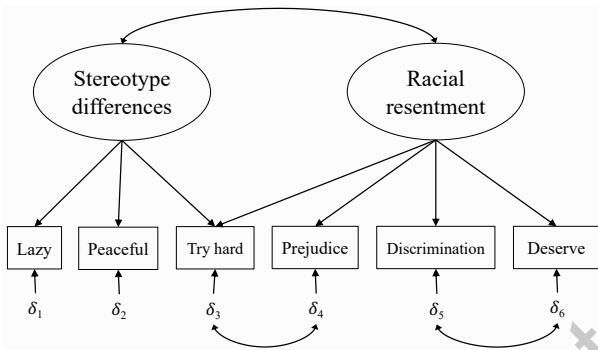


Figure 1.2: Proposed measurement model for racist attitudes.

variables. The first latent variable, which we might label “stereotype differences,” is measured by the first three indicators, and the second latent variable, which we might label “racial resentment,” is also measured by the third indicator as well as the remaining three indicators. Among the four indicators of racial resentment, one pair of the indicators is worded in terms of systemic oppression (e.g. “have gotten less than they deserve”) and another in terms of individual actions (e.g. “if Blacks would only try harder”). These similarities in wording have the potential to introduce some amount of shared variation among the pairs of the indicators that goes beyond what is accounted for by the latent variable racial resentment. To reflect this, an analyst can specify covariances between the errors for respective indicators.

Figure 1.2 depicts a diagram that illustrates this measurement model. By convention, in diagrams for CFAs, circles or ovals indicate latent variables, squares or rectangles indicate observed variables, directed arrows indicate effects from the variable at the origin of the arrow to the variable at the destination of the arrow, and curved two-headed arrows indicate a covariance between two variables. In this particular figure, the lower curved arrows point to deltas, or δ_1 to δ_6 . These Greek symbols represent measurement error or the portion of the variance in each indicator not explained by the latent variable. For this measurement model, we see the two latent variables, stereotype differences and racial resentment, represented in circles and allowed to be correlated. This specification captures the idea that these are two related but distinct dimensions of racist attitudes. We also see direct effects from the latent variables to the respective indicators that are measures of each. As noted above, the third indicator is thought to be a measure

for both latent variables, and thus we see an arrow from each pointing to it. Finally, the curved two-headed arrows among the measurement errors capture the correlated errors for the two sets of indicators that measure racial resentment. This figure presents measurement models with *effect indicators*, or indicators influenced by latent variables. Other measurement forms in which indicators influence latent variables are possible (see discussion in Chapter 2).

This measurement model for racist attitudes illustrates a number of features of CFA that capture the richness of the method. We see two dimensions of racist attitudes as reflected in the two latent variables, an indicator that serves as a measure for more than one latent variable and correlated errors that account for potential method effects from similarities in the wording of the questionnaire items (this is not the only source for correlated error terms; we discuss these in greater detail in Chapter 2).

Now that we have a specified measurement model for racist attitudes, we can test the overall fit of the model and estimate the model parameters given a dataset with these measures (e.g., the ANES). As we will discuss, for many measurement models CFA allows analysts to test and assess how well the proposed model specification reproduces the means, variances, and covariances among the observed indicators. This provides a holistic assessment of the measurement model and is a key advantage of CFA as compared with EFA.

In addition to evaluating the overall fit of the measurement model, we can estimate the parameters of the measurement model. The main parameters of any CFA include “factor loadings,” latent variable or factor variances and covariances if the model includes more than one latent variable, and error variances and covariances if the model specification permits as in our example. The factor loadings capture the effect of the latent variable on the indicators and are typically interpreted as regression coefficients. The factor variances capture the extent of variation across cases in the underlying latent variables. Similarly, the covariances among the factors capture the extent to which the latent variables covary with each other and can be an important component in assessing validity. For instance, if the correlation between stereotype differences and racial resentment is very close to 1.0, then we might question whether the items allow us to identify two distinct dimensions of racist attitudes. The error variances, sometimes referred to as “unique” variances, capture a combination of systematic sources of variation that affect only a given indicator and random error. We typically think of the combination of the two as “measurement error,” and estimates of error variances allow us to assess the reliabilities of the indicators. Finally, the error covariances capture shared sources of variation in pairs of indicators that

remain after accounting for any variation due to latent variables. We may also determine at this time whether to fix one loading per latent variable to one for identification purposes (the default in most recent software), or use other options to allow estimation of each loading (for more on these choices and model identification, see Chapter 3).

Once we have fit our measurement model to data, evaluated the overall fit of the model, and examined the parameter estimates, it is possible that we will obtain adequate model fit and reasonable parameter estimates and thus we may proceed with additional analyses or interpreting the various estimates. Alternatively, it is also possible that the model fit will be inadequate or parameter estimates will be unreasonable (e.g., a negative estimate for an error variance), in which case we may explore alternative measurement model specifications.

1.4 Statistical Software and Code

Applied researchers have a number of options for statistical software to fit CFA models. Most general packages, such as R, Stata, and SAS, have the capability to fit various types of CFA models. In addition, packages designed specifically for structural equation models, such as Mplus, LISREL, and EQS, can be used for an even wider range of CFAs. Because the capabilities and syntax of the various statistical software packages continue to evolve, we have decided not to include snippets of code or output in the book. Instead we have uploaded code for multiple software packages (primarily R, Stata, and Mplus) along with data for every example in the book to a companion website at study.sagepub.com/researchmethods/qass/roos-confirmatory-factor-analysis to serve as a resource for readers looking to replicate the examples or use as templates for their own work. We also indicate the specific software package used for the estimates reported for the examples in the book in the notes to each table.

1.5 Outline of the Book

Chapter 2 provides a detailed introduction to model specification in CFA that includes a consideration of different types of multidimensional measurement models and the distinction between effect and causal indicators. Chapter 3 covers model identification and estimation in CFA. Model identification concerns whether it is possible to obtain unique parameter estimates and is closely connected to tests for the fit of a specified model.

Chapter 4 details model testing, both overall or global model fit and item-specific fit, and nested models. That chapter also discusses strategies for model respecification if fit is unsatisfactory. Chapter 5 examines measurement invariance—that is, whether or not items perform similarly across different groups. Assessing measurement invariance is a key part of minimizing bias in measurement. Chapter 6 introduces categorical indicators, an important extension of traditional CFA models that addresses the categorical level of measurement of many observed indicators of latent variables. Chapter 7 closes the book with concluding remarks about the broader value of CFA, a discussion of how CFA fits within the more general SEM framework, and a consideration of advanced topics with CFA along with recommendations for further study.

1.6 Further Reading

For a more in-depth examination of social concepts and their measurement, see Goertz (2020). Alwin (2007) provides an extended treatment for readers interested in learning more about reliability, validity, and measurement error with survey data. For a fuller treatment of the measurement model presented here as an empirical example, see Roos et al. (2019). For another exemplar of a measurement model with multiple latent variables, see Manglos-Weber et al. (2016).